http://jhmi.sums.ac.ir
Health Man & Info Sci 2025, 12(4), 274–280

HMIS
Health Management and Information Science

Original Article

Secure Integration of Electronic Health Data Using Advanced Machine Learning and Blockchain Technology

Mostafa Kashani[1], **Seddigheh Barzekar**[2], **Asma Zare**[3*]

[1]Department of Health Information Technology, Sirjan School of Medical Sciences, Sirjan, Iran
[2]Department of Medicine, Sirjan School of Medical Sciences, Sirjan, Iran
[3]Department of Occupational Health Engineering, Sirjan School of Medical Sciences, Sirjan, Iran

Abstract

Introduction: Data integration and privacy preservation in electronic health records (EHRs) remain major challenges. This study combines advanced machine learning and blockchain to improve integration and security.

Methods: Using a synthetic multicenter EHR dataset (patient records, visits, diagnoses, medications, observations, procedures), we evaluated an Irregular Fuzzy Cellular Automata (IFCA) model—which incorporates fuzzy-logic rules—against XGBoost and LightGBM. Preprocessing included complete anonymization and 98.5% missing-value imputation. Machine learning addressed data integration, inconsistency resolution, and classification; HL7-FHIR–like formats and a Hyperledger Fabric consortium blockchain evaluated secure data exchange and access control. Analyses used Python 3.10 and R 4.2.

Results: Machine learning (data integrity & classification): IFCA achieved 92% accuracy (F1=0.90, AUC-ROC=0.92), outperforming XGBoost (89%) and LightGBM (90%); ANOVA indicated statistically significant differences (P<0.05). Blockchain & interoperability (security & exchange): data-exchange success was 94%, combined privacy/security score 95%, with 92% simulated attack prevention.

Conclusion: The combined approach shows promise for EHR integration and privacy preservation. Validation on real multisite EHR data is recommended to confirm generalizability.

Keywords: Electronic Health Records, Data Integration, Privacy, Machine Learning, Blockchain, Fuzzy Logic

Article History:
Received: 13 June 2025
Accepted: 11 September 2025

Please cite this paper as:
Kashani M, Barzekar S, Zare A. Secure Integration of Electronic Health Data Using Advanced Machine Learning and Blockchain Technology. Health Man & Info Sci. 2025; 12(4): 274-280. doi: 10.30476/jhmi.2025.108377.1309.

***Correspondence to:**
Asma Zare,
Department of Occupational Health Engineering, Ibn Sina Street, Postal code: 78168-44351, Sirjan, Iran
Tel: +98 9903330894
Email: a.zare@sirums.ac.ir

Introduction

Electronic Health Records (EHRs) are digital systems that collect, store, and manage patients' health information over time. EHRs centralize clinical data, support clinical decision-making, improve diagnostic accuracy, streamline workflows, and can reduce operational costs — benefits that accrue to patients, healthcare providers, and broader health systems by enabling better-coordinated care and faster public-health responses. During crises such as the COVID-19 pandemic, EHRs enabled rapid data sharing and real-time clinical insights, highlighting their pivotal role in modern healthcare (1, 2).

Despite these benefits, large-scale EHR deployments face two interrelated and critical challenges:

Data integrity and integration. Heterogeneous systems, differing data formats, and fragmented clinical workflows produce inconsistent and incomplete records across institutions, which can lead to duplication, increased costs, and potential clinical errors. Standards such as Fast Healthcare Interoperability Resources (FHIR) aim to improve interoperability, but technical and organizational barriers limit uniform adoption (3-5).

Data security and privacy. Health data are highly sensitive and increasingly targeted by cyberattacks. Traditional protections (encryption, access controls) are necessary but may be insufficient against sophisticated threats. Regulatory requirements (e.g., HIPAA, GDPR) further complicate design choices for data sharing. Advanced approaches, including blockchain for tamper-evident audit trails and privacy-enhancing technologies, have been proposed to strengthen security while preserving usability (6-8).

To address these two problem domains, two groups of solutions are particularly relevant:

Solutions for data integrity: standardized exchange formats (e.g., HL7 FHIR–like structures), harmonization and mapping procedures, and machine-learning techniques (including tree-based models such as XGBoost and LightGBM and graph/fuzzy-based methods such as IFCA) for inconsistency resolution, record linkage, and predictive tasks. Federated learning offers a way to train models across sites without moving raw records, thereby supporting integration while respecting privacy (5, 9-11).

