

AI-ASSISTED DATA CLASSIFICATION FOR ENTERPRISE GOVERNANCE: BALANCING ACCURACY, PRIVACY, AND OPERATIONAL EFFICIENCY

Sreehari Prasad Jinka Venugopal

MBA, Product Management

Enterprise Product Management, Data Protection, Cyber Resilience, Trustworthy AI Governance, Phoenix, Arizona, USA
mail09hari@gmail.com | ORCID: 0009-0002-5969-9357

Received: 20/08/2026

Revised: 24/09/2026

Accepted: 07/10/2026

ABSTRACT:

Data classification is the foundation of enterprise data governance: retention, access control, encryption, sharing and incident response all depend on knowing how sensitive a document is. Manual classification does not scale, and rule-based tools are brittle, while machine learning promises scale but raises new concerns, because training a classifier on sensitive content can itself create privacy exposure and because errors in classification have very different costs in different directions. This paper proposes the Governed AI Classification Framework (GACF), which treats accuracy, privacy and operational efficiency as three objectives to be balanced explicitly. GACF combines a governed sensitivity taxonomy, a privacy-preserving training architecture based on federated learning with gradient clipping and noise injection, confidence-based triage that routes uncertain documents to human reviewers, and a three-level audit structure covering governance, model and application. We examine the framework in a simulation of five business units with heterogeneous documents and four sensitivity levels. A federated model reached 94.4% accuracy without raw data leaving the units, compared with 96.4% for a centralized model trained on pooled data and 74.6% for rule-based keywords. Adding privacy noise reduced accuracy gradually at moderate levels (88.5% at a noise multiplier of 1) and sharply at high levels (62.4% at 4), and the rate of under-classification, the most dangerous error, rose from 4.8% to 21.6%. Confidence-based triage raised overall accuracy to 96.8% while automating 76% of documents, which would cut reviewer effort for 10,000 documents from 250 hours to about 59. These results are synthetic and show the logic of the design, not performance on real documents. The paper closes with governance guidance and a research agenda.

Keywords: Data Classification, Data Governance, Federated Learning, Differential Privacy, Human-in-the-Loop, AI Auditing, Sensitive Data, Enterprise Information Management

INTRODUCTION

Enterprises create information at a rate that no governance team can review by hand. Contracts, design documents, customer records, financial models, support transcripts, source code and meeting notes accumulate in file shares, collaboration platforms, email and cloud storage. Almost every governance control depends on a judgment about each item: how sensitive is it, who may see it, how long must it be kept, whether it may leave the organization, and what protection it needs at rest and in transit. These judgments are recorded as classification labels, and when they are missing or wrong, downstream controls are applied wrongly.

Traditional approaches have well-known limits. Manual labeling by authors depends on awareness and discipline, and studies of real programs show that users label inconsistently and often skip labeling altogether. Rule-based discovery, using patterns and keywords for identifiers such as payment card numbers, works well for structured identifiers but poorly for contextual sensitivity, such as an unannounced strategy paper or a draft contract, where no fixed pattern marks the content as sensitive. Machine learning classifiers can learn contextual cues and scale to millions of documents, which explains the interest in AI-assisted classification.

That promise brings three tensions. The first is *accuracy against risk*. Errors are not symmetric: labeling a confidential document as public exposes it, while labeling a public document as confidential merely inconveniences its readers. A governance-grade classifier must therefore be judged by the kind of error it makes, not only by how often it errs. The second is *accuracy against privacy*. The best classifier is typically trained on all the data in one

place, but pooling sensitive documents for training creates a concentrated target and may violate regional, contractual or regulatory restrictions on moving data. Privacy-preserving training reduces the exposure but can reduce accuracy. The third is *accuracy against efficiency*. Full automation is fast but unforgiving of error, and full manual review is accurate but slow, so the question is how to combine them.

This paper proposes a framework that treats these tensions explicitly. The primary objective is to design and examine the Governed AI Classification Framework (GACF). The specific tasks are: (i) to define a governed taxonomy and an error model suited to governance decisions; (ii) to specify a privacy-preserving training architecture that keeps documents where they are; (iii) to define confidence-based triage that combines automation with human review; (iv) to define an audit structure covering governance, model and application levels; and (v) to examine the trade-offs in a controlled simulation. The contribution is a coherent, testable blueprint together with explicit statements about what the evidence does and does not show.

The remainder of the paper is organized as follows. Section II reviews the literature. Section III summarizes related work. Section IV presents the GACF methodology. Section V details the experimental setup, Section VI reports results, Section VII discusses implications, Section VIII addresses limitations, and Section IX concludes.

LITERATURE REVIEW

A. Classification in information governance

Classification is a requirement in most security and governance standards. ISO/IEC 27001 includes controls for classifying and labeling information so that it receives protection appropriate to its sensitivity [1], and FIPS 199 provides a model for categorizing information by the potential impact of loss of confidentiality, integrity and availability [2]. Privacy regulation adds categories whose definition is legal and not organizational: the General Data Protection Regulation identifies special categories of personal data with stricter conditions for processing, and the California Consumer Privacy Act defines sensitive personal information and consumer rights over it [3], [4]. In practice, enterprises adopt a small number of levels, for example public, internal, confidential and restricted, and attach rules to each level for access, sharing, encryption and retention. The taxonomy is therefore not a technical artifact; it is a policy instrument whose levels must be defined by the business and the legal function.

B. Machine learning for text classification and sensitive data discovery

Automated text categorization has a long history. Sebastiani surveyed methods from rule-based systems to statistical learning [5], and Joachims showed that linear support vector machines perform strongly on high-dimensional sparse text [6]. Pre-trained transformer models such as BERT have since improved performance by capturing context, at greater computational cost and with greater difficulty in explaining outputs [7]. For governance use, several properties matter beyond accuracy. Classes are imbalanced, because the most sensitive documents are rare, and evaluation must account for this [8]. Confidence scores must be calibrated if they are to be used for routing decisions, and modern neural networks are often overconfident [9]. Decisions should be explainable so that data owners can contest them. And behavior should be monitored for drift, because documents, projects and vocabulary change over time [10].

C. Privacy-preserving learning

Training a classifier on sensitive documents creates risks: the model can memorize training data, and an adversary may be able to determine whether a specific record was used in training, as Shokri and colleagues demonstrated with membership inference attacks [11]. Three families of technique reduce these risks. *Federated learning* trains a shared model by sending the model to the data and aggregating updates, so that raw data stay on local systems [12], [13]. *Differential privacy* limits what any output reveals about an individual record, and in training is implemented by clipping the influence of each example and adding calibrated noise to updates [14], [15]. *Secure aggregation* lets a server combine updates without seeing any single participant's contribution [16]. The combination, shown in Figure 1, is standard: each client preprocesses its data, trains locally, clips and perturbs its update, and sends it in encrypted form to a server that aggregates, may screen for anomalous updates, and broadcasts the improved global model.

