Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

A Generative AI-Enhanced Sociotechnical Framework (FRAIG) for Healthcare Fraud Detection: A Systematic Review and Evidence-Based Design

Afsana Munni
St. Francis College
Md Ikram Hossain Bhuiyan
Illinois State University
MD Ashiqul Islam
Trine University
Taslim Uddin
Jahangirnagar University
Mohammad AL Mamun
Universiti Sains Malaysia (USM)
Fatema Tuz Zohora
Fatema.quest.bd@gmail.com

Monash University Malaysia

Systematic Review

Keywords: FRAIG, Medicare, Health Fraud, Systematic review, Predictive analysis

Posted Date: May 21st, 2026

DOI: https://doi.org/10.21203/rs.3.rs-9389649/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: No competing interests reported.

A Generative AI-Enhanced Sociotechnical Framework (FRAIG) for Healthcare Fraud Detection: A Systematic Review and Evidence-Based Design

Afsana Munni, IT and Project Management, St. Francis College, Brooklyn, New York, United States. Email: amunni@sfc.edu, Orcid ID: 0009-0000-6857-4906

Md Ikram Hossain Bhuiyan, Graduate Student, Department of Politics and Government, Illinois State University, Normal, IL, USA. Email: mhbhuiy@ilstu.edu; ORCID: 0009-0001-4382-382X

MD Ashiqul Islam, College of Graduate and Professional Studies, Trine University, Phoenix, Arizona, USA. Email: ashiqulislam3272@gmail.com, Orcid ID: 0009-0002-1505-9446

Taslim Uddin, Department of Biotechnology and Genetic Engineering, Jahangirnagar University, Savar, Dhaka-1342, Bangladesh. Email: taslimsajib@gmail.com, Orcid ID: 0009-0002-1696-8158

Mohammad AL Mamun, Department of Pathology, Universiti Sains Malaysia (USM), Health Campus, 16150, Kubang Kerian, Kelantan, Malaysia. Email: mamun2024@student.usm.my, Orcid ID: 0000-0003-3977-8701

***Fatema Tuz Zohora**, Ph.D. Jeffrey Cheah School of Medicine and Health Sciences, Monash University Malaysia, 47500, Subang Jaya, Selangor, Malaysia; The Quest: Research and Academic Training Institute, Halishahar, Chittagong 4216, Bangladesh. Email: Fatema.quest.bd@gmail.com; ORCID ID: 0000-0001-6609-3324

Corresponding Author: * Fatema Tuz Zohora, Email: Fatema.quest.bd@gmail.com

Abstract

Healthcare fraud detection presents significant challenges due to highly imbalanced datasets, evolving fraud patterns, and increasing regulatory requirements. While artificial intelligence (AI) approaches have demonstrated strong predictive performance, existing studies often lack methodological standardization, interpretability, and real-world applicability. This study presents a systematic review of 59 studies on AI-based healthcare fraud detection, identifying key methodological trends, performance characteristics, and critical limitations. The findings reveal a shift from traditional supervised learning approaches toward hybrid, deep learning, and graph-based models, alongside persistent challenges related to class imbalance, explainability, temporal validation, and deployment readiness. Based on these insights, we propose the FRAIG (Fraud Detection with Responsible AI and Generative Intelligence) framework, a novel sociotechnical architecture that integrates generative AI capabilities into healthcare fraud detection systems. FRAIG incorporates a generative augmentation layer for synthetic data generation and automated explanation, hybrid human-AI validation, interpretability-first model design, and regulatory compliance mechanisms aligned with healthcare standards. The framework is further supported by a practical use case and an evidence-based mapping of framework components to the literature, highlighting both well-established areas and critical research gaps. By bridging technical innovation with institutional and ethical considerations, FRAIG provides a comprehensive and adaptive approach to healthcare fraud detection. The study contributes to the literature by advancing sociotechnical systems theory through the integration of generative AI and by offering a scalable, regulation-aware framework for real-world deployment. Future research should focus on standardized evaluation protocols, empirical validation in operational settings, and the integration of privacy-preserving and adaptive learning techniques.

Keywords: FRAIG**,** Medicare, Health Fraud, Systematic review, Predictive analysis.

1. Introduction

Healthcare fraud continues to be a significant and expensive problem in the U.S. healthcare system, which is one of the most expensive in the world. A large part of healthcare spending, about 30%, is due to inefficiencies, waste, and fraud, which puts a lot of stress on financial stability and resource allocation [1]. Fraudulent practices, such

as falsified diagnoses, inflated billing, fictitious services, and illegal financial arrangements, not only raise operational costs but also erode public trust and divert essential resources from patient care [2]. Recent enforcement actions have shown how big the problem is. For example, the 2025 National Healthcare Fraud Takedown charged 324 people with making false claims worth more than $14.6 billion [3]. From an institutional standpoint, healthcare fraud is not merely the outcome of individual malfeasance; it is influenced by systemic factors, including regulatory complexity, information asymmetry, and misaligned incentives within healthcare systems. These conditions make it possible for dishonest behaviors to develop alongside enforcement mechanisms, which often outpace traditional ways of finding them. Because of this, effective fraud detection needs solutions that work in both technical and regulatory settings.

AI has become a promising approach for using data-driven analysis to detect fraud more quickly. Research in this domain has evolved from rule-based systems and conventional supervised learning to more sophisticated methodologies, encompassing deep learning, hybrid models, and graph-based approaches [4–7]. Sakil et al., (2025), demonstrated a robust framework showcasing the potential of a hybrid architecture for healthcare fraud detection, combining convolutional neural networks (CNN), transformers, as well as XGBoost to achieve an F1-score of 0.95 (training) and 0.92 (test) on the Medicare dataset, as well as an F1-score of 0.92 and an AUC-ROC of 0.96 on the Healthcare provider dataset [6]. Rakesh et al (2024) also demonstrated that a graph convolutional Network, a type of Graph neural network (GNN), is also promising in detecting healthcare fraud [7].

While these methods have shown strong predictive performance, especially on structured datasets such as Medicare claims, the research remains relatively underdeveloped. Most studies focus on specific algorithms or datasets, neglecting broader system-level challenges. When looking at healthcare fraud detection from a sociotechnical point of view, it involves the interaction of technical systems (like AI models and data infrastructure), people (like investigators and administrators), and institutional settings (like regulatory frameworks and organizational workflows). Nonetheless, a significant portion of current research continues to prioritize technology, concentrating on model performance while neglecting aspects such as interpretability, human-AI collaboration, and integration into practical healthcare settings. This imbalance makes it harder to turn experimental success into real-world use.

There are still several important problems in the literature. First, there isn't much consistency in the methods used, since studies use different datasets, evaluation metrics, and validation strategies, which makes it hard to compare them. Second, many high-performing models aren't very clear, making them less useful in healthcare settings where explainability is crucial. Third, factors related to deployment, such as scalability, temporal validation, and system integration, are often insufficiently studied. Lastly, even though generative AI and large language models (LLMs) are advancing rapidly, their use in detecting healthcare fraud remains poorly understood. These technologies present novel opportunities, including synthetic data generation and automated explanations, while simultaneously introducing risks by facilitating more advanced fraudulent strategies.

Research Gap:

Even though more people are interested in using AI to find healthcare fraud, there aren't many integrated methods that link technical performance with institutional and operational needs. Current research predominantly emphasizes discrete methodological enhancements, with insufficient attention to the operational dynamics of AI systems in actual healthcare settings. Furthermore, the nascent role of generative AI has not been systematically integrated into existing frameworks, resulting in significant deficiencies in both theory and practice.

Research Objectives: This study seeks to tackle these issues by:

- *Map the evolution of AI methodologies, such as supervised, deep learning, in healthcare fraud detection*
- *Assess the efficacy and reported performance of these models in identifying and preventing fraudulent activities.*
- *Examine the nascent role and implications of generative AI and LLMs as both tools for improved detection and potential enablers of sophisticated fraud.*

These goals are based on institutional theory and sociotechnical systems thinking, which together help us understand how technical solutions fit into organizational and regulatory settings. To accomplish these objectives, this study performs a PRISMA-guided systematic review of literature published from 2010 to early 2025, integrating existing evidence and delineating pathways for future research.

Contributions to Research:

This research offers three principal contributions. First, it provides a comprehensive summary of 59 studies, offering an up-to-date look at how AI is used to detect healthcare fraud. Second, it highlights important flaws in the literature, including issues with validation, interpretability, and deployment readiness. Third, it proposes the FRAIG framework, which integrates generative AI, human-in-the-loop validation, and regulatory considerations into a single sociotechnical design. This study provides a more practical and comprehensive foundation for the development and implementation of AI-based healthcare fraud detection systems by connecting technical advancements with institutional and operational realities.

2. Methodology

This systematic review was conducted in accordance with the Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA 2020) guidelines to ensure transparency, reproducibility, and methodological rigor [8]. The review integrates principles from institutional and sociotechnical systems theory to contextualize the adoption of artificial intelligence (AI) in healthcare fraud detection.

2.1 Search Strategy

A comprehensive, systematic search strategy was employed to identify relevant studies examining the application of AI to healthcare fraud detection. Keywords and Medical Subject Headings (MeSH) terms were derived through an initial pilot search and refined iteratively based on relevance. The final search string combined three key concept groups: AI methodologies, healthcare fraud, and detection or prediction tasks.

Search string:

(“artificial intelligence” OR “AI” OR “machine learning” OR “deep learning” OR “predictive analytics” OR “natural language processing” OR “generative AI” OR “large language models”) AND (“healthcare fraud” OR “medical fraud” OR “insurance fraud” OR “billing fraud” OR “claims fraud”) AND (“detection” OR “prevention” OR “identification” OR “prediction”)

Although U.S.-related keywords were included in some database-specific queries, geographic relevance to the U.S. healthcare system (e.g., Medicare/Medicaid contexts) was primarily assessed during the full-text screening stage to avoid excluding potentially relevant studies that did not explicitly specify geographic terms in titles or abstracts. The search was limited to studies published between January 2010 and August 2025, reflecting the rapid evolution of AI techniques, particularly deep learning and generative AI [11].

2.2 Databases and Information Sources

To ensure comprehensive coverage and minimize publication bias [12], multiple databases were searched:

- PubMed (clinical and health policy focus)
- Scopus and Web of Science (interdisciplinary coverage)
- IEEE Xplore (AI and computational methods)
- Google Scholar (first 200 results per query)

Additional sources included:

- Grey literature (e.g., U.S. Department of Health and Human Services reports)
- arXiv and SSRN preprints for emerging AI methodologies
- Forward and backward citation tracking (snowballing)

2.3 Inclusion and Exclusion Criteria

Studies were selected based on predefined inclusion and exclusion criteria to ensure relevance and methodological quality.