Solutions for data security: cryptographic protections, strict access-control policies, privacy-preserving training (e.g., differential privacy), and distributed ledger technologies (e.g., consortium Hyperledger Fabric) to provide provenance, access logging, and tamper evidence. Combining these mechanisms with standardized data exchange enables safer, auditable interoperability (6-8).

This study evaluates a combined approach that applies advanced machine-learning models (including an Irregular Fuzzy Cellular Automata, IFCA) for data-integration and classification tasks, alongside standardized exchange formats and blockchain-based controls for secure data exchange and auditability. We include a simulated multicenter case study to exercise the integration and security pipelines and to assess performance under realistic heterogeneity. The work aims to identify practical strategies to improve the reliability, usability, and safety of large-scale EHR systems.

Methods

This simulation-based experimental study (study year: 2024) addresses data integration and privacy preservation in large-scale Electronic Health Record (EHR) systems by combining quantitative and qualitative evaluations of machine-learning models, standardized data-exchange protocols, and security technologies. Reporting follows CONSORT-AI and STROBE guidance to promote transparency and reproducibility.

Data Source and Format

The experiments used a synthetic, simulated multicenter EHR dataset composed of patient records, visits, diagnoses, medication prescriptions, clinical observations, and procedures. Records were provided in both structured (CSV) and semi-structured (JSON) formats to emulate heterogeneity across sources while avoiding real-world privacy constraints. The dataset was augmented and balanced using techniques such as SMOTE, producing a sample size greater than 1,000 for model evaluation. (See Data Availability for generation scripts and repository links..

Data Integrity

This subsection describes the methods used to achieve data integration, harmonization, and model-based inference.

Preprocessing and harmonization. Preprocessing was implemented in Python 3.10 using Pandas and NumPy. Steps included schema validation, consolidation to a unified data model, code mapping (diagnoses and medications to consistent vocabularies), deduplication and record linkage, and timestamp normalization. Patient identifiers were anonymized and sensitive free-text fields were masked. Missing values were imputed using mean imputation for numerical features and mode imputation for categorical features (overall imputation success reported as 98.5%). Numerical features were scaled using Min–Max normalization. Patient ages were computed from recorded birth dates relative to the reference date September 1, 2025.

Feature engineering and balancing. Time-series observations were aligned to visit windows and transformed to summary features where appropriate. SMOTE was applied to training folds to mitigate class imbalance during supervised learning; resampling was restricted to training data to avoid leakage.

Modeling for integrity and classification. The Irregular Fuzzy Cellular Automata (IFCA) model was implemented to handle irregular graph-like record structures and uncertainty. IFCA models records as nodes with fuzzy states in [0,1]; node states are updated via local fuzzy rules using neighborhood information (implemented with NetworkX and NumPy). Cells update their state using local fuzzy rules according to the formula:

$$S_{i,t+1} = f\left(\sum_{j\epsilon N_i} \omega_j \,.\, S_{i,t}\right)$$

where f is a fuzzy membership function (we used a Gaussian membership, σ=0.5) and w_j are neighborhood weights based on graph distances (implemented with NetworkX). Neighborhood radii varied between 1 and 3, fuzzy learning rate=0.1, Gaussian membership function with σ=0.5, and iterations run until convergence (typically ≈ 50 iterations). IFCA was applied for record ranking, inconsistency resolution (e.g., conflicting discharge locations), and as a feature engineering step prior to classification.

Benchmark models (XGBoost, LightGBM) were trained for classification and integration tasks. Data were split 80:20 into training and test sets with 10-fold cross-validation on the training set for hyperparameter tuning (XGBoost: learning_rate=0.1, max_depth=6; LightGBM: num_leaves=31; GridSearchCV used to select parameters). Performance metrics included accuracy (primary metric for diagnosis classification), F1-score, AUC-ROC, processing speed (seconds, measured with the time module), and scalability (reported as data coverage percentage). Federated learning experiments were run to demonstrate training without exchanging raw records.

Data Exchange and Security Evaluation

This subsection describes the methods used to evaluate secure exchange, access control, and privacy preservation.