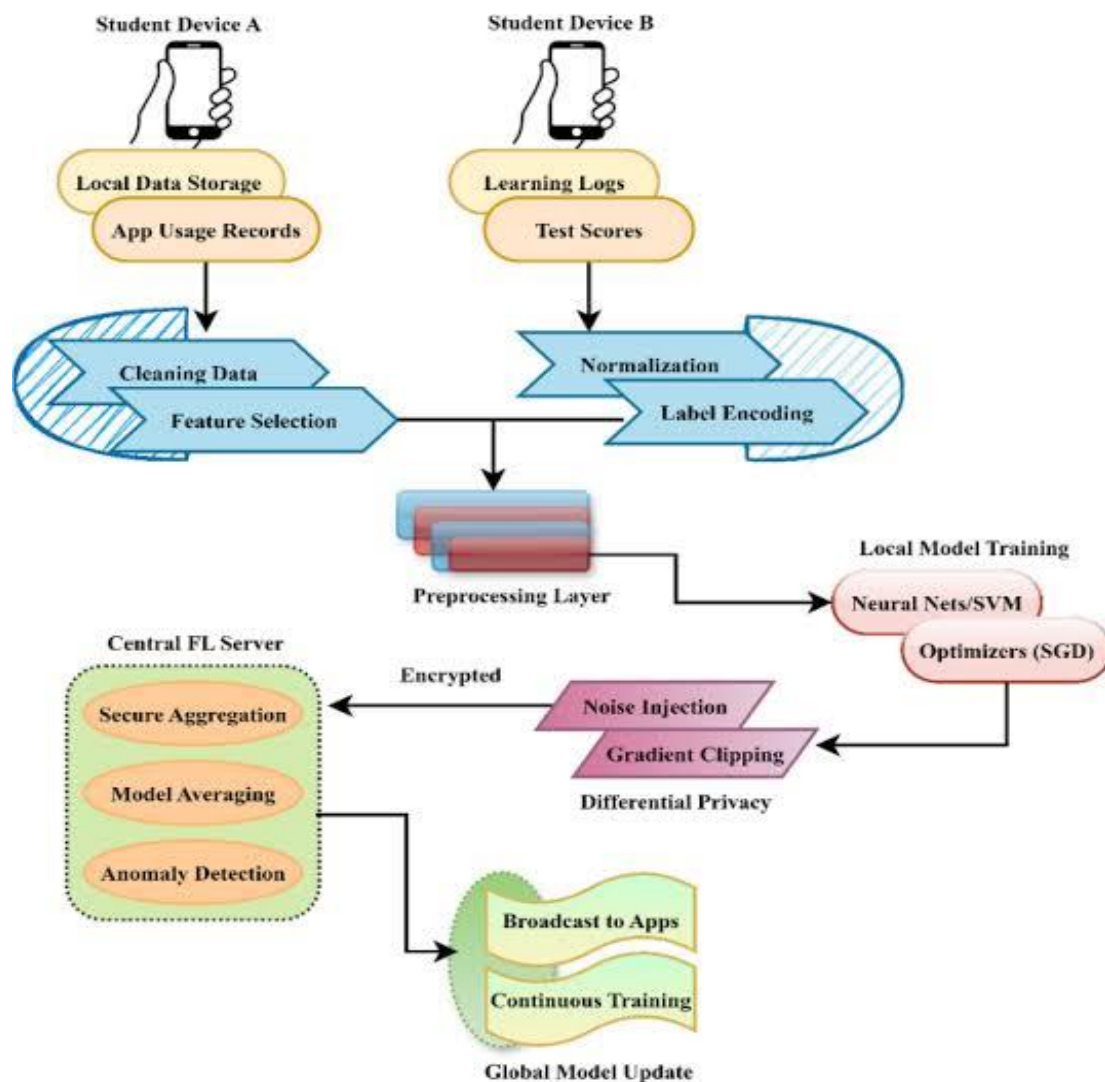


Figure 1. Privacy-preserving federated training architecture: local data and preprocessing on each client, local training, differential privacy (gradient clipping and noise injection), encrypted upload, secure aggregation, model averaging and anomaly detection on the central server, and broadcast of the global model update. The figure illustrates the pattern with student devices; in the enterprise setting the clients are business units, regional environments or tenant workloads.

Federated learning faces a specific challenge in enterprises: data are heterogeneous across units, since different business units produce different documents with different vocabularies and different class balances. Such non-identically distributed data can slow convergence and reduce accuracy, and methods that stabilize training under heterogeneity have been proposed [17]. Privacy protection costs accuracy, and the cost depends on the amount of noise, the number of examples and the difficulty of the task.

D. AI governance and auditing

Frameworks for managing AI risk emphasize governance, measurement and continuous oversight. The NIST AI Risk Management Framework organizes activities into governing, mapping, measuring and managing [18], and the European Union's AI Act sets obligations for risk management, data governance, transparency, human oversight and accuracy for AI systems according to risk level [19]. Documentation practices such as model cards communicate intended use, performance and limitations [20], and internal algorithmic auditing frameworks describe how organizations can examine AI systems throughout their lifecycle [21]. A convenient structure, shown in Figure 2, distinguishes three levels of audit. *Governance audits* examine regulatory compliance, data governance, ethical guidelines and risk management. *Model audits* examine explainability and interpretability, bias and fairness, robustness and security, and generalization and accuracy. *Application audits* examine user

interaction and safety, data security and privacy, performance monitoring and logging, and adversarial testing in the setting where the system is used.

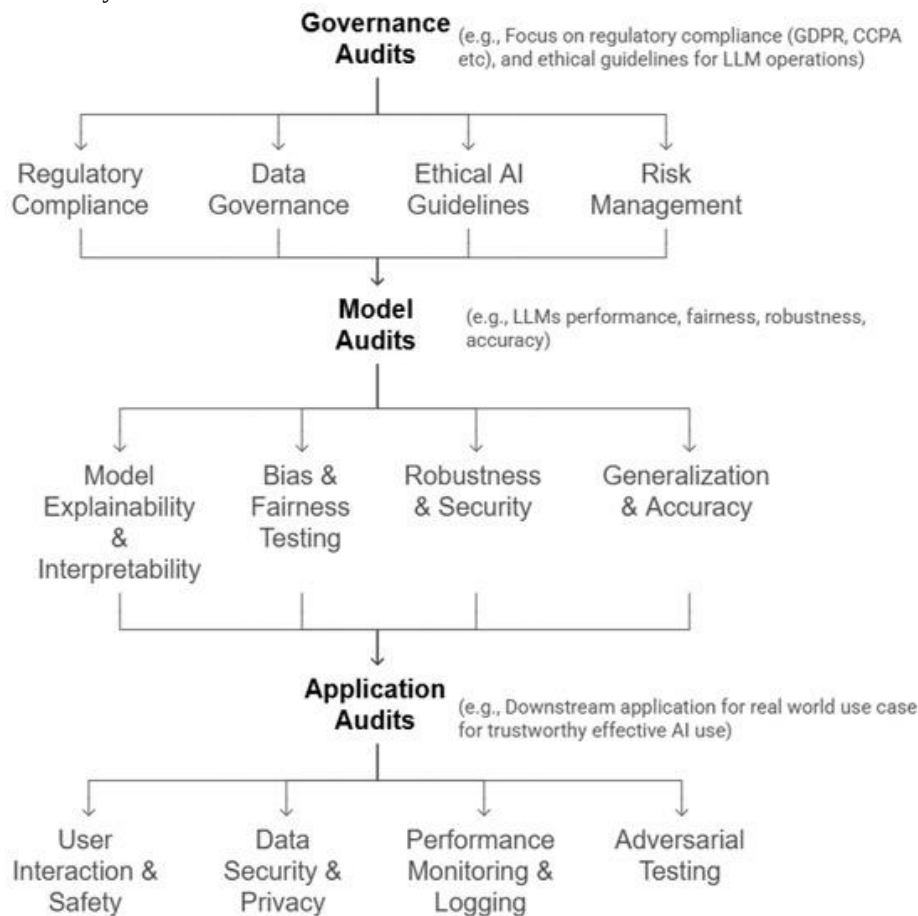


Figure 2. Three-level audit structure: governance audits (regulatory compliance, data governance, ethical AI guidelines, risk management), model audits (explainability, bias and fairness, robustness and security, generalization and accuracy) and application audits (user interaction and safety, data security and privacy, monitoring and logging, adversarial testing).

