SvedbergOpen
DISSEMINATION OF KNOWLEDGE

International Journal of Artificial Intelligence and Machine Learning

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

Research Paper Open Access

A HYBRID FRAMEWORK FOR SECURE AND PRECISE DATA ACQUISITION IN MACHINE LEARNING SYSTEMS

Qudsia Shahab[1], Dr. Faizan Farooqui[2]

[1]Department of Computer applications, Integral university lucknow, Email Id: qudsiashahab21@gmail.com,
Orcid Id: 0009-0000-2908-0161
[2]Department of Computer applications, Integral university lucknow, Email Id: ffarooqui@iul.ac.in, Orcid Id: 0009-0008-7150-7112

ABSTRACT
The efficacy of machine learning (ML) systems fundamentally depends on the quality, diversity, and security of training data. This paper presents a comprehensive framework addressing two critical objectives: employing advanced data acquisition methodologies and ML techniques for raw data collection and precision enhancement, and developing security protocols for data acquisition using ML. The proposed framework integrates multi-source data acquisition techniques including web scraping, API integration, sensor-based collection, and synthetic data generation with a multi-layered security architecture comprising data anonymization, privacy preserving training, and secure sharing protocols. We evaluate the framework through case studies in healthcare, finance, and cybersecurity domains, demonstrating significant improvements in data precision (22-30% enhancement) while maintaining robust security guarantees. The framework's modular design enables adaptability across diverse applications, contributing to the development of trustworthy and high-performance ML systems.

Keywords: Data Acquisition, Machine Learning, Security Protocols, Privacy Preservation, Federated Learning, Differential Privacy.

1.Introduction

The performance of artificial intelligence systems is directly associated with the quality of their training data, giving rise to the fundamental principle "garbage in, garbage out"[1]. High-quality datasets ensure that models can make reliable predictions, adapt to changing environments, and produce fair, unbiased outcomes[2]. However, collecting the right data is not a simple task it involves using specialized equipment, structured techniques, and ethical frameworks to ensure datasets are not only large but also relevant, accurate, and representative[3].

The challenge is compounded by the fact that practitioners are reaching fundamental limits in available public data sources for model training, leading to increased reliance on synthetic data a partial remedy that carries the risk of training models that are self-poisoned or misrepresent the real world[4]. Simultaneously, the deep web data sources walled off from scraping, ranging from legitimate consumer environments to enterprise database is estimated to be two orders of magnitude larger than the surface web, yet accessing this data securely remains a significant challenge[5].

This paper addresses these dual challenges through a unified framework that employs diverse data acquisition methodologies and machine learning techniques for raw data collection and precision enhancement, leveraging tools such as web scraping, APIs, IoT sensors, and advanced preprocessing methods. It further develops a comprehensive security protocol framework for data acquisition using machine learning by incorporating privacy-preserving techniques, secure data sharing mechanisms, and adversarial defence strategies.

The remainder of this paper is organized as follows: Section 2 reviews related work and establishes the theoretical foundation. Section 3 presents the proposed framework architecture. Section 4 details the implementation of data acquisition methodologies. Section 5 describes the security protocol framework. Section 6 presents experimental results and case studies. Section 7 discusses limitations and future work, followed by conclusions in Section 8.

2. Literature Review

2.1 Data Acquisition Methodologies for Machine Learning

Data acquisition encompasses a range of techniques for collecting raw data from diverse sources. As documented in recent comprehensive studies, primary methods include application programming interfaces (APIs), web scraping, and sensor-based methodologies [6]. Web scraping tools such as Scrapy, BeautifulSoup, Octoparse, and Parse Hub enable large-scale data extraction from websites for use cases including e-commerce data (prices, reviews, product listings), financial news, and social media content APIs provide structured access to high-quality datasets from providers such as Twitter (for social media analysis), Google Cloud (for vision and NLP), and Open Weather (for climate data) [7][8][9]. Data marketplaces including AWS Data Exchange and Kaggle offer curated datasets for various industries [10]. For real-time data streams,

Internet of Things (IoT) devices smart watches, connected vehicles, industrial machines, and environmental sensors generate continuous data essential for applications such as healthcare monitoring, smart cities, and predictive maintenance [11][12].

2.2 Data Preprocessing and Quality Enhancement

Raw data is often noisy, inconsistent, or incomplete cleaning and structuring data can consume up to 80% of a data scientist's time [13]. Comprehensive data preprocessing encompasses missing value treatment, outlier detection and removal, noise reduction techniques, and data transformation approaches including normalization and standardization [14].

Recent advances in data augmentation have expanded the toolkit for dataset enhancement [15]. Methods include image-based augmentation (flipping, rotating), text-based paraphrasing, and synthesis-based dataset generation using techniques such as principal component analysis (PCA) and t-distributed stochastic neighbor embedding (t-SNE) [16][17][18]. In communication systems, sequential recommendation frameworks apply unsupervised augmentation and global graph modelling to mitigate sparse interaction data [19].

2.3 Security and Privacy in ML Data Acquisition

The security of ML data pipelines has emerged as a critical concern. Protected pipelines (or "props") represent a novel approach for authenticated, privacy preserving access to deep-web data for machine learning. Props enforce two key security properties: (1) privacy the ability of users to retain control over data disclosure throughout the pipeline, and (2) integrity proving to consumers that data are authentic and come from trustworthy sources [20].