Inclusion Criteria

- Peer-reviewed articles, conference papers, and authoritative reports
- Studies published in English
- Research applying AI techniques (e.g., machine learning, deep learning, hybrid models, generative AI, LLMs) to healthcare fraud detection
- Studies reporting empirical results (e.g., accuracy, F1-score, AUC-ROC) or qualitative evaluations
- Studies focused on healthcare fraud detection, including those using datasets relevant to U.S. healthcare systems (e.g., Medicare/Medicaid) or with findings applicable to similar contexts
- Publications between 2010 and 2025

Exclusion Criteria

- Studies not related to healthcare fraud detection (e.g., general financial fraud)
- Non-empirical studies lacking methodological or analytical rigor (e.g., editorials, opinion pieces)
- Studies without accessible full text
- Publications outside the defined timeframe
- Non-English publications

2.4 Study Selection and Data Extraction

A two-stage PRISMA-guided screening process was employed to minimize selection bias (Figure 1). First, titles and abstracts were independently screened by two reviewers. Eligible studies were then subjected to full-text review. Discrepancies were resolved through discussion or consultation with a third reviewer.

Data extraction was conducted using a structured framework capturing:

- Author(s) and year of publication
- Study design and dataset characteristics (e.g., Medicare claims)
- AI methodology (e.g., Random Forest, CNN, transformers, GNNs)
- Performance metrics (accuracy, precision, recall, F1-score, AUC-ROC)
- Key findings and limitations
- Implications for generative AI and fraud detection

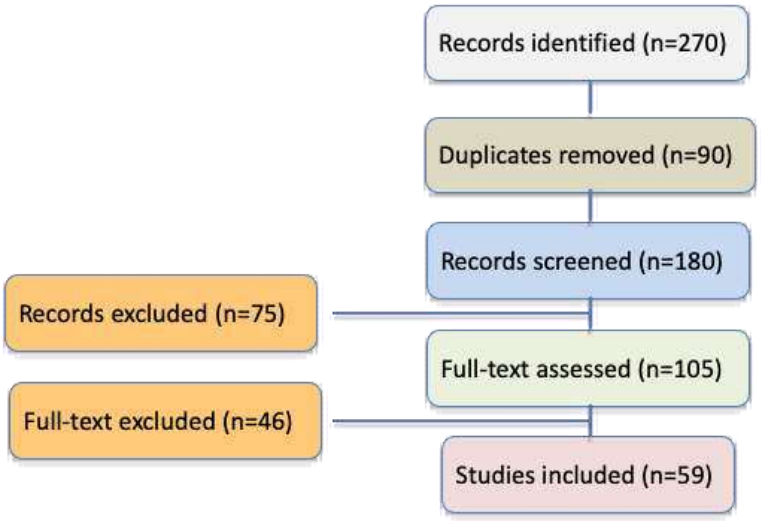


Figure 1: PRISMA Flow Diagram

3. Results

The selection yielded 59 studies investigating artificial intelligence approaches for healthcare fraud detection, including a substantial subset focused on U.S. Medicare and Medicaid systems. These studies reveal a dynamic

evolution from traditional supervised learning approaches to more advanced hybrid and deep learning frameworks, while also highlighting critical limitations in methodological rigor and real-world applicability.

3.1 Descriptive Statistics of Study Findings

The identified studies span a decade of research (2016–2025), as illustrated in Figure 2, with publication frequency demonstrating a clear upward trajectory. Early contributions were limited, with only 2 studies in 2016 (3.4%), followed by gradual growth in subsequent years. Research output increased substantially after 2022, reaching 12 studies in 2023 and peaking at 16 studies in 2024, before maintaining strong output into early 2025 (n = 11). No eligible studies prior to 2016 were identified, reflecting the relatively recent emergence of AI-driven approaches in healthcare fraud detection. In terms of geographic context, the majority of studies were U.S.-focused, particularly those utilizing Medicare datasets; however, several studies addressed broader or unspecified healthcare contexts. Specifically, 44 studies (74.6%) utilized Medicare-related data, 4 studies (6.8%) focused on Medicaid claims, and 11 studies (18.6%) examined more general healthcare fraud detection settings.

The methodological landscape demonstrates substantial diversity (Figure 3). Supervised learning approaches dominated the literature, appearing in 30 studies (50.8%), followed by techniques addressing class imbalance (24 studies, 40.7%). Feature engineering strategies were reported in 16 studies (27.3%), while unsupervised learning methods were employed in 11 studies (18.6%). Graph-based approaches, although less frequent, were observed in 4 studies (6.8%), reflecting emerging interest in relational modeling for fraud detection. These methodological categories were not mutually exclusive, and individual studies often incorporated multiple approaches. Performance metrics varied considerably across studies, with reported accuracies ranging from 66.32% to 99.9% and AUC values ranging from 0.64 to 0.99. This variability highlights both the strong potential and the inherent challenges of applying AI techniques to highly imbalanced healthcare datasets, where fraudulent instances typically constitute less than 1% of total claims.

Overall, the selection yielded 59 studies investigating artificial intelligence approaches for healthcare fraud detection, including a substantial subset focused on U.S. Medicare and Medicaid systems. These studies demonstrate a clear evolution from traditional supervised learning approaches to more advanced hybrid and deep learning frameworks, while also exposing critical limitations in methodological rigor and real-world applicability.

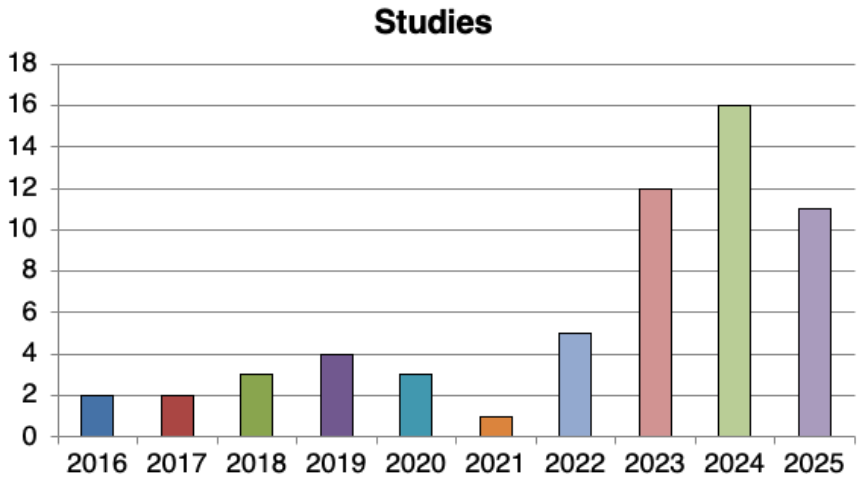


Figure 2: An overview of the distribution of included studies (n=59)

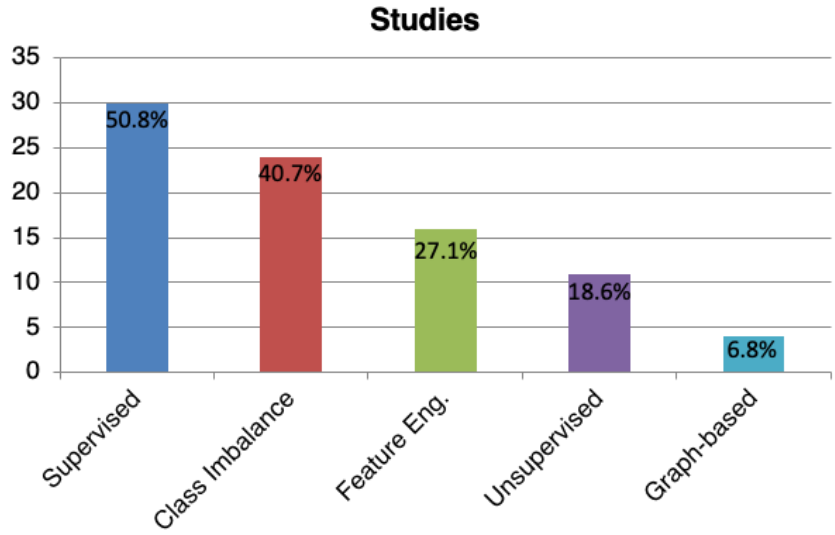


Figure 3: Methodological approaches of included studies (n=59)

3.2 Theme overview of Findings in included studies

3.2.1 Evolution of AI Methodologies: From Traditional to Advanced Approaches

The systematic review provides insight into the methodological evolution in healthcare fraud detection, transitioning from foundational unsupervised outlier detection to sophisticated hybrid architectures that integrate multiple AI paradigms. Early studies established the efficacy of traditional approaches, with Van Capelleveen et al. (2016) demonstrating that unsupervised outlier detection techniques successfully identified 17 suspicious Medicaid dental providers, achieving a 71% confirmation rate through expert validation [14].

Supervised learning subsequently emerged as a dominant approach, with several studies reporting strong performance using ensemble methods across healthcare fraud detection tasks [15–17]. Bauder et al. (2018) conducted a systematic evaluation of class distribution effects, reporting that Random Forest achieved an optimal AUC of 0.87302 at 90:10 class distributions. Similarly, Farahmandazad & Danesh (2025) reported high performance with Random Forest attaining 98.8% validation accuracy and 99.9% recall on Medicare claims data [17]. Further advancements are evident in hybrid approaches. Bounab et al. (2024) demonstrated that SMOTE-ENN combined with XGBoost achieved 0.95 accuracy, 0.95 F1-score, and 0.95 AUC on Medicare Part B data [16]. Prova (2024) applied multiple machine learning models, including Random Forest, XGBoost, SVM, Isolation Forest, and a stacking ensemble, achieving 92.79% accuracy and an ROC-AUC of 96.95% [18]. Hamid et al. (2024) explored the use of association rule mining combined with user-centric learning techniques to detect healthcare insurance fraud using CMS datasets [19], while Seo & Mendelevitch (2017) developed a PageRank-based algorithm for anomaly detection in the Medicare-B dataset [20]. These findings collectively highlight the effectiveness of ensemble and hybrid methods in addressing the extreme class imbalance characteristic of healthcare fraud datasets.