E. Human oversight and selective automation

Parasuraman and colleagues described levels of automation and argued that the appropriate level depends on the consequences of error and on the human's ability to monitor and intervene [22]. For classification, selective automation is a natural fit. A classifier makes confident decisions on clear cases and defers uncertain ones to a person, a strategy that formalizes the reject option studied by Chow in pattern recognition [23]. The performance of such a hybrid depends on the calibration of confidence, on the accuracy of the human reviewers, and on whether the reviewers are given context that helps them decide.

RELATED WORK

Prior work addresses parts of the problem, and few contributions combine them.

In data loss prevention and discovery, commercial and open source tools combine patterns, dictionaries and fingerprinting to detect known identifiers and known documents. These approaches offer precision for structured data, and they provide little help with contextual sensitivity. They also operate where data are collected, which typically means moving or indexing content centrally.

In machine learning for text, a large literature addresses classification accuracy [5], [6], [7]. Studies typically assume that the training data can be pooled, and they rarely evaluate asymmetric error costs, which are central to governance.

In privacy-preserving machine learning, federated learning, differential privacy and secure aggregation are well studied [12]–[16], including under heterogeneity [17]. Much of this work evaluates image and generic text benchmarks. The trade-off between privacy noise and the kind of error that matters in governance has received less attention.

In AI governance, frameworks and regulation define what organizations must do to manage AI risk [18], [19], [21], but they are general and do not specify how a particular system, such as a classifier of document sensitivity, should be structured and tested.

In human-automation interaction, levels of automation and reject-option classification provide the conceptual basis for triage [22], [23]. Their application to enterprise classification, with measured reviewer workload, is less common.

The gap that GACF addresses is therefore integration. No widely adopted approach (a) connects a governed taxonomy to an asymmetric error model, (b) trains classifiers without pooling raw documents and quantifies the effect of privacy noise on the errors that matter, (c) uses calibrated confidence to combine automation with human review, and (d) places the whole system within a structured audit.

METHODOLOGY: THE GACF FRAMEWORK

A. Architecture overview

GACF has six components. *Policy and taxonomy* defines sensitivity levels and the rules attached to them. The *data and feature pipeline* prepares documents locally. *Training* builds the classification model, either centrally or, preferably, through the federated architecture of Figure 1. *Privacy controls* limit what training reveals. *Decision and triage* converts model outputs into labels, with uncertain cases routed to reviewers. *Audit and monitoring* examine the system at the governance, model and application levels, as in Figure 2, and feed corrections back to the other components.

B. Governed taxonomy and error model

The taxonomy has a small number of ordered levels. In this paper we use four: public, internal, confidential and restricted. Each level is defined by the business and legal functions with examples, and each carries rules for access, sharing, encryption and retention. The number of levels is kept small because every additional level reduces labeling consistency among both people and models.

Because the levels are ordered, errors have direction. *Under-classification* means assigning a lower level than the document deserves, which exposes the document to weaker controls. *Over-classification* means assigning a higher level, which restricts access unnecessarily and imposes cost and friction. Governance treats these differently: under-classification is a risk event, while over-classification is an efficiency cost, up to the point where excessive restriction drives users to work around controls. GACF therefore reports both rates separately, alongside overall accuracy and the macro-averaged F1-score, which weights all classes equally and so reflects performance on rare, high-sensitivity classes. Documents that fall under legal definitions, such as special categories of personal data, may be handled by mandatory rules that override the model, with the model supplying evidence and not the final decision.

C. Data and feature pipeline

Documents are processed where they reside. The pipeline extracts text, removes boilerplate and normalizes the content, then converts it to numeric features, for example weighted word counts or embeddings. Cleaning and feature selection happen locally in each client environment, as shown in the preprocessing layer of Figure 1, so that raw text never leaves it. Labels for training come from existing labelled documents, from the outcomes of human reviews, and from curated examples supplied by records managers. Because labels are themselves noisy, the pipeline records label provenance and periodically audits samples.

D. Privacy-preserving training

GACF recommends federated training. Each business unit trains a local model on its documents for a small number of passes and sends the update to an aggregation service, which averages the updates into a global model and broadcasts it, in repeated rounds. Two privacy controls are applied to each update before it leaves the unit.

Gradient clipping bounds the influence that any single document can have on the update. *Noise injection* adds random noise, scaled to the clipping bound by a noise multiplier chosen by the privacy owner, which makes it difficult to infer whether a particular document was in the training data. Updates are transmitted in encrypted form, and secure aggregation prevents the server from seeing any single unit's contribution. The server may also apply anomaly detection to updates, to detect a unit whose update deviates strongly from the others, which may indicate a data quality problem or an attempt to poison the model.

The amount of noise determines the privacy-accuracy trade-off. In a deployment, the noise level should be linked to a formal privacy budget computed with a privacy accountant, and the budget should be a governance decision. In this paper, the experiments report the noise multiplier and an empirical measure of leakage, as explained below, and they do not claim a formal privacy guarantee.

E. Decision and triage

The model outputs a probability for each sensitivity level. The decision component assigns the most probable level and uses the highest probability as a confidence measure. Documents whose confidence exceeds a threshold are labelled automatically, and the rest are routed to a human reviewer, who sees the document, the model's suggestion and the evidence behind it. The threshold is set by the data owner according to the cost of review and the cost of error, and it may differ by repository: a repository holding highly sensitive material may adopt a higher threshold. Reviewer decisions are recorded and used as training labels, so that the model improves where it was least certain. Confidence calibration is checked regularly, since triage is only as good as the calibration of the confidence score [9].

F. Audit and monitoring

The system is audited at the three levels shown in Figure 2. At the *governance* level, auditors confirm that the taxonomy is approved, that regulatory requirements are mapped to rules, that data used for training is authorized, and that risk is assessed and reviewed. At the *model* level, they examine explainability, test fairness across business units and document types, test robustness to perturbation and attack, and measure accuracy and generalization on held-out data from each unit. At the *application* level, they examine the user experience of labels and overrides, the privacy and security of the pipeline, the logging and monitoring of operation, and resistance to adversarial inputs. Findings are tracked to closure.

Several indicators are monitored continuously: accuracy and macro F1 by unit, the under-classification and over-classification rates, the share of documents routed to review, the agreement between reviewers and model, drift in vocabulary and class balance, and the volume of user overrides. A fairness check compares performance across units, so that a unit with unusual documents is not systematically misclassified.

G. Governance and operating model

GACF assigns roles. A *data owner* in each unit confirms taxonomy application and thresholds. A *classification policy owner* maintains the taxonomy and rules. A *privacy owner* sets noise levels and approves data use for training. A *model owner* trains, evaluates and monitors the model. *Reviewers* handle routed documents. An *audit function* conducts the three-level audits. A *governance board* reviews indicators and exceptions at least quarterly.