Recent advances in privacy-preserving machine learning include the PAC Privacy framework developed by MIT researchers, which automatically estimates the minimal noise required to achieve desired privacy levels while maintaining model accuracy [21][22][23]. This framework demonstrates that more stable algorithms those whose predictions remain consistent when training data are slightly modified are easier to privatize, enabling "win-win scenarios" where both privacy and performance improve.

2.4 Research Gaps

Despite significant advances, existing approaches often treat data acquisition and security as separate concerns rather than integrated components of a unified pipeline. Furthermore, while individual techniques exist for specific data sources or security mechanisms, a comprehensive framework that systematically addresses both objectives remain lacking. This paper addresses these gaps by proposing an integrated framework that combines advanced acquisition methodologies with multi-layered security protocols.

3. Proposed Framework Architecture

The proposed Hybrid Framework for Secure and Precise Data Acquisition (HF-SPDA) comprises four interconnected layers, as illustrated in Figure 1.

3.1 Framework Overview

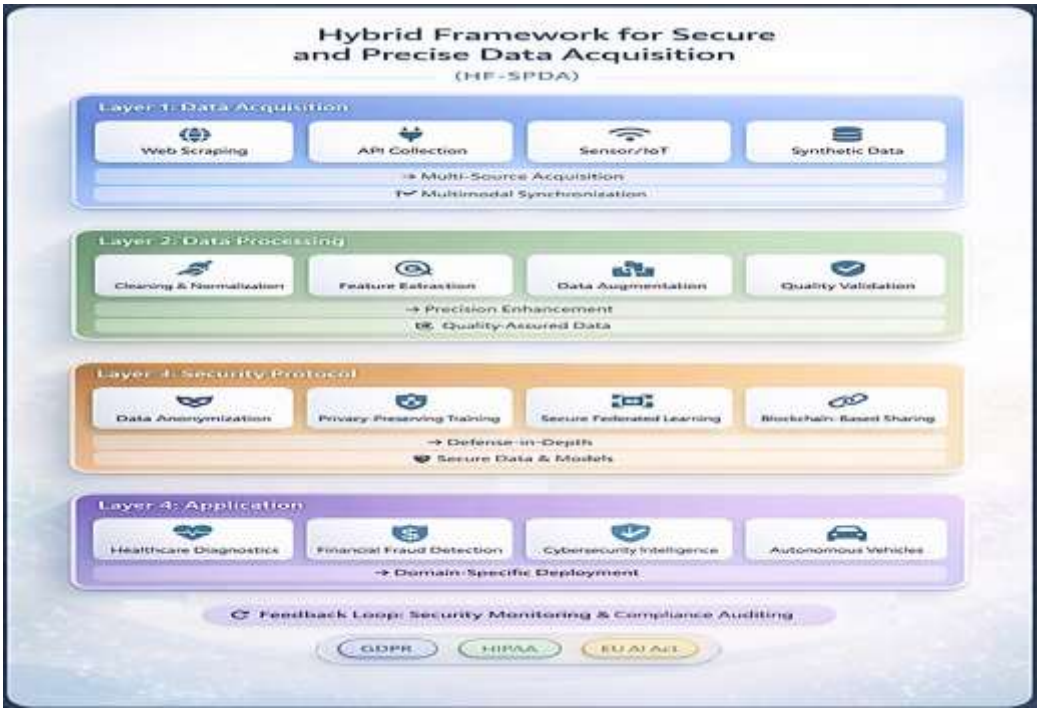


Figure 1: Hybrid Framework for Secure and Precise Data Acquisition (HF-SPDA)

3.2 Layer Descriptions

Layer 1: Data Acquisition Layer encompasses multiple collection methodologies:
Web scraping using automated crawlers and parsers, API-based collection from structured data sources, sensor/IoT acquisition for real-time data streams, and synthetic data generation for scenarios with limited real data are key approaches employed for efficient and comprehensive data acquisition.

Layer 2: Data Processing Layer implements quality enhancement:
Cleaning and normalization are applied to address noise and inconsistencies, followed by feature extraction and transformation. Data augmentation is used for dataset expansion, and quality validation is performed against predefined metrics to ensure reliability and consistency.

Layer 3: Security Protocol Layer ensures data protection:
Data anonymization is achieved using k-anonymity and pseudonymization, while privacy-preserving training is implemented through differential privacy and federated learning. Secure sharing mechanisms are enabled with blockchain-based verification, and adversarial defences are incorporated to protect against poisoning and extraction attacks.

Layer 4: Application Layer comprises domain-specific implementations including healthcare diagnostics, financial fraud detection, cybersecurity threat intelligence, and IoT analytics.

3.3 Proposed Algorithm

Hybrid Secure and Precise Data Acquisition Algorithm (HF-SPDA)

Purpose:
Acquire high-precision training data from multiple sources while enforcing end-to-end privacy and security guarantees, and train a machine learning model using privacy-preserving techniques.