Standards and interoperability testing. Interoperability was exercised using HL7 FHIR–like JSON resources and RESTful API endpoints. Data exchange success was measured as the percentage of records successfully transferred, parsed, and integrated by downstream nodes. Inconsistency reduction was quantified for selected fields (for example, discharge locations).

Privacy-preserving techniques. Differential privacy was applied to selected training experiments using a gradient-noise mechanism with an ε=1.0 privacy budget. Masking and irreversible hashing were applied to identifiers and sensitive fields during preprocessing. Federated learning experiments complemented DP to reduce information sharing of raw records.

Blockchain and access control. A consortium blockchain architecture based on Hyperledger Fabric was simulated to manage authorized nodes and to provide tamper-evident audit trails via smart contracts. We assessed the combined privacy/security posture by simulating common attack vectors (for example, man-in-the-middle and replay attacks) against REST endpoints and the blockchain network. Security outcomes presented in Results (e.g., attack-prevention rate and aggregated privacy/security score) are derived from these simulations.

Case Study Scope and Analysis

The case study consisted of a simulated EHR aggregation scenario designed to exercise the integration, harmonization, and security pipelines described above. This simulated multicenter project aggregates dispersed records to evaluate real-world heterogeneity and to quantify interoperability and security outcomes. Statistical analyses were performed in R version 4.2; ANOVA was applied to relevant comparisons with significance assessed at $P<0.05$. To support reproducibility while controlling access to synthetic materials, all code, data-generation scripts, and configuration files used in the case study are available from the corresponding author upon reasonable request.

Results

Results are presented in two parts that mirror the Methods: Data integrity (preprocessing & modeling) and Data security & exchange. Statistical significance is reported at $P<0.05$ where applicable.

Data Integrity — Preprocessing Outcomes

Patient identifier anonymization was applied to all records (100% anonymization).

Missing-value imputation (mean for numerical, mode for categorical) was applied with an overall imputation success of 98.5%.

Normalization and conversion to health-standard formats achieved a reported 95% integration rate.

Security measures (masking and noise addition) increased resistance to simulated attack scenarios by 93% according to our simulated tests.

These preprocessing and privacy steps demonstrate the pipeline's capacity to standardize heterogeneous EHR inputs while protecting sensitive attributes.

Model Performance

Model performance is summarized in Table 1. IFCA achieved the highest reported classification performance (Accuracy=92%, F1=0.90, AUC-ROC=0.92) compared with XGBoost (89%, 0.87,

Table 1: Model Performance Comparison

Model	Accuracy (%)	F1-score	AUC-ROC	Processing Speed (s)
IFCA	92	0.90	0.92	12
XGBoost	89	0.87	0.89	16
LightGBM	90	0.88	0.90	14

Table 2: Data Exchange and Security Performance

Metric	Proposed Method	Traditional Method
Data Exchange Success (%)	94	84
Inconsistency Reduction (%)	86	71
Attack Prevention (%)	92	77
Privacy Preservation (%)	95	79

0.89) and LightGBM (90%, 0.88, 0.90). Reported processing speeds were 12 s for IFCA, 16 s for XGBoost, and 14 s for LightGBM (measured on the experimental setup reported in Methods). One-way ANOVA was used to assess differences across models with significance tested at P<0.05 (see Table 1 and statistical reporting).

Data Security and Exchange

Interoperability and security outcomes are presented separately from modeling results:

Interoperability testing returned a 94% data exchange success rate and an 86% reduction in inconsistencies for the evaluated integration pipeline.

Security testing of the consortium blockchain combined with privacy-preserving methods reported a 92% attack-prevention rate and a combined 95% privacy/security score.

These metrics are summarized in Table 2.

Case Study

The case study applied the integration and security pipeline to aggregated simulated EHR records from multiple hospitals to exercise diagnosis diversity handling and discharge-location harmonization. The integration, interoperability, and security outcomes reported above were observed in this exercise. ANOVA analyses confirmed significant differences where reported (P<0.05). To support reproducibility while controlling access to synthetic materials, the experimental code, configuration files, and anonymized synthetic data used in the case study are available from the corresponding author upon reasonable request.