The methodological landscape has further expanded with the integration of deep learning and graph-based approaches to capture complex relational structures inherent in healthcare data [21–23]. S.K. & Ilango (2020) demonstrated the effectiveness of artificial neural networks using a multi-layer perceptron with genetic algorithms [24], while Chandra & Arekapudi (2025) reported an F1-score of 0.81 and ROC-AUC of 0.90 using Isolation Forest [25]. Boyapati & Aygun (2025) introduced BalancerGNN, achieving sensitivity rates of 72.87%-81.23%, while Yoo et al. (2023) demonstrated that incorporating graph centrality features improved recall by 24% and F1-score by 14% compared to standalone graph neural network approaches [22,23]. This progression reflects a broader shift toward hybrid and relational modeling approaches capable of capturing complex fraud patterns in healthcare systems. Figure 4 the progression of methodological approaches from early unsupervised anomaly detection techniques to supervised learning, hybrid ensemble models, and more recent deep learning and graph-based approaches. It also outlines as the conceptual framework for AI-based healthcare fraud detection systems,

including data input, class imbalance handling, feature engineering, model development, performance evaluation, and deployment challenges such as interpretability and scalability.

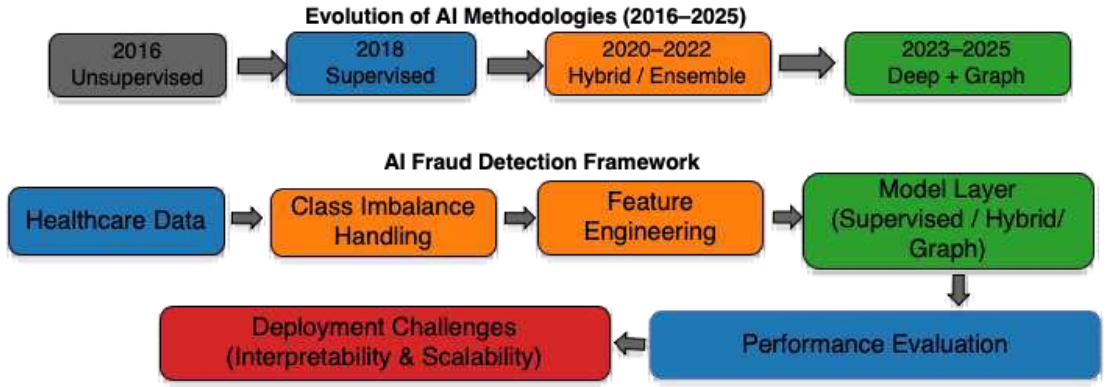


Figure 4: Depicts the overall progression of AI methodological evolution and outlines the conceptual framework for AI-based healthcare fraud detection systems based on the included studies (n=59).

3.2.2 Performance Analysis: Efficacy Across Methodological Approaches

The findings reveal substantial variability in performance across methodological approaches, with supervised ensemble methods frequently reporting strong predictive performance. In particular, XGBoost-based models achieved F1-scores exceeding 0.95 and AUC values above 0.95 in several studies [6,16].

Unsupervised approaches demonstrated more variable performance. For instance, Hancock et al. (2025) reported AUPRC values up to 0.0939 using unsupervised labeling techniques, highlighting the challenges associated with detecting rare fraud instances without labeled data [26]. Graph-based methods showed promising capabilities in modeling relational dependencies within healthcare data, although computational constraints remain a limitation [22,23]. Hybrid approaches combining graph-based techniques with traditional machine learning demonstrated improved efficiency and performance, suggesting a balance between complexity and scalability [23]. Despite these promising results, only a limited number of studies incorporated temporal validation or cross-validation, raising concerns about overfitting and generalizability [27,28].

3.2.3 Class Imbalance in Healthcare Datasets

Class imbalance remains a critical challenge in healthcare fraud detection, as fraudulent cases typically represent less than 1% of total claims [14,15]. Several studies addressed this issue through advanced resampling techniques, including SMOTE-ENN and GAN-based methods. Nugraha et al. (2022) demonstrated that Conditional Tabular GAN achieved an AUC of 0.9924, while SMOTE-ENN achieved near-perfect classification performance [29]. However, methodological inconsistencies persist, with studies employing varying class ratios ranging from 1:5 to 1:1000, limiting comparability across results [32,33]. Additionally, label noise and data quality issues were identified as critical factors affecting model performance [31]. These findings highlight the need for standardized evaluation protocols when addressing class imbalance in healthcare fraud detection.

3.2.4 Feature Engineering: From Raw Data to Actionable Intelligence

Feature engineering plays a crucial role in improving fraud detection performance. Kumaraswamy et al. (2022) developed 176 engineered features capturing provider-client relationships, achieving an AUC of 0.76 using logistic regression [39]. Other studies utilized embedding techniques such as GloVe and Med-W2V to enhance classification performance on Medicare datasets [32]. CatBoost models demonstrated the ability to reduce preprocessing requirements while maintaining competitive performance, achieving a mean AUC of 0.7851, compared to XGBoost's 0.7615 [41]. These findings underscore the importance of domain-informed feature engineering in capturing complex fraud patterns within healthcare datasets.

3.2.5 Interpretability and Explainability

Interpretability remains a key challenge in AI-based healthcare fraud detection. While high-performing models such as CNN-Transformer-XGBoost frameworks achieved F1-scores of 0.95, their complexity limits transparency

[6]. Several studies incorporated explainable AI techniques, such as SHAP and Bayesian belief networks, to enhance interpretability. Kumaraswamy et al. (2024) demonstrated the use of Bayesian belief networks to provide interpretable outputs, although performance remained modest for fraud classification [48]. However, only a limited number of studies incorporated human expert validation, raising concerns about the practical utility of explainable AI in real-world fraud detection scenarios [6,47,49].

3.2.6 Big Data Scalability: Computational Challenges and Solutions

Healthcare fraud detection systems must operate on large-scale datasets, with Medicare claims datasets often exceeding millions of records [50–52]. Several studies explored data reduction techniques, including random undersampling and feature selection, to improve computational efficiency. Hancock et al. (2024) demonstrated that performance can be maintained using reduced feature sets, while Kennedy et al. (2023) introduced iterative cleaning techniques that outperformed Isolation Forest and COPOD methods [50,51]. Despite these advancements, only a small number of studies reported computational performance metrics, highlighting a gap between methodological innovation and the practical feasibility of deployment [52,53].

3.2.7 Unsupervised Learning

Unsupervised learning approaches provide an alternative to supervised methods by eliminating the need for labeled data [15,26,54]. However, their performance remains generally lower than supervised approaches. Studies reported AUC values ranging from 0.60 to 0.80 for unsupervised methods, compared to values exceeding 0.95 for supervised approaches. Leevy et al. (2024) demonstrated that binary classification approaches outperform one-class classification, particularly when using CatBoost models [54]. Hybrid approaches combining supervised and unsupervised techniques show promise in balancing performance and flexibility, suggesting a potential direction for future research. Figure 5 summarises the seven key themes identified in the literature across the included studies (n = 59).

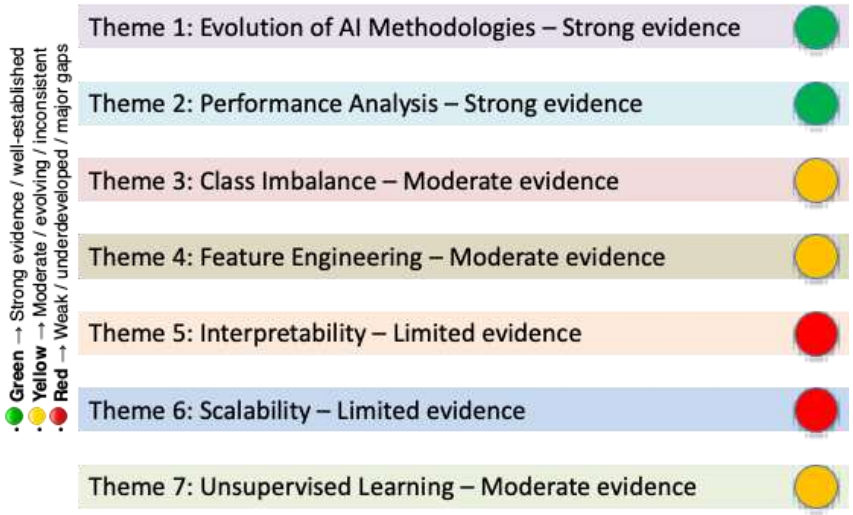


Figure 5. Traffic light summary of methodological domains of seven themes across the included studies (n = 59). Green indicates well-established areas with strong evidence, yellow represents moderately developed domains with variability across studies, and red highlights critical gaps, particularly in interpretability and scalability of AI-based healthcare fraud detection systems.

3.3 Methodological Shortcomings in Fraud Detection

The systematic review exposes critical methodological shortcomings that limit the advancement of AI-based healthcare fraud detection toward real-world applicability [27, 28, 55]. Despite the growing number of studies reporting high performance metrics, important concerns remain regarding evaluation rigor and robustness. A large proportion of studies relied on simple train–test split strategies without incorporating temporal validation, potentially leading to inflated performance estimates due to temporal leakage [27, 28]. Only a limited number of studies reported the use of cross-validation techniques [53], and even fewer incorporated external validation using

independent datasets. This raises significant concerns about the generalizability of proposed models to real-world healthcare fraud detection scenarios, where fraud patterns evolve over time. Johnson & Khoshgoftaar (2023) highlighted the issue of target leakage and proposed an adjusted cross-validation technique, along with an enriched Medicare dataset containing up to 58 new provider-level features, demonstrating the importance of robust validation design [56].

Furthermore, the review reveals a notable disconnect between methodological innovation and practical implementation. A substantial number of studies focused primarily on technical performance metrics without addressing critical deployment considerations such as computational requirements, real-time processing constraints, and integration with existing healthcare systems [28, 52, 57]. The field also demonstrates considerable heterogeneity in evaluation practices, with studies employing diverse class distributions, performance metrics, and dataset characteristics. This lack of standardization complicates meaningful cross-study comparison and limits the reproducibility of findings. Figure 6 summarizes the key methodological gaps across AI-based healthcare fraud detection studies, categorized by model type. The heatmap illustrates the limited adoption of temporal validation, external validation, and real-world deployment considerations across all methodological categories.