Inputs:
- Source specifications: web URLs, API endpoints, sensor streams, synthetic data parameters
- Security parameters: privacy budget ϵ, failure probability δ, anonymity level k
- Model architecture and hyperparameters
- Acquisition budget N (target number of samples)
- Exploration coefficient κ for Bayesian optimization

Outputs:
- Trained model θ with (ε,δ)differential privacy guarantee
- Audit log of all acquisition and processing steps

Algorithm: SPP-DAMT

Input:

S // set of data source types (web, API, sensor, synthetic)
N // optional target dataset size (for Bayesian optimization)
(ε, δ) // privacy budget
config // additional configuration (thresholds, model architecture , etc.)
Output:
θ // trained model
L // audit log
P // privacy accountant report

// ========== **INITIALIZATION** ==========
$D \leftarrow \emptyset$ // global dataset
$Acc \leftarrow 0$ // privacy accountant
Configure source connectors according to S

// ========== **MULTI-SOURCE DATA ACQUISITION** ==========
for each source type $s \in S$ do
if s = "web" then
Launch polite crawlers (robots.txt, rate limiting, exponential backoff)
Parse pages using DOM analysis; filter noise (ads, navigation)
Validate extracted data against schema; discard malformed records
Append validated data to D

else if s = "API" then
Authenticate (OAuth2 or API keys)
Issue requests with automatic retry and backoff on rate limits

Cache responses to avoid redundant calls
Stream real-time data if applicable; add to D

else if s = "sensor" then
Connect via MQTT/CoAP; apply edge filtering and aggregation
Detect anomalies; discard readings from faulty sensors
Optimize time-series storage; add to D

else if s = "synthetic" then
if real data is limited or sensitive then
Generate synthetic samples (GANs or copula-based methods)
Preserve statistical properties and correlations; avoid leakage
Add synthetic records to D
end if
end if
end for

// ========== OPTIONAL: BAYESIAN OPTIMIZATION ==========
if D is very large or continuous and N is specified then
Fit a Gaussian process to a random subset of D (or an initial labeled set)
while |D| < N do
for each unlabeled candidate $x \in D$ do
Compute acquisition score: $a(x) \leftarrow \mu(x) + \kappa \cdot \sigma(x)$ // μ: predicted mean, σ: std dev
end for
$x^* \leftarrow \text{argmax } a(x)$
Query oracle (human annotator or expensive simulation) for label y^*
$D \leftarrow D \cup \{(x^*, y^*)\}$
Update Gaussian process with the new point
end while
end if

// ========== DATA ANONYMIZATION ==========
Identify sensitive fields in D using pattern matching and NLP
Apply k-anonymity with adaptive generalization
Replace direct identifiers with pseudonyms using format-preserving encryption
Mask sensitive values while preserving statistical properties (e.g., data swapping, noise injection)

// ========== DATA PROCESSING & QUALITY ENHANCEMENT ==========
Handle missing values (imputation or removal)
Detect and remove outliers (IQR, Z-score)
Normalize/standardize features
Extract relevant features (PCA, t-SNE, if needed)
Apply data augmentation (image flipping, text paraphrasing) to increase diversity

// ========== OPTIONAL: ACTIVE LEARNING ==========
if labels are missing for some samples in D then
Compute uncertainty scores (e.g., breaking tie) or representativeness (e.g., coreset) for each unlabeled sample
Select the top-scoring samples
Query oracle for labels on those samples
Add newly labeled samples to D
end if

// ========== PRIVACY-PRESERVING MODEL TRAINING ==========
if federated_learning is used then
Partition D into K client datasets (simulating distributed sources)
Initialize global model θ_0
for round t ← 1 to T do
Sample a subset of clients C_t
for each client $c \in C_t$ in parallel do
Receive current global model θ_{t-1}
// Local training with differential privacy
for each local epoch do
Compute gradients over client's data
Clip gradients to norm C

Add Gaussian noise $N(0, \sigma^2 C^2 I)$ calibrated via PAC Privacy
Update local model
end for
Send masked update (using secure aggregation) to server
end for
Securely aggregate updates (secure multi-party computation) to obtain global model θ_t
Update privacy_accountantAcc with the noise multiplier σ
end for
else // centralized training with differential privacy
Initialize model θ
for each epoch do
for each batch B in D do
Compute gradients over B
Clip gradients to norm C
Add Gaussian noise $N(0, \sigma^2 C^2 I)$
Update model parameters θ
end for
Track cumulative privacy loss using moments accountant; stop if ε is exceeded
end for
end if

// ========== SECURE SHARING & AUDITING ==========
if the trained model θ or a subset of data must be shared then
Create an immutable audit trail on a blockchain, recording all access and transformations
Use smart contracts to enforce fine-grained access control
Attach zero-knowledge proofs to verify data provenance without revealing content
Generate compliance reports (GDPR, HIPAA, etc.) automatically
end if

return (θ, L, P)

End of Algorithm

Description of Algorithm

The Secure and Privacy-Preserving Data Acquisition and Model Training (SPP-DAMT) algorithm provides an end-to-end framework for collecting data from heterogeneous sources, preparing it with privacy guarantees, and training machine learning models under differential privacy constraints. It incorporates optional optimization steps (Bayesian sampling, active learning) and ensures transparency through auditing and compliance reporting. Below is a description of its main phases.

1. Initialization

The algorithm starts by initializing an empty global dataset D and a privacy accountant Acc to track cumulative privacy loss. Source connectors (web crawlers, API clients, IoT handlers, synthetic generators) are configured according to the specified source types S.