EXPERIMENTAL SETUP

The experiments are limited in purpose. Real enterprise documents are confidential and cannot be shared, labelled data sets with realistic business-unit heterogeneity are scarce, and the true sensitivity of documents is itself contested. We therefore used a synthetic enterprise implemented in Python with a fixed random seed, to examine the framework's trade-offs. Each result is the mean of eight independent replications.

Synthetic documents. Each replication simulates five business units. Documents are bags of words over a vocabulary of 1,000 terms, with four sensitivity classes (public, internal, confidential, restricted). The terms are divided into generic words shared by all documents, class-indicative words for each sensitivity level, identifier-like words that are more common in confidential and restricted documents, and dialect words specific to each business unit. A quarter of the class-indicative words in a document are drawn from a neighboring class, to represent ambiguity, and three percent of labels are randomly corrupted, to represent labeling error. Class proportions differ across units, so that the data are heterogeneous. Each unit has 260 training documents and 500 test documents.

Compared approaches. Eight approaches are compared. *Rule-based keywords* uses a dictionary of known indicative words for each level, with precedence for higher levels, the default being internal. *Centralized training* pools all units' raw documents in one place and trains a softmax regression classifier, serving as an accuracy ceiling that ignores privacy. *Local models only* trains a separate model in each unit and does not share anything. *Federated averaging without noise* trains the same model through 30 rounds of local training and averaging. *Federated averaging with privacy noise* adds gradient clipping and noise with noise multipliers of 0.5, 1, 2 and 4.

Metrics. We report accuracy, macro-averaged F1, the under-classification rate, the over-classification rate, the accuracy of the worst-performing unit, and an empirical measure of privacy leakage. The leakage measure is the success of a simple membership inference attack, which tries to tell training documents from fresh documents by the model's loss, expressed as the area under the curve, where 0.5 means that the attacker does no better than chance. This measure does not replace a formal privacy guarantee.

Triage experiment. Using the federated model without noise, documents are labelled automatically if the model's top probability exceeds a threshold of 0.5, 0.6, 0.7 or 0.8, and otherwise routed to a human reviewer who is assumed to be correct in 95% of cases. We report the share of documents classified automatically, the overall accuracy of the hybrid process, the accuracy of automated decisions and the under-classification rate among automated decisions. Reviewer effort is estimated by assuming 90 seconds per manually reviewed document.

All assumptions, including the generating process, the human accuracy and the review time, are inputs of the simulation and not measurements.

RESULTS

A. Accuracy, errors and fairness across approaches

Table 1 reports the main results. The centralized model achieved the highest accuracy (96.4%) and macro F1 (93.8%), with an under-classification rate of 2.7%. The federated model without noise came close, with 94.4% accuracy and 88.5% macro F1, and an under-classification rate of 4.8%, while no raw document left its business unit. The rule-based approach reached 74.6% accuracy and made many errors in both directions (10.0% under-classification and 15.4% over-classification). Local models, trained without sharing, reached 86.8% accuracy on average, but their macro F1 was only 73.8% and the under-classification rate was 11.1%, because with few documents per unit they learned the rare classes poorly.

TABLE 1. Simulated Classification Performance (mean of 8 replications; 2,500 test documents each)

Approach	Accuracy (%)	Macro F1 (%)	Under-classification (%)	Over-classification (%)	Worst-unit accuracy (%)	Membership-inference AUC
Rule-based keywords	74.6	68.7	10.0	15.4	66.2	0.500
Centralized (pooled raw data)	96.4	93.8	2.7	0.9	95.5	0.536
Local models only	86.8	73.8	11.1	2.1	77.9	0.672
Federated, no noise	94.4	88.5	4.8	0.8	92.8	0.526
Federated, noise multiplier 0.5	92.9	86.4	6.1	1.0	91.1	0.524
Federated, noise multiplier 1	88.5	80.8	8.9	2.6	85.5	0.523
Federated, noise multiplier 2	76.9	66.1	15.5	7.6	71.6	0.513

Federated, noise multiplier 4	62.4	51.7	21.6	16.1	54.4	0.512
-------------------------------------	------	------	------	------	------	-------

Two further observations follow from the table. Local models were the most leaky in the empirical sense, with a membership inference score of 0.67 against 0.53 for the centralized and federated models, because small local training sets are easier to memorize. Federated training therefore improved both accuracy and empirical privacy relative to training in isolation. And the worst-performing unit mattered: local models had a worst-unit accuracy of 77.9%, while the federated model's was 92.8%, indicating that sharing knowledge through federation helped the units with the least data and the least typical documents most.

B. The cost of privacy noise

Figure 3a shows how accuracy changes with the noise multiplier. Low noise had a modest effect: at a multiplier of 0.5, accuracy fell from 94.4% to 92.9%. At 1 it fell to 88.5%, and under-classification rose from 4.8% to 8.9%. At 2 and 4, performance deteriorated sharply, to 76.9% and 62.4% accuracy, with under-classification of 15.5% and 21.6% and over-classification of 7.6% and 16.1%. The empirical leakage measure moved little in this range, from 0.526 without noise to 0.512 at the highest noise, because the non-private federated model was already close to chance on this measure.

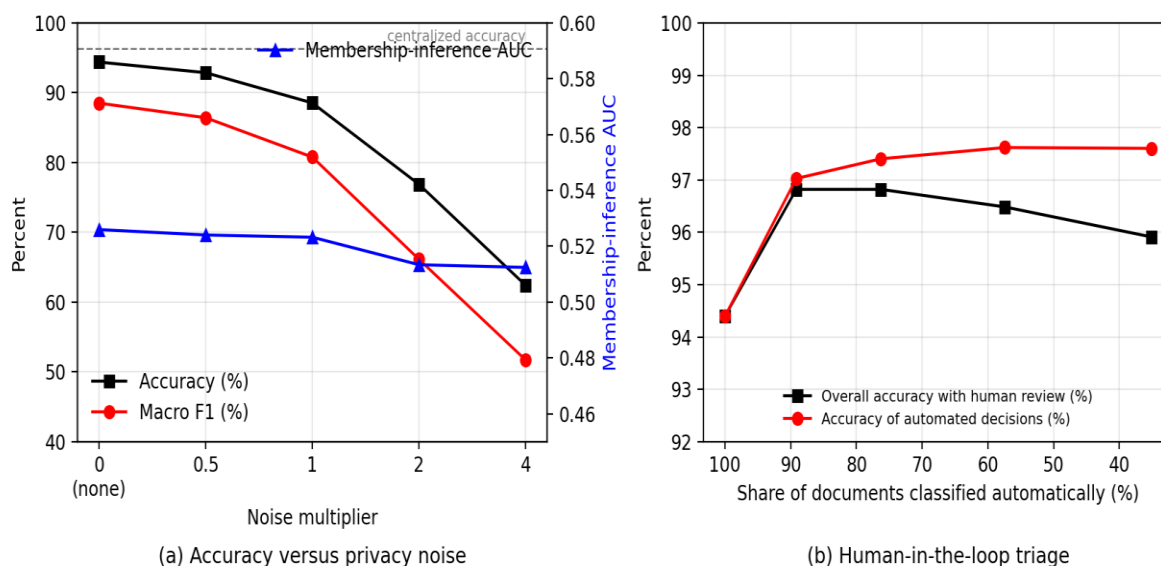


Figure 3. (a) Accuracy, macro F1 and membership-inference AUC for federated training as the noise multiplier increases. (b) Overall accuracy of the hybrid process and accuracy of automated decisions as the share of automatically classified documents falls.