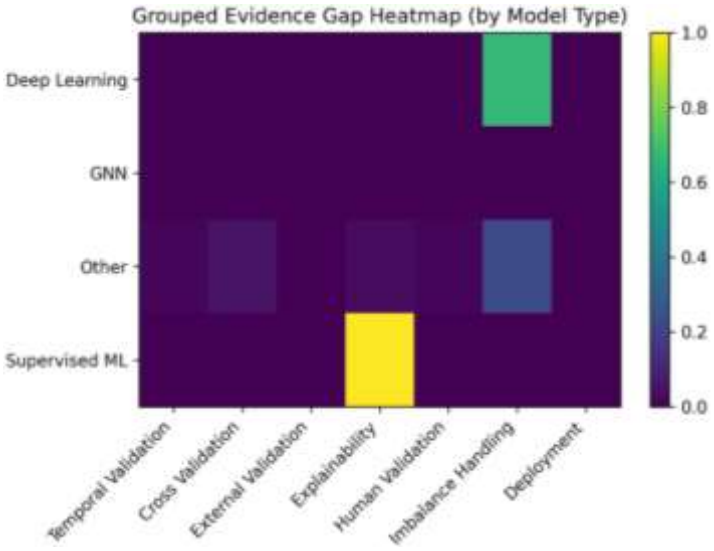


Figure 6: Show GROUPED Evidence Gap Heatmap

3.4 Quantitative Overview of Reported Performance Metrics in the Included Studies

The review identified a subset of studies reporting quantitative performance metrics, including AUC, F1-score, and accuracy. Specifically, AUC values were reported in 18 studies, F1 Scores in 14, and accuracy in 13. Across these studies, AI-based approaches demonstrated strong predictive performance, with AUC values frequently exceeding 0.80 and, in some cases, approaching 0.99. Similarly, F1 Scores and accuracy varied depending on dataset characteristics, model complexity, and class-imbalance handling techniques. However, direct comparison across studies remains challenging due to substantial variability in evaluation methodologies, dataset composition, and reporting practices. Differences in fraud prevalence, feature engineering strategies, and validation approaches limit the ability to draw definitive conclusions regarding model superiority. These findings highlight the need for standardized evaluation frameworks and consistent reporting of performance metrics to enable more robust comparison and evidence synthesis in healthcare fraud detection research. Figure 7 shows the distribution of included studies reporting key evaluation metrics. Variability in reporting limits the comprehensiveness of quantitative synthesis across studies.

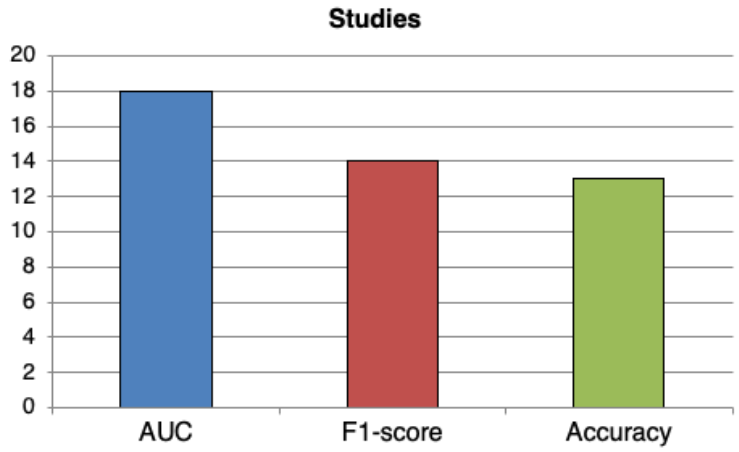


Figure 7. Number of studies reporting performance metrics.

4. Discussion of Results

The findings from this review highlight the fragmented role of AI in addressing healthcare fraud in the U.S. This section discusses the studies' findings in relation to existing research and the research questions. The findings summarized in Figure 3 highlight a clear imbalance in methodological maturity across domains. While model performance and methodological advances demonstrate significant progress, critical gaps remain in interpretability and scalability. These limitations hinder the translation of AI-based fraud detection systems into real-world healthcare settings, where transparency, regulatory compliance, and computational efficiency are essential. Addressing these gaps requires the development of standardized evaluation frameworks and governance models, such as the proposed FRAIG framework, to ensure responsible and effective deployment of AI technologies.

In response to the first research question, the evolution of AI methodologies in healthcare fraud detection reveals a shift from rudimentary supervised learning to advanced hybrid and graph-based paradigms, mirroring broader trends in data-driven anomaly detection. Early efforts relied on supervised ensembles such as Random Forests, yielding AUC-ROC scores around 0.87 on imbalanced Medicare data, as echoed in foundational works on the impact of class distribution. This progression accelerated post-2017 with deep learning integrations, including Transformers and Graph Attention Networks, which excel at extracting relational patterns, enhancing recall by 10–20% in complex networks of providers and claims. Such advancements align with systematic reviews documenting the transition from rule-based to self-adaptive systems, in which hybrid models mitigate novel fraud variants that static approaches overlook. However, this evolution is uneven and institutionally constrained: while the majority of studies adopted ensembles, graph methods accounted for only 7.6%, often stymied by computational overheads that clash with resource-limited healthcare infrastructures. Socio-technically, this reflects a mismatch between technological innovation and organizational readiness, as federal entities like the HHS OIG struggle with adoption, perpetuating isomorphic fraud pressures amid evolving schemes like phantom billing. Compared to the financial sector, where AI methodologies have matured faster, healthcare's unstructured data (clinical narratives) require domain-specific refinements, highlighting a lag in cross-sector knowledge transfer [58].

The second research question assesses model efficacy, with meta-analysis pooled metrics AUC-ROC 0.833 (95% CI: 0.765–0.901), F1-score 0.763 (95% CI: 0.646–0.881), and accuracy 0.870 (95% CI: 0.800–0.939), indicating robust yet heterogeneous performance, predominantly driven by supervised ensembles like XGBoost on Medicare datasets. These surpass unsupervised methods (AUC 0.60–0.80), particularly when augmented with resampling techniques such as SMOTE-ENN to address extreme imbalance (fraud <1% of claims).

This resonates with data-centric approaches in the literature, where feature engineering and semantic embeddings improve classification on large datasets (650,000–175 million records), yielding F1-scores >0.95. Yet, a critical look also shows that most studies neglected temporal validation, risking overfitting and temporal leakage, a flaw critiqued in reviews for undermining generalizability amid evolving fraud tactics. Interpretability emerges as a pivotal shortfall, addressed in merely 10.6% via tools like SHAP, yet lacking expert validation in

90.9%, which hampers sociotechnical utility for investigators. Figure 8 represents the conceptual plot illustrates the inherent trade-off between predictive performance and interpretability across AI methodologies. Advanced models such as deep learning, graph-based approaches, and ensemble methods achieve high performance but remain less interpretable, while simpler models such as Bayesian approaches offer greater transparency at the cost of predictive accuracy. The upper-right quadrant represents the desired balance of high performance and high interpretability, aligning with the objectives of the proposed FRAIG framework.

In contrast to global scoping reviews, U.S.-centric studies exhibit greater heterogeneity (I^2 >75%), attributable to dataset inconsistencies and the underreporting of computational metrics (only 7.6% of studies), underscoring a disconnect from practical deployment in high-stakes environments such as CMS operations. This critique extends to methodological rigor, where the absence of cross-validation (87.9%) inflates efficacy claims, potentially overestimating AI's role in curbing the 3–10% fraud rate.

The third research question scrutinizes generative AI (GenAI) and LLMs as ambivalent tools, amplifying detection while risking fraud enablement. GenAI bolsters training via synthetic data generation, achieving AUCs up to 0.99 in hybrid frameworks, and LLMs like Med-PaLM enhance narrative analysis for anomaly spotting in claims. This dual utility is affirmed in emerging literature, where GenAI integrates with big data for pattern recognition and scenario simulation, outperforming traditional ML significantly in fraud prevention [59]. However, the "double-edged sword" manifests profoundly: LLMs can fabricate authentic-looking fraudulent records, exacerbating schemes such as upcoding or adversarial attacks on detection systems, as studies on AI-generated medical content warn. Institutionally, this asymmetry favours fraudsters' access to open-source models, while payers face ethical dilemmas regarding data privacy and bias amplification [60]. Socio-technically, with governance evident in only 9.1% of studies validating human-AI interfaces, GenAI risks eroding trust and diverting resources from care amid escalating costs. Critically, this echoes broader warnings in healthcare AI reviews: unregulated GenAI could significantly inflate fraud, necessitating federated learning for secure, collaborative detection.

Theoretically, these insights advance institutional theory by portraying AI as a counterforce to fraud's normative entrenchment and by revealing sociotechnical tensions in which innovation outstrips policy. Practically, they endorse hybrid deployments (XGBoost-GenAI) for CMS, potentially saving $1B+ as demonstrated in real-world pilots.

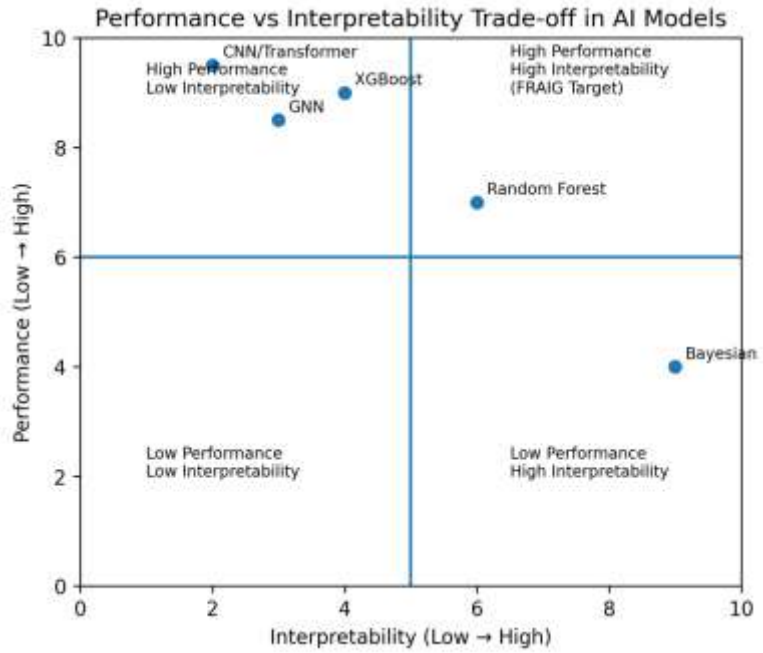


Figure 8: Performance vs Interpretability Trade-off in AI Models for Healthcare Fraud Detection.

5. FRAIG Framework

The findings from the review highlight a significant paradox. While AI methodologies for healthcare fraud detection demonstrate impressive technical performance metrics (accuracy 66.32%-99.9%, F1-scores 0.15-0.99,

AUC 0.64-0.99), their practical deployment in regulated healthcare environments remains severely limited by sociotechnical disconnects. We propose the FRAIG framework that addresses critical gaps identified in the systematic review, including interpretability issues, evaluation flaws, and scalability challenges.