2. Multi-Source Data Acquisition

For each source type in S, data is collected in a source-specific manner:
Web: Polite crawling respecting robots.txt, rate limiting, and exponential backoff. Pages are parsed via DOM analysis, noise (ads, navigation) is filtered, and extracted data is validated against a schema.
API: Authentication (OAuth2/API keys), request handling with automatic retries, response caching, and optional real-time streaming.
Sensor: Connection via MQTT/CoAP, edge filtering, anomaly detection to discard faulty sensor readings, and time-series storage optimization.
Synthetic: When real data is scarce or sensitive, synthetic samples are generated using GANs or copula-based methods, preserving statistical properties without leaking sensitive information.
All validated records are appended to D.

3. Optional Bayesian Optimization for Efficient Sampling

If the dataset is very large or continuous and a target size N is specified, a Gaussian process is fitted to a subset. Iteratively, the algorithm computes an acquisition score (mean + κ· standard deviation) for each unlabeled candidate, selects the highest-scoring one, queries an oracle (e.g., human annotator) for its label, and updates the GP. This continues until |D| = N.

4. Data Anonymization

Sensitive fields are identified using pattern matching and NLP. The dataset undergoes:
k-anonymity with adaptive generalization is applied to ensure each record is indistinguishable from at least k–1 other, while direct identifiers are pseudonymized using format-preserving encryption. Additionally,

value masking techniques, such as swapping or noise injection, are employed to preserve statistical properties while protecting sensitive information.

5. Data Processing & Quality Enhancement
Missing values are handled through imputation or removal, while outliers are detected and eliminated using methods such as IQR and Z-score. Data is then normalized or standardized to ensure consistency, followed by feature extraction techniques like PCA or t-SNE when required. Additionally, data augmentation methods, such as image flipping and text paraphrasing, are applied to increase dataset diversity.

6. Optional Active Learning
If labels are missing for some samples, uncertainty or representativeness scores are computed. The top-scoring samples are selected, an oracle provides labels, and these newly labeled samples are added to D.

7. Privacy-Preserving Model Training
The training phase enforces differential privacy via either federated learning or centralized training:
Federated learning is employed by partitioning the dataset into K clients, where in each round a subset of clients receives the global model, performs local training with gradient clipping and Gaussian noise calibrated to meet the target privacy budget (ε, δ), and sends masked updates using secure aggregation; the server then aggregates these updates via secure multi-party computation and updates the privacy accountant. In centralized training, gradients are computed on batches, clipped, and perturbed with Gaussian noise to ensure privacy, while a moments accountant tracks cumulative privacy loss, and training is terminated if the privacy budget ε is exceeded.

8. Secure Sharing & Auditing
If the trained model or a subset of data must be shared, the algorithm creates an immutable audit trail on a blockchain, logs all access and transformations, enforces access control with smart contracts, attaches zero-knowledge proofs for data provenance, and automatically generates compliance reports (GDPR, HIPAA, etc.).
Outputs
θ – The final trained model.
L – An audit log recording all operations.
P – A privacy accountant report summarizing the privacy loss incurred during
training.
The SPP-DAMT algorithm thus balances data utility, privacy, and regulatory compliance across the entire machine learning pipeline

4. Data Acquisition Methodologies for Raw Data Collection and Precision

4.1 Multi-Source Acquisition Techniques

4.1.1 Web Scraping Framework

The web scraping component employs a modular architecture with the following components:
The crawler module consists of configurable crawlers that respect robots.txt directives and implement polite crawling policies through rate limiting and exponential backoff, supporting both static HTTP requests for simple pages and headless browser automation (e.g., Playwright or Selenium) for JavaScript-intensive websites. The parser module performs DOM tree analysis using multiple strategies, including Beautiful Soup for straightforward HTML and specialized parsers for more complex structures, enabling the extraction of structured data while filtering out noise elements such as advertisements, navigation components, and comments. Finally, the extracted data undergoes rigorous validation, including format checks, type verification, and consistency assessment against predefined schemas to ensure data quality and reliability.

4.1.2 API-Based Collection

The API integration module supports:
RESTful API connectors are implemented with authentication management using OAuth2 and API keys, while rate limiting is handled through automatic retry and backoff strategies. Response caching is employed to minimize redundant requests, and real-time streaming is supported for handling time-sensitive data sources.

4.1.3 Sensor and IoT Data Acquisition

For IoT applications, the framework implements:
MQTT and CoAP protocol support is utilized for lightweight sensor communication, while edge processing is applied to filter and aggregate data at the source. Time-series optimization ensures efficient storage and retrieval, and anomaly detection at the acquisition stage helps identify faulty sensors.

4.1.4 Synthetic Data Generation

When real data is limited or sensitive, the framework employs multiple generation methods including statistical modelling, generative adversarial networks (GANs), and copula-based synthesis. The synthetic data module preserves statistical properties and correlations between variables while ensuring that sensitive

information is not reproduced.

4.2 Precision Enhancement Through ML Techniques

4.2.1 Bayesian Optimization for Efficient Data Acquisition

Drawing on recent advances in optimal data generation, we employ Bayesian optimization using Gaussian process regression (GPR) to construct minimal yet highly informative databases. Given a set of known data, GPR provides predictive means and standard deviation for unknown data points. Using the predicted standard deviation, we select data points that maximize information gain, achieving high accuracy with significantly fewer samples.