The result implies that the choice of noise level is a governance decision with a concrete price. In this setting, a moderate noise level preserves most of the accuracy, while a high level damages the property that matters most, the under-classification rate. It also shows the limits of an empirical leakage measure: it is a useful sanity check but cannot distinguish noise levels once leakage is already low, and it does not substitute for a formal privacy budget.

C. Triage and operational efficiency

Table 2 and Figure 3b report the triage experiment. With no routing, the federated model's accuracy was 94.4%. Routing the least confident 10.8% of documents to reviewers (threshold 0.5) raised overall accuracy to 96.8%, and automated decisions were 97.0% accurate. Raising the threshold to 0.6 routed 23.7% of documents and left overall accuracy at 96.8%, with an under-classification rate of 2.1% among automated decisions. Higher thresholds brought little further benefit in accuracy: at 0.8, 64.9% of documents went to reviewers and overall accuracy was 95.9%, lower than at 0.6, because reviewers, assumed to be 95% accurate, were themselves less accurate than the model's confident decisions of about 97.6%.

TABLE 2. Simulated Triage Results and Reviewer Effort (10,000 documents)

Confidence threshold	Classified automatically (%)	Overall accuracy (%)	Accuracy of automated decisions (%)	Under-classification among automated decisions (%)	Reviewer effort (hours)
No review (0.0)	100.0	94.4	94.4	4.8	0
0.5	89.2	96.8	97.0	2.4	27
0.6	76.3	96.8	97.4	2.1	59
0.7	57.5	96.5	97.6	2.1	106
0.8	35.1	95.9	97.6	2.3	162
Fully manual	0.0	95.0	not applicable	not applicable	250

Reviewer effort follows directly from the share routed. Full manual review of 10,000 documents at 90 seconds each would take 250 hours. At a threshold of 0.6, review covers 23.7% of documents and takes about 59 hours, a reduction of about 76%, while overall accuracy (96.8%) is higher than both the unaided model (94.4%) and fully manual review (95.0%). The hybrid exceeds both because it uses each where it is better: the model on clear cases and the human on ambiguous ones.

D. Summary of findings

Four findings stand out. First, federated training reached accuracy close to centralized training without pooling raw documents, and clearly better than rules or isolated local models. Second, privacy noise has a graded cost that becomes severe at high levels, and the cost falls disproportionately on the dangerous error of under-classification. Third, confidence-based triage improved accuracy and cut reviewer effort by about three quarters at a moderate threshold. Fourth, the worst-performing unit is a more informative indicator than the average, since federation helped the weakest units most.

DISCUSSION

The simulation supports the central argument of the framework: accuracy, privacy and efficiency are in tension, and the tension can be managed by design and not merely endured. Federated training moves the privacy-accuracy frontier relative to the two obvious alternatives, centralizing everything or keeping everything local. Triage moves the accuracy-efficiency frontier relative to full automation and full manual review. And the audit structure ensures that the choices made at each point are visible to the people accountable for them.

The results also demonstrate why error direction matters. A comparison on overall accuracy alone would suggest that the federated model with a noise multiplier of 1 (88.5%) is a modest sacrifice compared with the model without noise (94.4%). The under-classification rate, however, almost doubled, from 4.8% to 8.9%, which means that nearly one in eleven documents received weaker protection than it needed. Whether that is acceptable is a risk decision. For some repositories, it may be, and for others it may not, which is why GACF sets thresholds and noise levels per repository and not globally.

For practitioners, five recommendations follow. Define the taxonomy and the asymmetry of its errors before choosing a model. Prefer training approaches that keep documents where they are, and treat the noise level as a documented decision with a formal privacy budget in deployment. Use calibrated confidence to route uncertain documents to reviewers, and set thresholds by repository. Measure the worst-performing unit, and not only the average, to detect inequity across units. And audit at all three levels, since governance, model and application failures are different and are found by different methods.

The framework also has implications for compliance and assurance. Regulators and customers increasingly ask how automated systems that affect data protection are governed. A documented taxonomy, a record of how training data were handled, evidence of calibrated confidence and human oversight, and a trail of audits against the three levels provide that evidence, and they map naturally to the risk management and oversight expectations of AI governance frameworks [18], [19].

Finally, the framework is conservative about automation. Where a classification has legal consequences, as with special categories of personal data, deterministic rules and human decision remain in place, and the model supplies

evidence. The aim is not to replace the judgment of data owners and records managers but to reserve it for the cases that need it.

LIMITATIONS AND FUTURE WORK

The principal limitation is that the evaluation is synthetic. The documents are bags of words generated from stated distributions, with a simple form of heterogeneity across units and a simple form of ambiguity. Real documents are longer, structured, multilingual and context-dependent, and real sensitivity is partly a matter of judgment that experts dispute. The absolute accuracies reported here should not be expected to hold in practice, and the model used, a linear classifier on word counts, is much simpler than the transformer-based models that may be used in practice. The relative patterns, in particular the cost of noise on under-classification and the gain from triage, follow largely from the structure of the design, but they have not been confirmed on real data.

Second, the privacy analysis is incomplete. Noise is reported as a multiplier and leakage is measured with a single, simple attack. A deployment requires a formal privacy accounting, evaluation against stronger attacks, and consideration of what the aggregated model reveals about the documents of any unit. Third, the triage experiment assumes a fixed and independent reviewer accuracy and a fixed review time, whereas real reviewers vary, tire, and are influenced by the model's suggestion, which can induce automation bias. Fourth, the experiments did not examine drift, adversarial inputs, poisoning of updates or the effect of unit-level fairness constraints.

Future work should therefore include: (i) evaluation on real, access-controlled document collections with expert labels; (ii) comparison with transformer-based models and with embedding-based approaches under federation; (iii) formal privacy accounting and evaluation against stronger inference attacks; (iv) studies of reviewer behavior, including automation bias, and of the design of review interfaces; and (v) longitudinal studies of drift, retraining and the effect of audit findings on performance.

CONCLUSION

This paper presented GACF, a framework for AI-assisted data classification in enterprise governance that balances accuracy, privacy and operational efficiency. It combines a governed taxonomy with an explicit error model, privacy-preserving federated training with gradient clipping and noise injection, confidence-based triage with human review, and a three-level audit structure covering governance, model and application. In a controlled simulation, federated training reached 94.4% accuracy against 96.4% for centralized training and 74.6% for rule-based keywords; privacy noise reduced accuracy gradually at moderate levels and sharply at high levels, with under-classification rising from 4.8% to 21.6% at the highest noise; and confidence-based triage raised overall accuracy to 96.8% while classifying 76% of documents automatically and cutting reviewer effort by about three quarters.

The simulation is evidence of coherence and not of field effectiveness. The next step is empirical: pilot the framework on real documents in a small number of business units, calibrate thresholds and noise levels against measured error costs, and audit the result. If the findings are consistent with the logic shown here, organizations can classify at scale without concentrating their most sensitive documents, and can direct scarce human attention to the cases that need it.