Discussion

This We evaluated a combined approach that uses advanced machine-learning methods (including an Irregular Fuzzy Cellular Automata, IFCA), standardized data-exchange formats, and blockchain-based controls to address two core challenges in large-scale Electronic Health Record (EHR) systems: data integrity and data security. Below we compare our findings with related work, discuss implications, and describe limitations and directions for future research.

Data Integrity — Integration, Harmonization, and Modeling

Our preprocessing and harmonization pipeline reinforces the view in prior work that careful data curation is a prerequisite for reliable EHR analytics (see Yoon et al.). The adoption of HL7-FHIR–like resources in our pipeline reduced lexical heterogeneity and aided parsing, consistent with the interoperability-focused recommendations of Tachinardi et al. (12, 13). Where earlier studies relied primarily on rule-based linkage and mapping, our hybrid strategy (standards + ML for inconsistency resolution) aligns with literature advocating combined approaches for robust record linkage and harmonization (5, 14). On modeling, the IFCA approach—embedding fuzzy-logic rules within a graph-like representation—provided a natural way to handle irregular and uncertain record structures, echoing validations of fuzzy and graph-aware methods in heterogeneous health data contexts (15, 16). Tree-based gradient-boosting models (XGBoost, LightGBM) remain strong baselines for structured tasks and are widely used in comparable studies (17).

Our results support the common recommendation that model choice should be driven by the specific data characteristics and deployment constraints, in agreement with recent reviews (18). Federated learning also appears promising for enabling cross-site model training without centralizing raw records, consistent with current federated-learning literature (11, 19).

Taken together, this work supports three practical lessons emphasized in similar studies: (1) standardized data models ease downstream processing (13), (2) hybrid pipelines combining rules, mapping, and ML manage edge cases better than single-method pipelines (12, 15), and (3) privacy-preserving designs (federated learning, differential privacy) should be integrated early in pipeline design to avoid later rework (20, 19).

Data Security — Privacy, Access Control, and Distributed Ledgers

Our security evaluation examined how differential privacy, masking, and a consortium blockchain can provide provenance, tamper evidence, and controlled access—goals consistent with published proposals on ledger-based EHR governance (e.g., Stamatellis et al.) (21). The literature likewise highlights the practical trade-offs of blockchain deployments—most notably computational/storage overhead and the need to design appropriate on-chain/off-chain responsibilities (14, 22). Work on sharding and other scalability techniques suggests promising mitigations for these issues (14).

Combining blockchain with federated learning and differential privacy (as other studies have explored) offers a path toward auditability without centralizing sensitive records, but it requires careful orchestration: privacy mechanisms must be tuned to avoid undue loss in model utility, and ledger operations must avoid becoming system bottlenecks (19, 20). Prior work supports these observations and suggests governance frameworks and technical partitioning as important design considerations (20, 23).

Clinical and Operational Implications

Improved integration reduces the risk of incomplete or inconsistent records entering clinical workflows, which can accelerate and improve decision-making (24). Strong security and auditable logs can bolster stakeholder trust and regulatory compliance (25). For deployments, planners should evaluate resource constraints (compute, storage, network) and choose appropriate trade-offs (e.g., amount of on-chain logic, off-chain storage, and privacy-noise budgets) informed by the existing literature (14, 22).

Limitations and Suggestions for Future Research

A key limitation is the use of synthetic, simulated multicenter data to permit controlled integration and security testing without exposing real patient data. While synthetic data enables reproducible experimentation (12, 20), it cannot fully capture the idiosyncrasies and systematic biases of operational EHR systems; thus external validation on real multisite EHR data is necessary to determine generalizability (17, 20). Computational demands of IFCA and blockchain components may limit adoption by smaller or resource-constrained institutions; evaluating lightweight ledger approaches (for example, sharding) and optimized IFCA implementations is recommended (14, 26). Future work should also explore adaptive privacy–utility trade-offs (e.g., tuning differential-privacy budgets in federated contexts), broader governance models for consortiums, and deployment studies that measure clinical impact in real operational environments (19, 20, 27).

Conclusion

This study demonstrated that the Irregular Fuzzy Cellular Automata (IFCA) model and blockchain technology can effectively address data integration and privacy challenges in electronic health record (EHR) systems. IFCA achieved a 92% accuracy rate and faster processing, handling complex and heterogeneous data better than XGBoost and LightGBM. Additionally, data exchange standards like FHIR and blockchain technology enabled a 94% data exchange success rate and 95% security. In simple terms, this means a system that quickly and accurately integrates hospital data while keeping patient information highly secure, which is invaluable for hospitals and patients alike.