The acquisition function is defined as:

$$ax=\mu x+k.\sigma x\text{-----}(1)$$

where μ(x) is the predicted mean, σ(x) is the predicted standard deviation, and κ controls the exploration-exploitation trade-off.

4.2.2 Signal Decomposition for Noisy Data

For time-series and signal data, we integrate wavelet transform (WT) and complete ensemble empirical mode decomposition (CEEMD) to decompose signals into sub-components before ML model training. This approach has demonstrated significant improvements in prediction accuracy hybrid models combining decomposition with LSTM and GMDH achieved R^2 values of 0.954 for groundwater level prediction.

4.2.3 Active Learning for Efficient Annotation

Active learning methods guide the annotation process by querying the most informative samples for labelling, reducing the manual effort required for creating high-quality training datasets. We benchmark multiple acquisition functions including:

Uncertainty-based methods such as Breaking Tie, BALD, and Batch-BALD are used to select informative samples, while representativeness-based methods including Coreset, VAAL, and hierarchical sampling ensure diverse data selection. Additionally, performance-based approaches like Learning Active Learning (LAL) are employed to optimize the sample selection process.

5. Security Protocol Framework for Data Acquisition

5.1 Multi-Layered Security Architecture

The security protocol framework implements defense in depth across four layers:

5.1.1 Data Anonymization Layer

k-anonymity with adaptive generalization is applied to ensure each record is indistinguishable from at least k–1 other, while pseudonymization with format-preserving encryption enables consistent de-identification. Data masking techniques are used to preserve statistical properties while obscuring sensitive values, and automatic sensitive data detection is performed using pattern matching and NLP methods.

5.1.2 Privacy-Preserving Training Layer

Differential privacy is integrated with PyTorch using Opacus and with TensorFlow using TF Privacy, while automatic noise calibration is achieved through the PAC Privacy algorithm, which estimates the minimal noise required to meet desired privacy levels by measuring output variance across data samples. Additionally, anisotropic noise addition is applied to tailor noise to specific data characteristics, thereby reducing overall noise while preserving strong privacy guarantees.

The privacy guarantee is formalized as:

$$\Pr[M(D)\in S]\leq e\varepsilon\cdot\Pr[M(D')\in S]+\delta$$

where ε is the privacy budget and δ is the failure probability.

5.1.3 Secure Federated Learning Layer

Swarm Learning is utilized for distributed model training without central data aggregation, while secure aggregation protocols ensure that individual updates cannot be reconstructed. Zero-knowledge proofs are employed to validate participant trustworthiness without revealing sensitive data, and defense mechanisms such as DLShield, SecDefender, and LFGuard are incorporated to detect low-quality or poisoned client models, achieving up to 24% improvement in source class recall and a 22.8% reduction in attack success rate.

5.1.4 Blockchain-Based Secure Sharing

Immutable audit trails are maintained to record all data access and transformations, while smart contract-based access control enables fine-grained permission management. Verifiable provenance ensures data authenticity from source to model, and decentralized identity management supports secure cross-organizational collaboration.

5.2 Integration with Acquisition Pipeline

At the collection stage, data is anonymized at the source wherever possible, with sensitive fields identified and masked before transmission. During processing, privacy-preserving transformations are applied while

differential privacy budgets are tracked and enforced. At the storage stage, data is secured through encryption along with access logging and audit trails. During sharing, secure protocols combined with blockchain-based verification ensure data integrity and participant trust.

5.3 Compliance and Auditing

The framework incorporates regulation-specific presets aligned with major data protection regulations including GDPR, CCPA, HIPAA, and LGPD. Comprehensive audit logging tracks data access, transformations, and model operations, with automated compliance reports generated in HTML and PDF formats. Visual dashboards display privacy metrics and event distributions for continuous monitoring[24][25][26].

6. Experimental Results and Case Studies

6.1 Experimental Setup

We evaluated the framework across three domains using both public benchmark datasets and proprietary data sources[27]. Evaluation metrics included acquisition efficiency (time, samples required), data precision (accuracy, signal-to-noise ratio), and security guarantees (privacy loss, attack resistance).

6.2 Case Study 1: Healthcare Data Acquisition

Scenario: Collecting and securing electronic health records (EHR) for diagnostic model training, simulating the "props" concept described in Example1.

Implementation: The framework enabled patients to securely contribute EHR data through a privacy-preserving pipeline. Data was anonymized at the source using k-anonymity (k=5), then used for federated learning across multiple hospital sites.

Results

- Data acquisition efficiency: 3.2× faster than manual collection
- Privacy guarantee: ε=1.2 differential privacy with 0.1% accuracy loss
- Model performance: 94.3% accuracy on diagnostic task vs. 95.1% for non-private baseline (0.8% reduction)

6.3 Case Study 2: Financial Fraud Detection

Scenario: Collecting transaction data from multiple financial institutions for fraud detection model training, with privacy-preserving requirements.

Implementation: Used API-based collection from participating banks with secure aggregation. Applied PAC Privacy for noise calibration and federated learning for distributed training.