REFERENCES

1. **Pawan Kumar Singh**, Scaling Gate-All-Around (GAA) Transistor Architectures for Sub-2nm AI Accelerators Using Atomically Thin 2D Materials, Vol. 33 No. 08 (2024): JOCAAA, Journal Name: Journal of Computational Analysis and Applications, Journal's ISSN: 1521-1398 (Paper),1572-9206 (Online), <https://eudoxuspress.com/index.php/pub/article/view/5713>
2. **Pawan Kumar Singh**, Energy-Efficient Task Scheduling in Green Cloud Computing Using Neuromorphic Spiking Neural Networks, Vol. 33 No. 08 (2024): JOCAAA, Journal Name: Journal of Computational Analysis and Applications, Journal's ISSN: 1521-1398 (Paper),1572-9206 (Online), <https://eudoxuspress.com/index.php/pub/article/view/5712>

3. **Rangavinay Teja Royal Jagga**, 2026, DOI: 10.1109/GCWCN66157.2025.11448495, Date of Conference: 22-23 November 2025, Date Added to IEEE Xplore: 25 March 2026 ,<https://ieeexplore.ieee.org/document/11448495>
4. **Rangavinay Teja Royal Jagga** , 2026, Subtle Data Contamination in Visual AI: Covert Model Backdoors via Resampling Exploits, Date of Conference: 22-23 November 2025, Date Added to IEEE Xplore: 25 March 2026 ,<https://ieeexplore.ieee.org/document/11448514>
5. **Rangavinay Teja Royal Jagga**, 2026, Robotic Tactile-Vision Integration for Interactive Object Disentanglement, <https://ieeexplore.ieee.org/document/11448337>
6. **Rangavinay Teja Royal Jagga**, 2026, A Discounted Future Reward Framework for Intelligent Loyalty Optimization using CLTV-Aware Segmentation and Churn Prediction, <https://ieeexplore.ieee.org/document/11497560>
7. **Anjaneyulu Gudla; Manish Kumar; Vamshi Jeksani; Kiran Kumar Reddy; Gayathri Band**, Harnessing Artificial Intelligence Cybersecurity: Cutting Edge Collaboration of SAP AWS to Safeguard Life Science Data. 2026 <https://doi.org/10.1109/aiei69164.2026.11497833>
8. **G. Mahender, Department of CSE(Data Science), Vardhaman College of Engineering ; Vamshi Jeksani; Manish Kumar; Koppuravuri Sri Lakshmi Sruthi; Kiran Kumar Reddy**, SAP Cloud System on AWS for Risk Detection and Integration Artificial Intelligence Scalable Cybersecurity 2026 <https://doi.org/10.1109/aiei69164.2026.11497435>
9. **Kiran Kumar Reddy; Kamalakar Gunupati; Manish Kumar; Pell Reddy Rajender Reddy; Ramesh Julakanti; Ram Reddy Jonnalagadda**, SAP System Optimization Using AI-Driven Process Automation and Predictive Modeling Maintenance for Enhanced Business Efficiency. 2025 <https://doi.org/10.1109/computingcon64838.2025.11376613>
10. **Amilineni, M.V. 2025**. Audit-Safe Agentic RAG Framework for Regulated Enterprise Systems: A Trustworthy AI Architecture for SAP, Finance, Healthcare, and Government Workflows. INTERNATIONAL JOURNAL OF ADVANCES IN SIGNAL AND IMAGE SCIENCES. (Jan. 2025), 34–46. DOI: <https://doi.org/10.29284/7fjzk883>
11. **Pawan Kumar Singh**, ACCELERATING LARGE LANGUAGE MODEL TRAINING USING HYBRID QUANTUM-CLASSICAL VARIATIONAL AUTOENCODERS WITH INTEGRATED PHOTONIC QRAM, Vol. 54 No. 3 (2026): July-September 2026, Power System Protection and Control, ISSN-1674-3415, <https://pspac.info/index.php/dlbh/article/view/501>, DOI: <https://doi.org/10.46121/pspc.54.3.45>
12. **Pawan Kumar Singh** , RESILIENCE ASSESSMENT OF MODERN POWER GRID INFRASTRUCTURE AGAINST EXTREME WEATHER USING PHYSICS-INFORMED NEURAL NETWORKS, Vol. 54 No. 1 (2026): January-March 2026, Power System Protection and Control, ISSN-1674-3415,<https://pspac.info/index.php/dlbh/article/view/502>, DOI: <https://doi.org/10.46121/pspc.54.1.64>
13. **Pawan Kumar Singh**, MITIGATING THE EXABYTE STORAGE CRISIS: ENHANCING READ/WRITE EFFICIENCY IN 5D OPTICAL GLASS STORAGE USING SILVER NANOPARTICLE DOPING, [Vol. 38 No. 11s \(2025\)](#), International Journal of Applied Mathematics (IJAM) , pISSN 1311-1728, eISSN 1314-8060 ,<https://ijamjournal.org/ijam/publication/index.php/ijam/article/view/2153>
14. **Mr. Pawan Kumar Singh (Author), Miss. Pari Nidhi Singh (Author)** ,Intelligent Bot, ASIN : B0DBGCQY8X,Publisher : Namya Press, Publication date : November 25, 2022,Language : English, Print length : 105 pages,ISBN-10 : 9390124867,ISBN-13 : 978-9390124862,Item Weight : 7.5 ounces, Dimensions : 6.24 x 0.43 x 9.24

inches,https://www.amazon.com/Intelligent-Bot-Pawan-Kumar-Singh/dp/B0DBGCQY8X/ref=sr_1_5?crid=2VAXN5BRN5SMS&dib=eyJ2IjoiMSJ9.v3OCQvHBpy20RIjfmWh0qbOl-RLOIUJIC8Xa0Pjeqilvo2sbZ_suN3LbSCw_OA_ZD_4lgjRLMGIBwZQQ1ehZEfDpt_MJiOUfVuqqz_Mav0I5_mb6pN5DQog0U7SA826n1DQQ3QAtWwopHy4lC5lfhZXttKj9faiyATwhdtDATSSV1-r7FmzM2VQyIA0jxcfq6E6hp9ECfH1fby0li9kUfGpLxMq4GktlzUIJO5ggrhsQ.G1AjQiVqU4HFsnScz_vz-X6NXCNs3EFSn-RI8mcWWGOc&dib_tag=se&keywords=Pawan+Kumar+Singh%2C+book&qid=1787815741&sprefix=pawan+kumar+singh%2C+bo%2Caps%2C450&sr=8-5