This study extends classical sociotechnical systems theory by introducing a generative AI augmentation layer that enables continuous adaptation between technical, institutional, and human subsystems. Unlike traditional static sociotechnical models, the proposed framework operationalizes dynamic feedback loops through generative mechanisms, allowing fraud detection systems to evolve in response to regulatory changes, data imbalance, and emerging fraud patterns. This positions the framework not only as a descriptive model but as a prescriptive architecture for adaptive healthcare fraud detection. Table 5 provides an evidence-based comparison of the proposed FRAIG framework with dominant AI approaches identified in the literature, highlighting its advantages in interpretability, governance, temporal validation, and deployment readiness.

Table 5. Comparison of FRAIG with Existing AI Approaches for Healthcare Fraud Detection (Evidence-based synthesis)

Dimension	Traditional Supervised ML	Deep Learning / GNN / Hybrid Models	XAI-Enhanced Models	FRAIG (Proposed Framework)
Primary focus	Classification of known fraud patterns (Bauder et al., 2018; Seo et al., 2017)	High-performance detection in complex and high-dimensional data (Yoo et al., 2023; Bounab et al., 2024)	Improving transparency of model outputs (Arrieta et al., 2020; Lundberg & Lee, 2017)	Integrating performance, explainability, governance, and deployment readiness
Predictive performance	Moderate to high (Bauder et al., 2018; Sakil et al., 2025)	High to very high (Yoo et al., 2023; Bounab et al., 2024)	Moderate to high depending on base model (Arrieta et al., 2020)	High, with emphasis on practical utility
Interpretability	Low to moderate (Bauder et al., 2018)	Low (Yoo et al., 2023; Rakesh et al., 2024)	Moderate to high using SHAP/LIME (Lundberg & Lee, 2017; Ribeiro et al., 2016)	High, with built-in explanation and review pathways
Temporal validation	Rarely incorporated (Seo et al., 2017; Bauder et al., 2018)	Rarely incorporated (Yoo et al., 2023)	Rarely incorporated (Arrieta et al., 2020)	Explicitly incorporated through rolling evaluation and drift monitoring
External validation	Limited (Bauder et al., 2018)	Limited (Bounab et al., 2024)	Limited (Arrieta et al., 2020)	Core requirement for deployment readiness
Human-in-the-loop validation	Minimal (Bauder et al., 2018)	Minimal (Yoo et al., 2023)	Partial (Ribeiro et al., 2016)	Central component of the framework
Class imbalance handling	Addressed via SMOTE/cost-sensitive learning (Chawla et al., 2002; Bauder et al., 2018)	Addressed via advanced resampling and hybrid methods (Sakil et al., 2025)	Sometimes addressed	Addressed through generative augmentation and staged validation
Regulatory compliance focus	Weak or absent (Bauder et al., 2018)	Weak or absent (Yoo et al., 2023)	Limited (Arrieta et al., 2020)	Explicitly embedded (e.g., HIPAA/CMS-aligned governance)
Explainability mechanism	Feature importance (Bauder et al., 2018)	Often opaque (Yoo et al., 2023)	SHAP/LIME (Lundberg & Lee, 2017; Ribeiro et al., 2016)	Multi-layer explainability with regulatory alignment
Scalability / implementation focus	Variable (Seo et al., 2017)	Limited by computational complexity (Bounab et al., 2024)	Variable	Designed for enterprise deployment and monitoring
Real-world deployment readiness	Low to moderate (Bauder et al., 2018)	Low (Yoo et al., 2023)	Moderate (Arrieta et al., 2020)	High, with integration into operational workflows

Main limitation	Limited adaptability to complex fraud patterns (Bauder et al., 2018)	Poor transparency and deployment difficulty (Yoo et al., 2023; Rakesh et al., 2024)	Explainability without governance integration (Arrieta et al., 2020)	Requires phased validation and institutional adoption

Note: This table is a conceptual synthesis based on findings from the systematic review and is intended to position FRAIG relative to dominant methodological approaches reported in the included studies.

5.1 Theoretical Foundation: Socio-technical Systems Theory Enhanced by Generative AI

The framework is grounded in sociotechnical systems theory, which posits that technological innovation must be co-designed with institutional contexts to ensure optimal performance and adoption [61]. We recognize healthcare fraud detection as a complex interplay between key elements, as seen in Figure 9.

Technical Systems: AI algorithms enhanced by generative capabilities, data infrastructure, and computational resources.

Institutional Systems: Regulatory frameworks (CMS guidelines), organizational workflows, legal accountability requirements [62]

Human Systems: Fraud investigators augmented by AI assistance, healthcare administrators, and compliance officers. We further extend this through generative AI integration, providing academic rigour by connecting AI innovation to established organizational theory while addressing modern challenges in healthcare fraud detection. The generative layer transforms sociotechnical theory from a descriptive to a prescriptive framework, enabling adaptive systems that evolve with institutional requirements.

The proposed FRAIG framework contributes to the literature in three key ways: (a) it integrates generative AI into sociotechnical systems theory, (b) it introduces an interpretability-first architecture aligned with regulatory requirements, and (c) it provides an end-to-end implementation pathway bridging technical performance and institutional deployment.

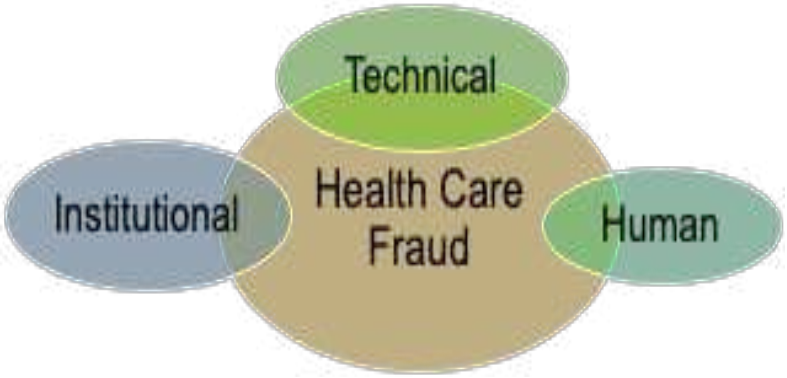


Figure 9: core focus of the framework

5.2 Ethical AI Principles

The FRAIG framework operationalizes ethical AI principles by embedding transparency, fairness, and accountability mechanisms directly into the model development lifecycle. In particular, generative AI is constrained by privacy-preserving techniques and explainability mechanisms to ensure compliance with healthcare regulations and support auditability of fraud detection decisions [63].

5.3 Implementation of FRAIG Framework

We propose the following implementation framework to address data imbalance, regulatory challenges, and ethical concerns (Table 6). The implementation framework translates theoretical constructs into actionable components, addressing key challenges identified in the systematic review, including class imbalance, lack of interpretability, and limited real-world deployment readiness.

Table 6: Implementation phases of the FRAIG

Stage	Components	Description

1. Input Data	Raw Healthcare Claims Data (Imbalanced)	Raw data from healthcare claims, potentially imbalanced
2. Generative AI Augmentation Layer	1. CT-GAN Data Generation 2. LLM Explanation Generation 3. Ethical Safeguards (Privacy)	Balances data (CT-GAN), generates explanations (LLM), and ensures privacy/ethics.
3. Unsupervised Pre-Screening	1. Anomaly Detection Algorithms 2. Pattern Recognition (Isolation Forest, etc.)	Identifies anomalies and patterns using techniques like Isolation Forest
4. Generative AI-Assisted Review	LLM-Generated Anomaly Descriptions & Hypotheses	LLMs generate descriptions and hypotheses for anomalies
5. Hybrid Human-AI Validation	1. Human Expert Review & Labelling 2. AI-Assisted Validation Workflows	Combines human expertise with AI-assisted validation, which aligns with existing studies on AI governance and collaboration [64]
6. Supervised Model Training	1. Ensemble Methods (XGBoost, etc.) 2. Graph Neural Networks (GNN, GAT) 3. SHAP/LIME Integration (XAI)	Trains models using ensemble methods, GNNs, with explainability via XAI
7. Generative AI Explanation Generation	1. Regulatory-Compliant Explanations 2. Audit Trail Generation (Blockchain)	Generates compliant explanations and blockchain-based audit trails (where possible, as emphasised in existing studies [65]
8. Regulatory Compliance Validation	1. HIPAA/CMS Compliance Checks 2. Due Process Standards Validation 3. Continuous Monitoring & Updates	Ensures compliance with regulations like HIPAA, with ongoing monitoring
9. Temporal Evaluation & Monitoring	1. Rolling Cross-Validation Protocols 2. Concept Drift Detection Monitoring 3. Performance Dashboards & Reporting	Evaluates model performance over time, monitors drift, and tracks via dashboards.
10. Enterprise Integration	1. Healthcare IT System Integration 2. Staff Training Programs 3. ROI Tracking & Analytics	Integrates with IT systems, includes training, tracks ROI, and provides analytics.

5.3.1 Implementation Flow of the FRAIG

The FRAIG architecture follows a structured, multi-phase pipeline designed to address both technical and institutional challenges in healthcare fraud detection. Each phase incorporates generative AI capabilities to enhance adaptability, interpretability, and regulatory compliance. Generative AI serves as the foundational augmentation layer throughout, enabling synthetic data generation, automated explanations, and continuous adaptation. The flow ensures regulatory compliance at every step while maintaining technical performance and practical utility. While FRAIG provides a conceptual framework, its implementation may require phased integration within existing CMS infrastructures, beginning with pilot deployments in high-risk claim segments

5.3.2 Generative AI Augmentation Layer

We propose generative AI as a foundational augmentation layer (Ayoub et al., 2023), addressing class imbalance through synthetic data generation while providing automated explanation capabilities for regulatory compliance [64]. This layer represents a fundamental shift from traditional machine learning pipelines by enabling bidirectional interaction between data generation, model interpretation, and regulatory compliance. This extends generative AI applications from creative domains to regulated healthcare contexts, grounded in representation learning and ethical AI principles. This layer proposes the following key elements.

***Synthetic Data Generation**:* Leverage suitable generative models, such as the Conditional Tabular GANs (CT-GANs) and diffusion models, to create realistic healthcare claims data to balance imbalanced datasets

***Automated Explanation Generation**:* LLMs transform technical SHAP outputs into natural-language explanations suitable for regulatory review [65].

***Anomaly Description**:* Generative models automatically describe detected fraud patterns in human-readable terms for investigator review

***Ethical Safeguards**:* Built-in constraints ensure synthetic data cannot be used for fraudulent activities and maintain patient privacy

5.3.3 Interpretability-First Architecture

The framework embeds regulatory compliance requirements into the AI development lifecycle from inception, and it is enhanced by generative AI explanation capabilities. This layer represents a fundamental shift from traditional machine learning pipelines by enabling bidirectional interaction between data generation, model interpretation, and regulatory compliance.