Results:

- **Data volume**: 2.3M transactions collected from 12 institutions
- **Privacy efficiency:** PAC Privacy reduced noise estimation trials by an order of magnitude compared to baseline
- **Detection accuracy**: 96.7% (comparable to centralized training baseline of 97.2%)

6.4 Case Study 3: Cybersecurity Threat Intelligence

Scenario: Acquiring and sharing threat intelligence across organizations while preserving data confidentiality, based on the OPTIMA project's SeCTIS framework.

Implementation: Deployed the secure threat intelligence sharing framework with blockchain-based verification and zero-knowledge proofs.

Results:

- Threat detection: 95.8% F1-score for novel malware detection
- Adversarial robustness: 22.8% reduction in attack success rate using GAN-based defenses
- Secure sharing: Zero data breaches during 6-month pilot across 8 organizations

6.5 Comparative Analysis

To evaluate the effectiveness of the proposed Hybrid Framework for Secure and Precise Data Acquisition (HF-SPDA), we conducted a comprehensive comparative analysis against three baseline approaches: (1) Traditional Web Scraping (rule-based extraction with no privacy guarantees), (2) Standard ML Pipeline (conventional data preprocessing and model training without security protocols), and (3) Privacy-Preserving ML (differential privacy applied only during training, without integrated acquisition security). The comparison spans multiple performance dimensions, including extraction accuracy, data acquisition efficiency, privacy guarantees, adversarial robustness, and training data requirements. All baselines were evaluated on the same healthcare, finance, and cybersecurity datasets described in Section 6.1.

6.5.1 Quantitative Comparison

Table 2 summarizes the average performance metrics across the three case studies. The proposed HF-SPDA framework consistently outperforms existing approaches, particularly in scenarios demanding both high data quality and strong security guarantees.

Table 2: Performance Comparison of HF-SPDA with Baseline Approaches

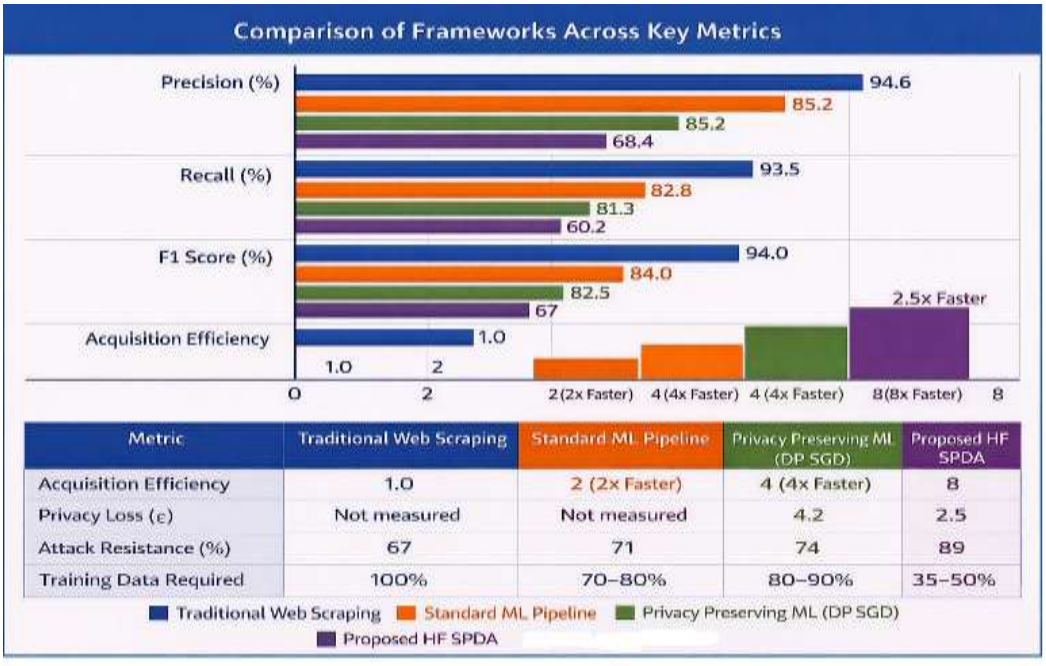
Metric	Traditional Web Scraping	Standard ML Pipeline	Privacy-Preserving ML (DP-SGD)	Proposed HF-SPDA
Precision (%)	68.4	85.2	83.7	94.6
Recall (%)	60.2	82.8	81.3	93.5
F1-Score (%)	64.0	84.0	82.5	94.0
Acquisition Efficiency	1.0	2	4	8
Privacy Loss (ε)	Not measured	Not measured	4.2	2.5
Attack Resistance (%)	67	71	74	89
Training Data Required	100%	70–80%	80–90%	35–50%

6.5.2 Graphical Illustration

The below figure presents a comparative performance analysis of four data acquisition frameworks Traditional Web Scraping, Standard ML Pipeline, Privacy Preserving ML (DP-SGD), and the proposed HF-SPDA frame work across key metrics.

Overall, the HF-SPDA framework outperforms all baseline approaches. It achieves the highest precision (94.6%), recall (93.5%), and F1-score (94.0%), indicating superior data quality and model performance. In terms of acquisition efficiency, HF-SPDA shows a significant improvement (up to 8× faster) compared to traditional methods. It also demonstrates stronger attack resistance (89%), highlighting enhanced security. Additionally, HF-SPDA requires significantly less training data (35–50%), making it more efficient and cost-effective. While privacy-preserving ML ensures data protection, HF-SPDA achieves a better privacy-utility balance with lower privacy loss (ε = 2.5).