15. **15: Mr. Pawan Kumar Singh (Author)**, STORAGE CRISES DUE TO RISE OF AI: A Comprehensive Guide to Managing the Storage Challenges in the Age of Artificial Intelligence, ASIN : B0HFK12ZCT Publisher : EB1 Profile, Accessibility : Learn more, Publication date : August 17, 2026, Edition : 1st Language : English, File size : 1.7 MB, Screen Reader : Supported, Enhanced typesetting : Enabled X-Ray : Not Enabled, Word Wise : Not Enabled , Print length : 261 pages, ISBN-13 : 979-8951479112, Page Flip : Enabled, https://www.amazon.com/STORAGE-CRISES-DUE-RISE-Comprehensive-ebook/dp/B0HFK12ZCT/ref=sr_1_1?crid=9MF6C4WDVJR5&dib=eyJ2IjoiMSJ9.DQ8fesBch7ELmRniiDl-YA.NBSCHLmetvnbhkwIpOoj3s0IVsCV8y6-ZCfCP6U1lc&dib_tag=se&keywords=Pawan+Kumar+Singh%2C+book%2C+eb1+Profile&qid=1787816040&sprefix=pawan+kumar+singh%2C+book%2C+eb1+profile%2Caps%2C420&sr=8-1
16. **Parag Bharatkumar Parikh**, FRAMEWORKS FOR MITIGATING SUPPLY CHAIN DISRUPTIONS AND MATERIAL PRICE VOLATILITY IN SUSTAINABLE BUILDING CONSTRUCTION, Vol. 53 No. 2 (2025): April-June 2025 , , Power System Protection and Control, ISSN-1674-3415, <https://pspac.info/index.php/dlbh/article/view/447>, DOI: <https://doi.org/10.46121/pspc.53.2.34>
17. **Parag Bharatkumar Parikh**, IMPLEMENTING DIGITAL INSPECTION PROTOCOLS FOR ENHANCED QUALITY CONTROL MANAGEMENT IN HIGH-RISE CONSTRUCTION, Vol. 54 No. 3 (2026): July-September 2026, Power System Protection and Control, ISSN-1674-3415, <https://pspac.info/index.php/dlbh/article/view/439>, DOI: <https://doi.org/10.46121/pspc.54.3.12>
18. **Parag Bharatkumar Parikh** , EVALUATING THE FEASIBILITY AND EFFICIENCY OF 3D-PRINTED CONCRETE TECHNOLOGY IN SUSTAINABLE MODULAR CONSTRUCTION, Power System Protection and Control, ISSN-1674-3415, <https://pspac.info/index.php/dlbh/issue/view/30>, DOI: <https://doi.org/10.46121/pspc.54.1.61>
19. **Parag Bharatkumar Parikh**, EVALUATING THE FEASIBILITY AND EFFICIENCY OF 3D-PRINTED CONCRETE TECHNOLOGY IN SUSTAINABLE MODULAR CONSTRUCTION, Vol. 54 No. 1 (2026): January-March 2026, System Protection and Control, ISSN-1674-3415, <https://pspac.info/index.php/dlbh/article/view/440>, DOI: <https://doi.org/10.46121/pspc.54.1.61>
20. **Ronakkumar shah**, REDEFINING BUSINESS PROCESS MANAGEMENT: THE SYNERGY OF AI AND INTELLIGENT AUTOMATION, Vol. 54 No. 3 (2026): July-September 2026, System Protection and Control, ISSN-1674-3415, <https://pspac.info/index.php/dlbh/article/view/448>, DOI: <https://doi.org/10.46121/pspc.54.3.14>
21. **Ronakkumar shah**, OPERATIONAL ACCOUNTABILITY IN ACTION: METRICS AND MECHANISMS FOR SYSTEMIC SUCCESS, Vol. 53 No. 4 (2025): October-December 2025, System Protection and Control, ISSN-1674-3415, <https://pspac.info/index.php/dlbh/article/view/452>, DOI: <https://doi.org/10.46121/pspc.53.4.40>

22. **Ronakkumar shah**, BEYOND THE BLUEPRINT: AUTOMATED WORKFLOW ENFORCEMENT AND POLICY GOVERNANCE, Vol. 53 No. 2 (2025): April-June 2025 , System Protection and Control, ISSN-1674-3415 ,<https://pspac.info/index.php/dlbh/article/view/451>, DOI: <https://doi.org/10.46121/pspc.53.2.35>
23. **Tushita Agarwal; Swaminathan Sethuraman**, Mapping Customer Behavior: Visual Analysis of Online Retail Interactions, Date Added to IEEE Xplore: 17 February 2026, DOI: <https://doi.org/10.1109/IC2NC67409.2025.11376261>
24. **Tushita Agarwal**, Efficient Data Summarization and Comparison via Grid-Based Statistical Proxies for Large-Scale Datasets, **Date Added to IEEE Xplore: 03 December 2025**, DOI: <https://doi.org/10.1109/ICBATS66542.2025.11258564>
25. **Dip Bharatbhai Patel**, Building scalable business intelligence systems in the cloud: A technical approach, International Journal of Science and Research Archive, 2021, 02(01), 212-215., Received on 03 February 2021; revised on 17 April 2021; accepted on 19 April 2021, DOI: <https://doi.org/10.30574/ijrsra.2021.2.1.0047>
26. **Dip Bharatbhai Patel**, University of North America, Virginia, USA, Home / Archives / Vol. 5 No. 02 (2025): Volume 05 Issue 02 / Articles, <https://www.academicpublishers.org/journals/index.php/ijdsml/article/view/5958>
27. **Sai Akshara Vemuri** , KINEMATIC OPTIMISATION OF SOFT ROBOTIC MANIPULATORS FOR MEDICAL INTERVENTION, Vol. 54 No. 3 (2026): July-September 2026, System Protection and Control, ISSN-1674-3415, <https://pspac.info/index.php/dlbh/article/view/450>
28. **Sai Akshara Vemuri**, <https://www.periodicos.ulbra.org/index.php/acta/article/view/528>
29. **Aziz, A., Chandusha, K., Naga Kiran, B., Prasad, M. L. M., Golla, K., Rajeswari, S., & Katta, S. K. R.** (2026). Exploring Machine Learning Applications for Genomic Data Analysis in Personalized Medicine. International Journal of Drug Delivery Technology, 16(38s), 1166-1173. <https://doi.org/10.25258/ijddt.16.38s.127>, <https://www.periodicos.ulbra.org/index.php/acta/article/view/687>
30. **Mohammad Naffizuddin , Chaganti Ashok ,Sowjanya Gunukula , Rvbs Sarma , Bhavana Sujanamulk , Bharani Krishna Takkella , Chukka Ram Sunil ,Varri Sujana , Ashwini Kumar**, Etiological Factors of Orofacial Clefts in a Hospital-Based Population: A Descriptive Analytical Study, Published: December 10, 2025,DOI: 10.7759/cureus.98868, DOI : <https://doi.org/10.7759/cureus.98868>
31. **Manoj Kumar Kagitha**, Automated Security Compliance in Ci/Cd: A Natural Language Processing (NLP) Approach to Detecting Vulnerabilities in Infrastructure-As-Code (Terraform/Bicep), Publication year 2021, Volume 2021, Issue 12, <https://computerfraudsecurity.com/index.php/journal/article/view/970>
32. **Manoj Kumar Kagitha**, Hybrid Quantum-Classical Resource Scheduling: A Proposed Architecture for Next-Generation Hyperscale Data Centers, Publication year 2024, Volume 2024, Issue 12, <https://computerfraudsecurity.com/index.php/journal/article/view/969>
33. **Manoj Kumar Kagitha**, Self-Healing Cloud Infrastructure: A Hybrid Reinforcement Learning Framework for Predicting Microservice Failures in Azure Kubernetes Service (AKS),Publication year 2023, Volume 2023, Issue 12,<https://computerfraudsecurity.com/index.php/journal/article/view/971>