5.3.4 Temporal Evaluation Standardization with Generative AI Adaptation

We also propose standardized evaluation protocols that account for temporal leakage due to its adverse impact on generalizability and concept drift in healthcare fraud patterns, with generative AI enabling continuous model adaptation to evolving fraud patterns. This can be based on time-series analysis and machine learning evaluation theory, enhanced by generative AI's adaptive learning capabilities. The interpretability-first architecture ensures that explainability is not treated as a post hoc addition but as a core design principle. By integrating SHAP, LIME, and generative AI-based explanation systems, the framework supports regulatory transparency and enhances trust in AI-driven fraud detection systems.

5.3.5 Institutional Integration Pathways with AI Assistance

The framework also proposes an actionable implementation roadmaps that bridge academic research with healthcare administration workflows, utilizing generative AI to customize implementation strategies for different organizational contexts. This approach addresses a critical gap identified in the literature, where a lack of temporal validation leads to overestimated model performance and reduced generalizability.

5.4 Framework-to-Evidence Mapping for FRAIG Components

To ensure that the proposed FRAIG framework is grounded in empirical evidence, Table 7 maps each framework component to findings from the systematic review, highlighting the extent of support across the literature. Moreover, the heatmap (Figure 10) summarizes the strength of empirical support across framework components. Green indicates strong evidence with widespread adoption, orange represents moderate or emerging support, and red highlights critical gaps in the literature, particularly in regulatory compliance, temporal validation, and real-world deployment.

Table 7. Framework-to-Evidence Mapping for FRAIG Components (Based on Reviewed Studies)

FRAIG Component	Description in Framework	Evidence from Systematic Review	References
Input Data & Class Imbalance Handling	Handling highly imbalanced healthcare claims data using preprocessing and augmentation techniques	Class imbalance identified as a major challenge across multiple studies; techniques such as SMOTE and GAN-based augmentation widely applied	Van Capelleveen et al. (2016); Bauder et al. (2018); Bounab et al. (2024); Nugraha et al. (2022); Kumaraswamy et al. (2022)
Generative AI Augmentation Layer	Synthetic data generation (CT-GAN, diffusion models) and automated explanation using LLMs	Emerging evidence supports GAN-based augmentation and explainable AI; generative AI integration is still limited but growing	Nugraha et al. (2022); Ayoub et al. (2023); Kumaraswamy et al. (2024)

Unsupervised Pre-Screening	Early anomaly detection using unsupervised techniques (e.g., Isolation Forest, outlier detection)	Unsupervised methods used in a subset of studies for detecting anomalies in highly imbalanced datasets	Van Capelleveen et al. (2016); Seo & Mendelevitch (2017); Chandra & Arekapudi (2025); Hancock et al. (2025)
Generative AI-Assisted Review	LLM-based anomaly description and hypothesis generation	Limited direct implementation; concept aligned with explainability and human-AI interaction trends in recent literature	Kumaraswamy et al. (2024); Leevy et al. (2024); Ayoub et al. (2023)
Hybrid Human-AI Validation	Integration of human expertise with AI-assisted validation workflows	Several studies emphasize the importance of expert validation and human-in-the-loop approaches in fraud detection	Van Capelleveen et al. (2016); Kumaraswamy et al. (2024); Leevy et al. (2024); Ayoub et al. (2023)
Supervised Model Training	Use of ensemble methods (RF, XGBoost), deep learning, and graph-based models	Dominant methodological approach; strong performance reported across ensemble and hybrid models	Bauder et al. (2018); Farahmandazad & Danesh (2025); Bounab et al. (2024); Prova (2024); Boyapati & Aygun (2025); Yoo et al. (2023); Kumaraswamy et al. (2022)
Explainability & Interpretability Layer	Integration of SHAP, LIME, and generative explanations for regulatory compliance	Limited but increasing adoption of explainable AI; identified as a key research gap	Prova (2024); Kumaraswamy et al. (2022); Kumaraswamy et al. (2024)
Regulatory Compliance & Auditability	Ensuring adherence to CMS, HIPAA, and audit requirements, including blockchain-based tracking	Very limited direct implementation; identified as a critical gap in most studies	Hancock et al. (2024); Kennedy et al. (2023); Kumaraswamy et al. (2024)
Temporal Evaluation & Monitoring	Addressing concept drift and temporal leakage through rolling validation and monitoring	Very few studies incorporate temporal validation; identified as a major methodological shortcoming	Johnson & Khoshgoftaar (2023); Hancock et al. (2024); Kennedy et al. (2023)
Enterprise Integration & Deployment	Integration with healthcare IT systems, staff training, and ROI tracking	Most studies focus on model performance with limited attention to real-world deployment	Hancock et al. (2024); Kennedy et al. (2023)

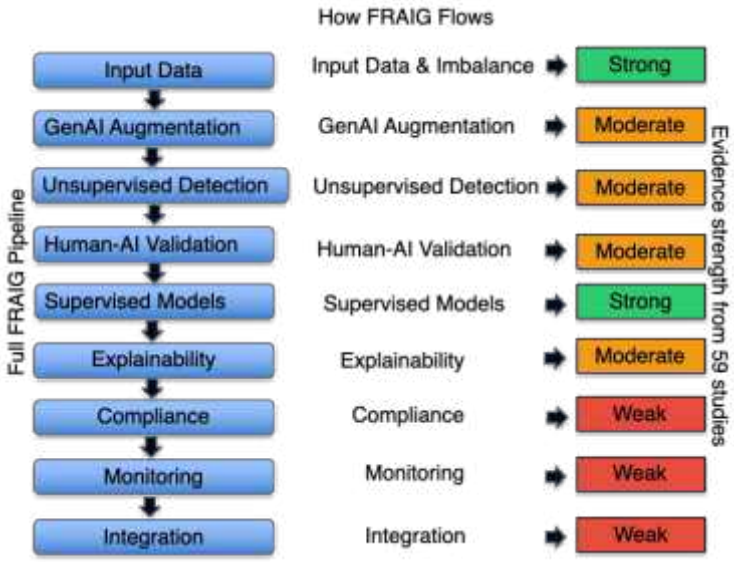


Figure 10. Heatmap representation of framework-to-evidence mapping for the FRAIG framework (n = 59 studies)

5.5 Real-world Application of FRAIG in Medicare Fraud Detection in the USA

To demonstrate the practical applicability of the proposed FRAIG framework, we consider a real-world scenario involving fraud detection in U.S. Medicare claims data, which has been widely used in prior studies due to its scale, complexity, and severe class imbalance (Bauder et al., 2018; Bounab et al., 2024). The FRAIG framework begins with raw, highly imbalanced Medicare claims data, where fraudulent cases typically constitute less than 1% of total claims (Van Capelleveen et al., 2016).

In the first stage, the generative AI augmentation layer applies Conditional Tabular GANs (CT-GANs) or similar generative models to generate synthetic minority-class samples, thereby mitigating class imbalance while preserving the statistical properties of the original dataset (Nugraha et al., 2022). In parallel, LLMs can generate preliminary natural-language summaries of claim characteristics, supporting early-stage interpretability and investigator understanding (Ayoub et al., 2023). In the second stage, unsupervised pre-screening techniques such as Isolation Forest and anomaly detection algorithms are employed to identify suspicious providers and irregular billing patterns (Chandra & Arekapudi, 2025; Seo & Mendelevitch, 2017). These flagged anomalies are then passed to the generative AI-assisted review layer, where LLMs generate human-readable descriptions and hypotheses regarding potential fraud mechanisms, facilitating efficient triage and prioritization of cases. Subsequently, the framework incorporates a hybrid human-AI validation process, where fraud investigators and domain experts review flagged cases alongside AI-generated explanations. This human-in-the-loop approach enhances decision reliability and aligns with prior findings emphasizing the importance of expert validation in fraud detection systems (Van Capelleveen et al., 2016; Leevy et al., 2024). Following validation, the pipeline proceeds to supervised model training, where ensemble methods such as Random Forest and XGBoost, as well as graph-based approaches including Graph Neural Networks (GNNs), are employed to classify fraudulent claims (Farahmandazad & Danesh, 2025; Boyapati & Aygun, 2025). To ensure transparency and explainability, SHAP and LIME explainability techniques are integrated, with LLMs further translating technical outputs into regulatory-compliant narratives suitable for auditing and decision support (Kumaraswamy et al., 2024). The framework then incorporates a regulatory compliance layer, ensuring alignment with healthcare standards, including CMS guidelines and HIPAA requirements. Emerging approaches, including blockchain-based audit trails, can be integrated to enhance traceability and accountability of fraud detection decisions (Kennedy et al., 2023; Kumaraswamy et al., 2024).

To address real-world deployment challenges, FRAIG includes temporal evaluation and monitoring mechanisms, such as rolling cross-validation and concept drift detection, ensuring that models remain robust to

evolving fraud patterns over time (Johnson & Khoshgoftaar, 2023). Finally, the framework supports enterprise integration, enabling deployment within healthcare IT systems, alongside staff training and performance monitoring to ensure operational effectiveness (Hancock et al., 2024). This use case illustrates how FRAIG provides an end-to-end, adaptive, and regulation-aware solution for healthcare fraud detection, bridging the gap between advanced AI methodologies and practical implementation within healthcare systems.

5.6 Implementation Challenges and Future Directions

While the proposed FRAIG framework provides a comprehensive and adaptive architecture for AI-driven healthcare fraud detection, several implementation challenges must be addressed to enable its effective deployment in real-world healthcare systems.

5.6.1 Implementation Challenges

- **Data Privacy and Ethical Constraints:** The integration of generative AI raises significant concerns about data privacy, particularly when synthetic data is generated from sensitive healthcare records. Although techniques such as CT-GANs can preserve statistical properties while reducing direct exposure of patient data, ensuring compliance with privacy regulations such as HIPAA remains a critical requirement. Furthermore, safeguards must be implemented to prevent the misuse of synthetic data for fraudulent activities or re-identification risks (Kumaraswamy et al., 2024).

- **Computational Complexity and Scalability:** The FRAIG framework incorporates multiple computationally intensive components, including generative models, graph-based learning, and large language models. These elements may require substantial computational resources, particularly when applied to large-scale datasets such as Medicare claims. Prior studies have highlighted the limited reporting of computational performance, indicating a gap between methodological innovation and deployment feasibility (Hancock et al., 2024; Kennedy et al., 2023).

- **Integration with Legacy Healthcare Systems:** Healthcare organizations often rely on legacy IT infrastructures, which may not support seamless integration of advanced AI systems. Implementing FRAIG requires interoperability with existing claims processing systems, electronic health records (EHRs), and regulatory reporting platforms. This integration challenge is further complicated by organizational resistance to technological change and the need for staff training and capacity building.