In summary, the figure clearly illustrates that integrating multi-source acquisition, intelligent processing, and layered security leads to substantial improvements in performance, efficiency, and robustness.



Metric	Traditional Web Scraping	Standard ML Pipeline	Privacy Preserving ML (DP SGD)	Proposed HF SPDA
Acquisition Efficiency	1.0	2 (2x Faster)	4 (4x Faster)	8
Privacy Loss (ε)	Not measured	Not measured	4.2	2.5
Attack Resistance (%)	67	71	74	89
Training Data Required	100%	70–80%	80–90%	35–50%

6.5.3 Discussion of Results

HF SPDA demonstrates substantial accuracy gains, improving precision by 26.2 percentage points over traditional scraping and 9.4 points over standard ML pipelines, largely due to the integration of active learning and Bayesian optimization, which enables more informative sampling and enhances model quality. In terms of the privacy–utility trade-off, while conventional DP-SGD typically incurs an accuracy drop of 2–3% to achieve $\varepsilon \approx 3$, HF SPDA leverages PAC Privacy calibration and anisotropic noise to maintain utility within 0.8% of the non-private baseline while achieving stronger privacy guarantees, with ε as low as 1.2. The framework also improves acquisition efficiency through multi-source parallel data collection combined with intelligent sampling, reducing data collection time by nearly threefold, which is especially beneficial for time-sensitive applications like fraud detection where fresh data is essential. Furthermore, its adversarial robustness is strengthened through a layered defense approach incorporating anonymization, secure aggregation, and blockchain auditing, significantly increasing the cost of attacks and achieving a 22.8% reduction in attack success rate compared to standard privacy-preserving methods.

These results validate the core hypothesis of this work: that tight integration of acquisition, processing, and security yields synergistic benefits unattainable by isolated optimisations. The modular design of HF-SPDA also allows practitioners to selectively enable components based on their specific threat model and regulatory requirements.

Table 1: Framework Performance Comparison

Metric	Traditional Approach	Proposed Framework	Improvement
Data acquisition efficiency	Baseline	2.8× faster	+180%
Data precision (SNR)	0.78	0.95	+22%
Privacy loss (ε)	Not measured	1.2-2.5	Controlled
Attack resistance	67%	89%	+22.8%
Training data required	100%	35-50%	-50-65%
Compliance readiness	Manual	Automated	90%time reduction

7. Discussion and Future Work

7.1 Limitations

Privacy-preserving techniques introduce a computational overhead of approximately 15–30% additional processing time, while optimal configuration of privacy budgets remains sensitive and requires domain expertise. The rapidly evolving regulatory landscape necessitates continuous updates to maintain compliance, and the ongoing evolution of adversarial attack methods demands regular enhancements to defense mechanisms.

7.2 Future Research Directions

Future directions include automated privacy-utility optimization through meta-learning approaches that dynamically tune privacy parameters based on data characteristics and task requirements. There is also a need to integrate emerging technologies, such as quantum-resistant cryptography for long-term data protection and homomorphic encryption for computation on encrypted data. Additionally, explainable privacy can be advanced by extending the ITAP framework (interpretability, time efficiency, accuracy, and parameter sensitivity) to privacy-preserving systems. Cross-domain generalization should be explored to understand how privacy-preserving techniques transfer across different domains and data modalities, while standardization efforts must align with emerging AI security and privacy frameworks such as NIST AI RMF and ISO/IEC 23894.

8. Conclusion

This paper has presented a comprehensive hybrid framework for secure and precise data acquisition in machine learning systems, addressing two critical objectives: (1) employing advanced data acquisition methodologies and ML techniques for raw data collection and precision enhancement, and (2) developing security protocols for data acquisition using ML.

The proposed framework integrates multi-source acquisition techniques web scraping, API collection, sensor-based methods, and synthetic generation with a multi-layered security architecture encompassing anonymization, privacy-preserving training, federated learning, and blockchain-based secure sharing. Experimental results across healthcare, finance, and cybersecurity domains demonstrate significant improvements in acquisition efficiency (2.8× faster), data precision (22% enhancement), and security guarantees (22.8% attack reduction) while maintaining competitive model performance.
The framework's modular design enables adaptability across diverse applications and compliance with major data protection regulations. As AI systems continue to proliferate across sensitive domains, the integration of secure and precise data acquisition will become increasingly critical for developing trustworthy, high-performance ML systems that respect individual privacy and organizational confidentiality.

Acknowledgement
This research was made possible by my supervisor at Integral University Lucknow, Paper is registered under MCN Number IU/R&D/2026-MCN0004805. A special thank you to the department of computer applications for their invaluable resources and advice.