However, using synthetic data instead of real-world data and high computational costs are limitations that need addressing. We recommend future research focus on real-world datasets and cost-effective methods like federated learning to make these solutions practical. This study marks a significant step toward efficient and secure digital health systems, paving the way for improved healthcare through innovative technologies.

Acknowledgement

The authors express their sincere gratitude to the participants for their cooperation in this research. This study was supported by Sirjan School of Medical Sciences (No.402000030).

Authors' Contribution

M K collected data, performed statistical analysis, performed software processes, and contributed to the writing of the initial draft of the paper. SB contributed to the modeling of the results and edited the final draft of the paper. AZ developed the research methodology, was responsible for data validation, and also contributed to the writing of the final draft of the paper. All authors read and approved the final manuscript.

Funding

Not applicable

Ethics Approval

The protocol of the present study was approved by the Ethics Committee of the Sirjan School of Medical Sciences (IR.SIRUMS.REC.1403.019).

Conflict of Interest

There are no conflicts of interest.

References

1. Modi S, Feldman SS. The Value of Electronic Health Records Since the Health Information Technology for Economic and Clinical Health Act: Systematic Review. *JMIR Med Inform*. 2022;10(9):e37283. doi: 10.2196/37283.
2. Hsu H, Greenwald PW, Laghezza MR, Steel P, Trepp R, Sharma R. Clinical informatics during the COVID-19 pandemic: Lessons learned and implications for emergency department and inpatient operations. *J Am Med Inform Assoc*. 2021;28(4):879-89. doi: 10.1093/jamia/ocaa311.
3. Ewoh P, Vartiainen T. Vulnerability to Cyberattacks and Sociotechnical Solutions for Health Care Systems: Systematic Review. *J Med Internet Res*. 2024;26:e46904. doi: 10.2196/46904.
4. Amar F, April A, Abran A. Electronic Health Record and Semantic Issues Using Fast Healthcare Interoperability Resources: Systematic Mapping Review. *J Med Internet Res*. 2024;26:e45209. doi: 10.2196/45209.
5. Ayaz M, Pasha MF, Alzahrani MY, Budiarto R, Stiawan D. The Fast Health Interoperability Resources (FHIR) Standard: Systematic Literature Review of Implementations, Applications, Challenges and Opportunities. *JMIR Med Inform*. 2021;9(7):e21929. doi: 10.2196/21929.
6. Fonseca ALA, Barbalho IMP, Fernandes F, Arrais Junior E, Nagem DAP, Cardoso PH, et al. Blockchain in Health Information Systems: A Systematic Review. *Int J Environ Res Public Health*. 2024;21(11).:1512. doi: 10.3390/ijerph21111512.
7. Fang HSA, Tan TH, Tan YFC, Tan CJM. Blockchain Personal Health Records: Systematic Review. *J Med Internet Res*. 2021;23(4):e25094. doi: 10.2196/25094.
8. Williamson SM, Prybutok V. Balancing privacy and progress: a review of privacy challenges, systemic oversight, and patient perceptions in AI-driven healthcare. *Applied Sciences*. 2024;14(2):675. doi: 10.3390/app14020675.
9. Khosravi B, Rouzrokh P, Erickson BJ. Getting More Out of Large Databases and EHRs with Natural Language Processing and Artificial Intelligence: The Future Is Here. *J Bone Joint Surg Am*. 2022;104(Suppl 3):51-5. doi: 10.2106/JBJS.22.00567.
10. Fisher S, Rosella LC. Priorities for successful use of artificial intelligence by public health organizations: a literature review. *BMC Public Health*. 2022;22(1):2146. doi: 10.1186/s12889-022-14422-z.
11. Hiwale M, Walambe R, Potdar V, Kotecha K. A systematic review of privacy-preserving methods deployed with blockchain and federated learning for the telemedicine. *Healthc Anal (N Y)*. 2023;3:100192. doi: 10.1016/j.health.2023.100192.
12. Yoon J, Mizrahi M, Ghalaty NF, Jarvinen T, Ravi AS, Brune P, et al. EHR-Safe: generating high-fidelity and privacy-preserving synthetic electronic health records. *NPJ Digit Med*. 2023;6(1):141. doi: 10.1038/s41746-023-00888-7.
13. Tachinardi U, Grannis SJ, Michael SG, Misquitta L, Dahlin J, Sheikh U, et al. Privacy-preserving record linkage across disparate institutions and datasets to enable a learning health system: The national COVID cohort collaborative (N3C) experience. *Learn Health Syst*. 2024;8(1):e10404. doi: 10.1002/lrh2.10404.
14. Ettaloui N, Arezki S, Gadi T. Blockchain-Based Electronic Health Record: Systematic