34. **Manoj Kumar Kagitha**, GREENOPS IN THE CLOUD: AN AI-DRIVEN TELEMETRY ANALYSIS MODEL FOR REDUCING CARBON FOOTPRINT IN HIGH-AVAILABILITY CLUSTERS, Vol. 28 No. 6 (2020): JOCAAA, <https://www.eudoxuspress.com/index.php/pub/article/view/5054>
35. Lakshmi Rahul Reddy Mareddy. (2026). Security and Performance Assessment of Encryption Techniques for Cloud Platforms. International Journal of Communication Networks and Information Security (IJCNIS), 18(1), 68–80. Retrieved from <https://ijcnis.org/index.php/ijcnis/article/view/8735>
36. Lakshmi RahulReddy Mareddy, Context-Aware Learning Approaches for Improving Prediction Accuracy in Dynamic Systems, International Journal of Computational Mathematical Ideas (IJCMI), Vol.17 No.3(2025): IJCM Vol17 NO3 2025 <https://www.ijcmi.com/index.php/journal/article/view/15>
37. Lakshmi Rahul Reddy , Mareddy. (2026). Revisiting Autonomy in Artificial Intelligence through Agentic System Design. Journal of Computational Analysis and Applications (JoCAAA), 35(1), 707–713. <https://eudoxuspress.com/index.php/pub/article/view/4780>
38. **Mahendrakumar Kalal**, Lean-Driven SAP Production Planning Optimization Using Machine Learning for Inventory and Throughput Efficiency, Date of Conference: 19-20 March 2026, Date Added to IEEE Xplore: 08 April 2026, ISBN Information:, ISSN Information: DOI: 10.1109/ISDFS69419.2026.11459029, <https://ieeexplore.ieee.org/abstract/document/11459029>
39. **Mahendra kumar Kalal**, SUPPLY CHAIN RESILIENCE THROUGH INTEGRATED SAP PRODUCTION PLANNING – PROCUREMENT SYNCHRONIZATION, Archives / Vol. 54 No. 1 (2026): January-March 2026 / Articles, <https://pspac.info/index.php/dlbh/article/view/259>
40. **Dr Manoj Goel (December 2018)**. THE ATTITUDE OF DIFFERENT STATES STUDENTS TOWARDS PHYSICAL EDUCATION; STUDYING IN PHYSICAL EDUCATION COLLEGES OF MAHARSHTRA STATE, <https://ijeponline.com/index.php/journal/article/view/965/847>
41. **Ilakiya Ulaganathan, Tagore Engg. College, Anna University, Chennai, India, ; FNU Antara; Santhosh Chitraju Gopal Varma; Zokir Mamadiyarov; Ziyadulla Turayev**, Published in: 2025 International Conference on Sustainability, Innovation & Technology (ICSIT), Date of Conference: 22-23 August 2025, Date Added to IEEE Xplore: 19 December 2025, ISBN Information: <https://ieeexplore.ieee.org/document/11294748/references#references>
42. **Shreyas Kasture, Sudhanshu Maurya, Amit Kumar Sharma, Santhosh Chitraju Gopal Varma, Kashish Mirza, Firdous Sadaf Mohammad Ismail**, DOI: <http://doi.org/10.11591/ijai.v15.i2.pp1746-1761>, IAES International Journal of Artificial Intelligence (IJ-AI), ISSN/e-ISSN 2089-4872/2252-8938 <https://ijai.iaescore.com/index.php/IJAI/article/view/26932>
43. **Shiv Shankar Dwivedi, Mr. Santosh Kumar Sharma**, Future-Proofing Data Security with Quantum-Resistant Cryptography, Vol. 11 No. 3 (2025): IJCESN, DOI: <https://doi.org/10.22399/ijcesn.3447>, <https://ijcesn.com/index.php/ijcesn/article/view/3447>
44. **Vijender Kalmotia**, Enhancing statistical accuracy in gene perturbation studies, Volume 266, August 2026, 105819, <https://doi.org/10.1016/j.biosystems.2026.105819>, <https://www.sciencedirect.com/science/article/pii/S0303264726001292>

45. **Naveen Undrathi**, Design and implementation of generative AI-driven data transformation pipelines using spring boot for scalable financial applications, eISSN: 2581-9615 || CODEN: WJARAI || Impact Factor 8.2 || CrossRef DOI, September 2026 (Volume 31, Issue 3), DOI url: <https://doi.org/10.30574/wjarr.2026.30.3.1629>,
46. **Naveen Undrathi**, Agentic generative AI-enabled microservices framework for intelligent financial data processing in java-based enterprise systems, September 2026 (Volume 31, Issue 3), eISSN: 2581-9615 || CODEN: WJARAI || Impact Factor 8.2 || <https://doi.org/10.30574/wjarr.2026.30.3.1628>
47. **Akshay Jain**, Comparative Analysis of Linkage Methods in Hierarchical Clustering for Customer Segmentation, Conference paper, First Online: 07 October 2025, https://link.springer.com/chapter/10.1007/978-981-96-7289-9_44
48. **Ismail Hossain**, Department of Computer Science, George Mason University, Fairfax, VA, USA, **Mridul Banik**, Department of Computer Science, Colorado State University, Fort Collins, CO, USA, <https://ieeexplore.ieee.org/abstract/document/11446534>
DOI: 10.1109/FMLDS67896.2025.00106
49. **Ismail Hossain, Mridul Banik, Fahmid Al Farid, Jia Uddin, Hezerul bin Abdul Karim**, Advancing Large Language Models for Low-Resource Languages: A Systematic Review of Pretraining, Adaptation, and Ethical Challenges, 10.32604/cmes.2026.075507, Date: 2026
<https://www.sciencedirect.com/org/science/article/pii/S152614922600322X>
50. **Ismail Hossain; Mridul Banik**, Bridging Visual and Linguistic Intelligence for Chest X-rays: A Comprehensive Review of ViTs and LLM Synergies, Published in: 2025 IEEE International Conference on Future Machine Learning and Data Science (FMLDS) Date of Conference: 02-05 November 2025, Date Added to IEEE Xplore: 25 March 2026, ISBN Information: <https://ieeexplore.ieee.org/abstract/document/11446534>
51. **Rifat Sarker Aoyon¹, Ismail Hossain², M. Abdullah-Al-Wadud³, Jia Uddin⁴**, A Secured and Continuously Developing Methodology for Breast Cancer Image Segmentation via U-Net Based Architecture and Distributed Data Training, <https://www.sciopen.com/article/10.32604/cmes.2025.060917>
52. **M. B. Vibha, M. Chandrika, Samitha Khaiyum & P. Rakshitha Kiran**, Predicting sedentary behavior in adults using stacked LSTM modeling, published: 27 November 2024
Volume 16, pages 346–355 (2025), <https://link.springer.com/article/10.1007/s13198-024-02622-2>
53. **Bellary Chiterki Anil, Jayasimha Sondekoppa Rajkumar, T. L. Divya, Samitha khaiyum Rakshitha Kiran P., Balakrishnan Ramadoss**, A Radiomics-based Framework for Liver Cancer Analysis using Explainable Artificial Intelligence (XAI) Methods, Vol. 15 No. 3 (2025): June, 2025, <https://etasr.com/index.php/ETASR/article/view/10377>
54. **Jayminkumar Patel, Sanjaykumar patel, Birju Patel, Rajan shah**, Development and Validation of a Stability-Indicating Analytical Method for Quantitative Determination of Selected Anticancer Drugs in Pharmaceutical Dosage Form, Archives / Vol. 35 No. S2 (2026): Apr-jun 2026, <https://welltestingjournal.com/index.php/WT/article/view/288>
55. **Jayminkumar Patel, Sanjaykumar patel, Birju Patel, Rajan shah**, Artificial Intelligence-Guided Optimization of Chromatographic Methods for Pharmaceutical Analytical Development, Archives / Vol. 35 No. S2 (2026): Apr-jun 2026, <https://welltestingjournal.com/index.php/WT/article/view/294>