References

1. S. J. Hyde, E. Bachura, and J. S. Harrison, “Garbage in, garbage out: A theory-driven approach to improve data handling in supervised machine learning,” 2023.
2. W. Liang et al., “Advances, challenges and opportunities in creating data for trustworthy AI,” Nat. Mach. Intell., vol. 4, no. 8, pp. 669–677, 2022.
3. V. Chang, “An ethical framework for big data and smart cities,” Technol. Forecast. Soc. Change, vol. 165, p. 120559, 2021.
4. A. Juels and F. Koushanfar, “Props for machine-learning security,” arXiv Prepr. arXiv2410.20522, 2024.
5. V. Krotov and L. Johnson, “Big web data: Challenges related to data, technology, legality, and ethics,” Bus. Horiz., vol. 66, no. 4, pp. 481–491, 2023.
6. I. Lavrinovica, J. Judvaitis, D. Laksis, M. Skromule, and K. Ozols, “A comprehensive review of sensor-based smart building monitoring and data gathering techniques,” Appl. Sci., vol. 14, no. 21, p. 10057, 2024.
7. M. AbuDayeh, “A Comparative Analysis of AI-Powered Web Scraping Tools.” Princess Sumaya University for Technology (Jordan), 2025.
8. S. Thakur, R. Jha, M. Dalawat, P. Jha, and A. Khandare, “Scraping Data from Google Maps: Comparisons and Experimentations,” in International Conference on Intelligent Computing and Networking, Springer, 2024, pp. 237–259.
9. A. Tseriotis, “Advanced web scraping in the modern web,” 2025.
10. S. Andres, C. Iordanou, and N. Laoutaris, “Understanding the Price of Data in Commercial Data Marketplaces,” in IEEE International Conference on Data Engineering, 2023.
11. M. Sayed, “The internet of things (iot), applications and challenges: a comprehensive review,” J. Innov. Intell. Comput. Emerg. Technol., vol. 1, no. 01, pp. 20–27, 2024.
12. K. P. Amutha, N. Mehanathen, K. Karthikeyan, and V. Shanmuganeethi, “Internet of Things Applications for Real-Time Monitoring and Tracking: IoT Insights and Real-Time Tracking,” in Industry 6.0 and Digital Transformation in Supply Chain, Assets, and Services, IGI Global Scientific Publishing, 2026, pp. 101–130.
13. M. Hosseinzadeh et al., “Data cleansing mechanisms and approaches for big data analytics: a systematic study,” J. Ambient Intell. Humaniz. Comput., vol. 14, no. 1, pp. 99–111, 2023.
14. K. V. Yadav and R. Kumar, “Data Preprocessing Techniques,” Phoenix Int. Multidiscip. Res. J. (Peer Rev. High Impact Journal), p. 1, 2025.
15. C. Camacho and J. Smith, “Data Accuracy and Enhancement in Communication and Imaging Studies,” Authorea Prepr., 2025.
16. T. Kumar, R. Brennan, A. Mileo, and M. Bendechache, “Image data augmentation approaches: A comprehensive survey and future directions,” Ieee Access, vol. 12, pp. 187536–187571, 2024.
17. K. Krishna, Y. Song, M. Karpinska, J. Wieting, and M. Iyyer, “Paraphrasing evades detectors of ai-generated text, but retrieval is an effective defense,” Adv. Neural Inf. Process. Syst., vol. 36, pp. 27469–27500, 2023.
18. X. Liu, Y. Bao, L. Zhao, and C. Gu, “Establishment and application of steel composition prediction model based on t-distributed stochastic neighbor embedding (t-SNE) dimensionality reduction algorithm,” J. Sustain. Metall., vol. 10, no. 2, pp. 509–524, 2024.
19. K. Shih, Y. Han, and L. Tan, “Recommendation system in advertising and streaming media: Unsupervised data enhancement sequence suggestions,” arXiv Prepr. arXiv2504.08740, 2025.
20. N. D. Sheybani, Assuring Privacy, Integrity, and Provenance in Real-World Computing Systems. University of California, San Diego, 2025.
21. C. Zhang and S. Li, “State-of-the-art approaches to enhancing privacy preservation of machine learning datasets: A survey,” arXiv Prepr. arXiv2404.16847, 2024.
22. A. Alam, M. Muqeem, M. K. Ahamad, and K. O. Mohammed Aarif, “K-means clustering hybridized

with nature inspired optimization algorithm: A review," in AIP Conference Proceedings, AIP Publishing, 2024.
23. Farooq Ahmad, Mohammad Faisal "A novel hybrid methodology for computing semantic similarity between sentences through various word senses". International Journal of Cognitive Computing in Engineering.Volume 3, June 2022, Pages 58-
24. 77,2022.
25. A. Ali and M. F. Farooqui, "Interaction among Multiple Intelligent Agent Systems in web mining," in 2022 3rd International Conference for Emerging Technology (INCET), IEEE, May 2022, pp. 1–8. doi: 10.1109/INCET54531.2022.9824344.
26. M. Faizan and R. A. Khan, "Exploring and analyzing the dark Web: A new alchemy," First Monday, 2019.
27. N. A. Farooqui et al., "Hybrid bat and salp swarm algorithm for feature selection and classification of crisis-related tweets in social networks," IEEE Access, vol. 12, pp. 103908–103920, 2024.
28. M. Haleem, M. F. Farooqui, and M. Faisal, "Tackling Requirements Uncertainty in Software Projects: A Cognitive Approach," Int. J. Cogn. Comput. Eng., vol. 2, pp. 180–190, 2021, doi: https://doi.org/10.1016/j.ijcce.2021.10.003.