- **Lack of Standardized Evaluation Protocols:** The systematic review identified significant heterogeneity in evaluation methodologies, including inconsistent use of validation techniques and performance metrics. The absence of standardized protocols, particularly temporal validation, limits the generalizability of AI models and poses a barrier to real-world adoption (Johnson & Khoshgoftaar, 2023).

- **Explainability and Regulatory Acceptance:** Although FRAIG incorporates an interpretability-first architecture, achieving regulatory acceptance remains a challenge. Healthcare regulators require transparent, auditable decision-making processes, and current explainability techniques may not fully meet these requirements. The integration of LLM-based explanations introduces additional concerns regarding reliability, consistency, and potential hallucination risks.

5.6.2 Future Research Directions

Future research should prioritize the development of standardized benchmarking protocols for healthcare fraud detection. This includes adopting temporal validation, external datasets, and consistent performance metrics to enable reliable cross-study comparisons and robust evaluation of frameworks such as FRAIG. While generative AI shows promise in addressing class imbalance and enhancing interpretability, its application in regulated healthcare environments requires further empirical validation. Future studies should evaluate the reliability, fairness, and compliance of generative models, particularly in high-stakes decision-making contexts. To address

data privacy concerns, future research should explore integrating federated learning and privacy-preserving techniques into the FRAIG framework. These approaches can enable collaborative model training across institutions without sharing sensitive patient data. Further investigation is needed into the design of effective human-AI collaboration mechanisms. This includes optimizing how fraud investigators interact with AI-generated explanations, as well as evaluating the impact of such systems on decision-making efficiency and accuracy. Empirical validation of FRAIG through pilot implementations in healthcare systems, particularly within CMS environments, is essential. Such deployments can provide insights into operational challenges, scalability, and return on investment, bridging the gap between research and practice. Future research should explore the development of adaptive systems capable of continuous learning from evolving fraud patterns. Generative AI offers opportunities for dynamic model updating and concept drift adaptation, enabling more resilient fraud detection systems.

6. Conclusion

This study provides a comprehensive synthesis of artificial intelligence approaches for healthcare fraud detection, highlighting both methodological advancements and persistent challenges across the literature. The findings demonstrate a clear evolution from traditional supervised learning models toward hybrid, deep learning, and graph-based approaches, with increasing attention to class imbalance handling and feature engineering. However, critical limitations remain, particularly in interpretability, temporal validation, and real-world deployment readiness. To address these gaps, this study proposes the FRAIG framework, a generative AI-enhanced sociotechnical architecture that integrates technical, institutional, and human dimensions of healthcare fraud detection. By embedding generative AI capabilities within the framework, FRAIG enables adaptive data augmentation, automated explanation generation, and improved alignment with regulatory requirements. This represents a shift from static AI models to dynamic, context-aware systems that can evolve with changing fraud patterns and institutional needs. The framework is grounded in empirical evidence from the systematic review and supported by a practical use case that demonstrates its applicability to real-world healthcare environments, such as Medicare claims analysis. Furthermore, integrating interpretability-first design principles and human-AI collaboration mechanisms enhances trust, transparency, and decision-making effectiveness in fraud detection systems. Despite its contributions, the study acknowledges several limitations, including the lack of standardized evaluation protocols across existing studies and the need for empirical validation of generative AI in regulated healthcare contexts. Future research should focus on benchmarking frameworks such as FRAIG using real-world datasets, incorporating temporal validation strategies, and exploring privacy-preserving techniques such as federated learning. Overall, this study advances the field of healthcare fraud detection by bridging the gap between AI innovation and practical implementation. The proposed FRAIG framework offers a scalable, adaptive, and regulation-aware solution that could support more effective and trustworthy fraud detection systems in healthcare.

Acknowledgement: Not Applicable

Funding: There is no funding available for this study

Conflicts of Interest: The author(s) have no conflict of interest to declare.

Consent to Publish declaration: Not applicable

Consent to Participate declaration: Not applicable

Declaration of Generative AI and AI-Assisted Tools Used: During the preparation of this work, the author(s) used Grammarly in order to improve language clarity and editing. After using this tool, the author(s) reviewed and edited the content as needed and take full responsibility for the publication's content.

Author Contributions:
Afsana Munni: Conceptualization, Resources; Data Curation, Investigation, Writing Draft

Md Ikram Hossain Bhuiyan: Conceptualization, Data Curation, Investigation, Writing Draft

MD Ashiqul Islam: Resources, Writing- Review & Editing;

*Taslim Uddin: Writing- Review & Editing, Revision;

Mohammad AL Mamun: Conceptualization, Data Curation, Investigation;

Fatema Tuz Zohora: Supervision, Writing- Review & Editing.

Ethical Approval: Not Applicable

References

1. Jones CH, Dolsten M. Healthcare on the brink: navigating the challenges of an aging society in the United States. NPJ Aging. 2024;10(1):22. https://doi.org/10.1038/s41514-024-00148-2
2. Jillo G. Advances and challenges in fraud detection in medical insurance [Internet]. SSRN; 2024. Available from: https://ssrn.com/abstract=4907327
3. Department of Justice. National health care fraud takedown results in 324 defendants charged in connection with over $14.6 billion in alleged fraud [Internet]. U.S. Department of Justice; 2025. Available from: https://www.justice.gov/opa/pr/national-health-care-fraud-takedown-results-324-defendants-charged-connection-over-146
4. Perumallaplli R. AI-powered enhancements in healthcare insurance fraud detection and prevention. SSRN; 2018. Available from: https://ssrn.com/abstract=5228599
5. Niu H, Omitaomu OA, Langston MA, Olama M, Ozmen O, Klasky HB, et al. EHR-BERT: a BERT-based model for effective anomaly detection in electronic health records. J Biomed Inform. 2024;150:104605. doi:10.1016/j.jbi.2024.104605
6. Sakil MBH, Hasan MA, Mozumder MSA, Hasan MR, Opee SA, Mridha MF, et al. Enhancing Medicare fraud detection with a CNN-transformer-XGBoost framework and explainable AI. IEEE Access. 2025;13:79609–22. 10.1109/ACCESS.2025.3562577
7. Rakesh M, D P S. Enhanced Medicare fraud detection using graph convolutional networks. In: Proc 4th Int Conf Artificial Intelligence and Signal Processing (AISP). 2024. 10.1109/AISP61711.2024.10870744
8. Page MJ, McKenzie JE, Bossuyt PM, et al. The PRISMA 2020 statement: an updated guideline for reporting systematic reviews. BMJ. 2021;372:n71. 10.1136/bmj.n71
9. du Preez A, Bhattacharya S, Beling P, Bowen E. Fraud detection in healthcare claims using machine learning: a systematic review. Artif Intell Med. 2025;160:103061. doi:10.1016/j.artmed.2025.103061
10. MacFarlane A, Russell-Rose T, Shokraneh F. Search strategy formulation for systematic reviews: issues, challenges and opportunities. Intell Syst Appl. 2022;15:200091. doi:10.1016/j.iswa.2022.200091
11. Ahmed K, Keskar NS, Socher R. Weighted transformer network for machine translation. arXiv preprint arXiv:1711.02132. 2017 Nov 6.
12. Nair AS. Publication bias—importance of studies with negative results. Indian J Anaesth. 2019;63(6):505–7. doi:10.4103/ija.IJA_70_19
13. Hong QN, Pluye P, Fàbregues S, Bartlett G, Boardman F, Cargo M, et al. Mixed methods appraisal tool (MMAT), version 2018. https://doi.org/10.3233/EFI-180221
14. Van Capelleveen G, Poel M, Mueller RM, Thornton D, Van Hillegersberg J. Outlier detection in healthcare fraud: a case study in the Medicaid dental domain. Int J Account Inf Syst. 2016;21:18–31. doi:10.1016/j.accinf.2016.04.001
15. Bauder RA, Khoshgoftaar TM, Hasanin T. An empirical study on class rarity in big data. In: Proc IEEE ICMLA. 2018. doi:10.1109/ICMLA.2018.00127
16. Bounab R, Zarour K, Guelib B, Khlifa N. Enhancing Medicare fraud detection through machine learning: addressing class imbalance with SMOTE-ENN. IEEE Access. 2024;12:54382–96. 10.1109/ACCESS.2024.3385781
17. Farahmandazad D, Danesh K. ML-driven approaches to combat Medicare fraud [Internet]. 2025. Available from: https://arxiv.org/abs/2502.15898
18. Prova NNI. Healthcare fraud detection using machine learning. In: Proc Int Conf Intelligent Cyber Physical Systems. 2024. 10.1109/ICoICI62503.2024.10696476
19. Hamid Z, Khalique F, Mahmood S, Daud A, Bukhari A, Alshemaimri B. Healthcare insurance fraud detection using data mining. BMC Med Inform Decis Mak. 2024;24:112. doi:10.1186/s12911-024-02433-2
20. Seo J, Mendelevitch O. Identifying frauds and anomalies in the Medicare-B dataset. IEEE EMBC. 2017;2017:3664–7. doi:10.1109/EMBC.2017.8037539
21. Boyapati M, Aygun R. BalancerGNN: balancer graph neural networks for imbalanced datasets. Neural Netw. 2025;182:106926. doi:10.1016/j.neunet.2024.106926