Literature Review. *Human Behavior and EmergingTechnologies*.2024;2024(1):4734288. doi: 10.1155/hbe2/4734288.

15. Kiania K, Jameii SM, Rahmani AM. Blockchain-based privacy and security preserving in electronic health: a systematic review. *Multimed Tools Appl*. 2023:1-27. doi: 10.1007/s11042-023-14488-w.
16. Wu Z, Wang H, Wan J, Zhang L, Huang J. An inner product predicate-based medical data-sharing and privacy protection system. *IEEE Access*. 2024;12:68680-96doi: 10.1109/ACCESS.2024.3400611.
17. Wu G, Wang S, Ning Z, Zhu B. Privacy-Preserved Electronic Medical Record Exchanging and Sharing: A Blockchain-Based Smart Healthcare System. *IEEE J Biomed Health Inform*. 2022;26(5):1917-27. doi: 10.1109/JBHI.2021.3123643.
18. Al Mamun A, Azam S, Gritti C. Blockchain-based electronic health records management: a comprehensive review and future research direction. *IEEE Access*. 2022;10:5768-89. doi: 10.1109/access.2022.3141079.
19. Alzubi JA, Alzubi OA, Singh A, Ramachandran M. Cloud-IIoT-based electronic health record privacy-preserving by CNN and blockchain-enabled federated learning. *IEEE Transactions on Industrial Informatics*. 2022;19(1):1080-7.
20. Jonnagaddala J, Wong ZS. Privacy preserving strategies for electronic health records in the era of large language models. *NPJ Digit Med*. 2025;8(1):34. doi: 10.1038/s41746-025-01429-0.
21. Stamatellis C, Papadopoulos P, Pitropakis N, Katsikas S, Buchanan WJ. A Privacy-Preserving Healthcare Framework Using Hyperledger Fabric. *Sensors (Basel)*. 2020;20(22):6587.. doi: 10.3390/s20226587.
22. Liu J, Fan Y, Sun R, Liu L, Wu C, Mumtaz S. Blockchain-aided privacy-preserving medical data sharing scheme for e-healthcare system. *IEEE Internet of Things Journal*. 2023;10(24):21377-88. doi: 10.1109/JIOT.2023.3287636.
23. Vanin F, Policarpo LM, Righi RDR, Heck SM, da Silva VF, Goldim J, et al. A Blockchain-Based End-to-End Data Protection Model for Personal Health Records Sharing: A Fully Homomorphic Encryption Approach. *Sensors (Basel)*. 2022;23(1).:14. doi: 10.3390/s23010014.
24. Tan L, Yu K, Shi N, Yang C, Wei W, Lu H. Towards secure and privacy-preserving data sharing for COVID-19 medical records: A blockchain-empowered approach. *IEEE Transactions on Network Science and Engineering*. 2021;9(1):271-81. doi: 10.1109/tnse.2021.3101842.
25. Boyd JH, Thompson S, Schull M, Park AL, Hachey B, Akbari A. Synthetic Data, Common Data Models and Federation: Holy Trinity or unholy mess? *International Journal of Population Data Science*. 2024;9(5). doi: 10.23889/ijpds.v9i5.2933.
26. Zaghloul E, Li T, Ren J. d-EMR: Secure and distributed Electronic Medical Record management. *High-Confidence Computing*. 2023;3(1):100101.doi: 10.1016/j.hcc.2022.100101
27. Ruotsalainen P, Blobel B. Transformed Health Ecosystems-Challenges for Security, Privacy, and Trust. *Front Med (Lausanne)*. 2022;9:827253. doi: 10.3389/fmed.2022.827253.