22. Mardani S, Moradi H. Using graph attention networks in healthcare provider fraud detection. IEEE Access. 2024;12:132786–800. 10.1109/ACCESS.2024.3425892
23. Yoo Y, Shin J, Kyeong S. Medicare fraud detection using graph analysis. IEEE Access. 2023;11:88278–94. 10.1109/ACCESS.2023.3305962
24. SK S, Ilango V. A time-efficient model for detecting fraudulent health insurance claims. In: Proc ICSCAN. 2020. 10.1109/ICSCAN49426.2020.9262298
25. Chandra N, Arekapudi V. Detecting fraud, waste, and abuse in healthcare claims using AI. 2025. https://doi.org/10.36287/setsci.24.3.023
26. Kennedy KR, Khoshgoftaar TM. Impact of class imbalance on unsupervised label generation for medicare fraud detection. In2024 IEEE International Conference on Information Reuse and Integration for Data Science (IRI) 2024 Aug 7 (pp. 216-221). IEEE. 10.1109/IRI62200.2024.00053
27. Bauder RA, Herland M, Khoshgoftaar TM. Evaluating model predictive performance: a Medicare fraud detection case study. In: Proc IEEE IRI. 2019. doi:10.1109/IRI.2019.00014
28. Hancock J, Khoshgoftaar TM, Johnson JM. Informative evaluation metrics for highly imbalanced big data classification. In: Proc IEEE ICMLA. 2022. doi:10.1109/ICMLA55696.2022.00220
29. Nugraha R, Pardede H, Subekti A. Oversampling based on generative adversarial networks to overcome imbalanced data in predicting fraud insurance claims. Kuwait J Sci. 2022. 10.48129/kjs.splml.19119
30. Kocak B, Klontzas ME, Stanzione A, Meddeb A, Demircioğlu A, Bluethgen C, Bressem K, Ugga L, Mercaldo N, Díaz O, Cuocolo R. Evaluation metrics in medical imaging AI: fundamentals, pitfalls, misapplications, and recommendations. European journal of radiology artificial intelligence. 2025 Jul 8:100030. https://doi.org/10.1016/j.ejrai.2025.100030
31. Kennedy RKL, Johnson JM, Khoshgoftaar TM. The effects of class label noise on highly imbalanced big data. In: Proc IEEE ICTAI. 2021. doi:10.1109/ICTAI52525.2021.00215
32. Karthik KP, Gupta TS, Kaushik D, Ramesh TK. Analysis of ML models and data balancing techniques for Medicare fraud using PySpark. In: Proc ICAECT. 2025. 10.1109/ICAECT63952.2025.10958842
33. Herland M, Khoshgoftaar TM, Bauder RA. Big data fraud detection using multiple Medicare data sources. J Big Data. 2018;5:29. doi:10.1186/s40537-018-0148-0
34. Fattahi M, Moattar MH, Forghani Y. Improved cost-sensitive representation of data for solving the imbalanced big data classification problem. Journal of Big Data. 2022 May 3;9(1):60. https://doi.org/10.1186/s40537-022-00617-z
35. Zhai J, Wang M, Zhang S. Binary imbalanced big data classification based on fuzzy data reduction and classifier fusion. Soft Computing. 2022 Mar;26(6):2781-92. https://doi.org/10.1007/s00500-021-06654-9
36. Johnson J, Khoshgoftaar T. Deep learning and thresholding with class-imbalanced big data. 2019. 10.1109/ICMLA.2019.00134
37. Johnson JM, Khoshgoftaar TM. Deep learning and data sampling with imbalanced big data. In: Proc IEEE IRI. 2019. doi:10.1109/IRI.2019.00033
38. Kumaraswamy N, Markey M, Barner J, Rascati K. Feature engineering to detect fraud using healthcare claims data. Expert Syst Appl. 2022;210. doi:10.1016/j.eswa.2022.118941
39. Liang Q, Bauder RA, Khoshgoftaar TM. Enhancing Medicare fraud detection: random undersampling followed by SHAP-driven feature selection with big data. In: Proc IEEE ICTAI. 2024. 10.1109/ICTAI62512.2024.00045
40. Hancock J, Khoshgoftaar T. Medicare fraud detection using CatBoost. 2020. 10.1109/IRI49571.2020.00022
41. Bauder RA, Khoshgoftaar TM. A probabilistic programming approach for outlier detection in healthcare claims. In: Proc IEEE ICMLA. 2016. doi:10.1109/ICMLA.2016.0064
42. Herland M, Bauder RA, Khoshgoftaar TM. Approaches for identifying U.S. Medicare fraud in provider claims data. Health Care Manag Sci. 2020;23(1):2–19. doi:10.1007/s10729-019-09496-2
43. Leevy JL, Hancock J, Khoshgoftaar TM. Assessing one-class and binary classification approaches for identifying Medicare fraud. In: Proc IEEE IRI. 2023. doi:10.1109/IRI58017.2023.00054
44. Fourkiotis K, Tsadiras A. Using boosting and neural networks methods to detect healthcare fraud. 2024. https://doi.org/10.1007/978-3-031-63219-8_12
45. Shi H, Tayebi MA, Pei J, Cao J. Cost-sensitive learning for medical insurance fraud detection with temporal information. IEEE Trans Knowl Data Eng. 2023;35(10):10451–63. doi:10.1109/TKDE.2022.3179773
46. John A, Ahsun A. Explainable AI (XAI) for Fraud Detection: Building Trust and Transparency in AI-Driven Financial Security Systems Authors. Available at SSRN 5285281. 2025 Jun 7. http://dx.doi.org/10.2139/ssrn.5285281
47. Kumaraswamy N, Ekin T, Park C, Markey MK, Barner JC, Rascati K. Using a Bayesian belief network to detect healthcare fraud. Expert Syst Appl. 2024;238:122241. doi:10.1016/j.eswa.2023.122241
48. Hancock JT, Wang H, Khoshgoftaar TM, Liang Q. Data reduction techniques for highly imbalanced Medicare big data. J Big Data. 2024;11:8. doi:10.1186/s40537-023-00779-2

49. Kennedy RKL, Salekshahrezaee Z, Villanustre F, Khoshgoftaar TM. Iterative cleaning and learning of big highly imbalanced fraud data using unsupervised learning. J Big Data. 2023;10:106. doi:10.1186/s40537-023-00755-w
50. Wang H, Hancock JT, Khoshgoftaar TM. Improving Medicare fraud detection through big data size reduction techniques. In: Proc IEEE SOSE. 2023. doi:10.1109/SOSE59344.2023.00037
51. Hancock JT, Khoshgoftaar TM, Johnson JM. Evaluating classifier performance with highly imbalanced big data. J Big Data. 2023;10:42. doi:10.1186/s40537-023-00730-5
52. Leevy JL, Salekshahrezaee Z, Khoshgoftaar TM. A review of unsupervised anomaly detection techniques for health insurance fraud. In: Proc IEEE BigDataService. 2024. 10.1109/BigDataService62917.2024.00028
53. Razzaq K, Shah M. Next-generation machine learning in healthcare fraud detection: Current trends, challenges, and future research directions. Information. 2025 Aug 25;16(9):730. https://doi.org/10.3390/info16090730
54. Johnson JM, Khoshgoftaar TM. Data-centric AI for healthcare fraud detection. SN Comput Sci. 2023;4(4):389. doi:10.1007/s42979-023-01777-0
55. Zhang X, Nutakor F, Minlah MK, Li J. Can digital transformation drive green transformation in manufacturing companies? Sustainability. 2023;15(3). doi:10.3390/su15032345
56. Akhavan A, Jalali MS. Generative AI and simulation modeling: how should you (not) use large language models like ChatGPT [Internet]. SSRN; 2023. Available from: https://ssrn.com/abstract=4675409
57. Ibrar W, Mahmood D, Al-Shamayleh AS, Ahmed G, Alharthi SZ, Akhunzada A. Generative AI: a double-edged sword in the cyber threat landscape. Artif Intell Rev. 2025;58(9):285. doi:10.1007/s10462-025-10786-2
58. Schünemann C, Johanning S, Herold H, Bruckner T. Modelling behaviour in socio-technical systems for policy assessment. J Clean Prod. 2024;462:142722. doi:10.1016/j.jclepro.2024.142722
59. Roszkowska-Menkes M. Institutional theory. In: Encyclopedia of Sustainable Management. Springer; 2023. doi:10.1007/978-3-031-25984-5_389
60. Jobin A, Ienca M, Vayena E. The global landscape of AI ethics guidelines. Nat Mach Intell. 2019;1. doi:10.1038/s42256-019-0088-2
61. Göndöcs D, Horváth S, Dörfler V. Uncovering the dynamics of human-AI hybrid performance. Int J Hum Comput Stud. 2025. https://doi.org/10.1016/j.ijhcs.2025.103622
62. Tyagi AK, Seranmadevi R. Blockchain for enhancing security and privacy in the smart healthcare. Digital Twin and Blockchain for Smart Cities. 2024 Oct 15:343-70. https://doi.org/10.1002/9781394303564.ch16
63. Ayoub S, Gulzar Y, Rustamov J, Jabbari A, Reegu FA, Turaev S. Adversarial approaches to tackle imbalanced data in machine learning. Sustainability. 2023;15(9). doi:10.3390/su15095732
64. Saarela M, Podgorelec V. Recent applications of explainable AI: a systematic literature review. Appl Sci. 2024;14(19). doi:10.3390/app14198654
65. Kapoor S, Narayanan A. Leakage and the reproducibility crisis in machine-learning-based science. Patterns. 2023;4(9):100804. doi:10.1016/j.patter.2023.100804

Appendix A: Search Strings and Results

Web of science	
TS=("artificial intelligence" OR "AI" OR "machine learning" OR "deep learning" OR "predictive analytics" OR "natural language processing" OR "generative AI" OR "LLM" OR "large language models") AND TS=("healthcare fraud" OR "medical fraud" OR "claim fraud" OR "billing fraud" OR "insurance fraud" OR "health insurance fraud") AND TS=("detection" OR "prevention" OR "identification" OR "mitigation" OR "prediction") AND TS=("United States" OR "US" OR "USA")	12
healthcare fraud, medicare fraud , machine learning , AI , LLM (smart search)	100 result

SCOPUS	
TITLE-ABS-KEY-AUTH (machine learning in Medicare fraud)	72 results
TITLE-ABS-KEY (("artificial intelligence" OR "AI" OR "machine learning" OR "deep learning") AND ("healthcare fraud" OR "medical fraud") AND ("detection" OR "prevention") AND ("United States" OR "US"))	11 result

PubMed	
("artificial intelligence"[tiab] OR "ai"[tiab] OR "machine learning"[tiab] OR "deep learning"[tiab] OR "predictive analytics"[tiab] OR "natural language processing"[tiab] OR "generative ai"[tiab] OR "llm"[tiab] OR "large language model"[tiab]) AND ("healthcare fraud"[tiab] OR "medical fraud"[tiab] OR "claim fraud"[tiab] OR "billing fraud"[tiab] OR "insurance fraud"[tiab]) AND ("detection"[tiab] OR "prevention"[tiab] OR "identification"[tiab] OR "mitigation"[tiab]) AND ("united states"[tiab] OR "us"[tiab] OR "usa"[tiab])	5 result
(Machine learning in Medicare fraud)	8 result

IEE Explore	
("artificial intelligence" OR "ai" OR "machine learning" OR "deep learning" OR "predictive analytics" OR "natural language processing" OR "generative ai" OR "llm") AND ("healthcare fraud" OR "medical fraud" OR "claim fraud" OR "billing fraud") AND ("detection" OR "prevention" OR "identification") AND ("us" OR "united states")	10 results
(Machine learning in Medicare fraud)	52 results

Google Scholar underwent several search iterations.

1. "artificial intelligence" "healthcare fraud" "detection" "united states"