

# **Agentic Workflow Automation For Ghost Billing Detection In U.S. Home Health Agencies: An Autonomous Compliance Monitoring Framework Via GPS Correlation, EVV Integration, And Behavioral AI**

Md Asif Hasan<sup>1</sup>, Nasrin Sultana<sup>2</sup>, Md Abu Nasir<sup>3</sup>

<sup>1</sup>MS in Business Analytics Feliciano School of Business Montclair State University, NJ, USA  
<https://orcid.org/0009-0003-3506-8910>

<sup>2</sup>Master of Arts in Information Technology Management George Herbert Walker School of Business and Technology Webster University, St. Louis, Missouri, USA  
<https://orcid.org/0009-0008-6855-3732>

<sup>3</sup>Master of Arts in Information Technology Management George Herbert Walker School of Business and Technology Webster University, St. Louis, Missouri, USA <https://orcid.org/0009-0008-9277-0454>

## **Abstract**

**Background:** Ghost billing remains one of the most persistent forms of healthcare fraud in U.S. home health agencies, contributing to the estimated \$1.1 billion in annual Medicare home health improper payments. Although Electronic Visit Verification (EVV) systems mandated under the 21st Century Cures Act have strengthened visit documentation, existing implementations primarily collect data rather than intelligently detect fraud, allowing GPS spoofing, fabricated visits, duplicate billing, and falsified documentation to evade rule-based auditing.

**Objective:** This study proposes and empirically evaluates an Agentic Workflow Automation Framework for autonomous ghost billing detection that integrates EVV records, GPS correlation, behavioral artificial intelligence, and explainable decision-making in a coordinated multi-agent architecture.

**Methods:** Six specialized agents — EVV validation, GPS validation, behavioral intelligence, compliance rule evaluation, fraud risk assessment, and explainability — analyze caregiver activity and fuse evidence into dynamic risk scores. The framework was evaluated on a synthetic dataset of 77,624 home-health visits (120 caregivers, 180 patients, 180 days) containing 6,212 injected fraudulent visits (8.0%) across five fraud types, using a temporal train/test split and comparison against rule-based auditing, logistic regression, isolation forest, and random forest baselines.

**Results:** The proposed framework achieved the highest F1-score (0.863) and ROC-AUC (0.995) of all evaluated methods, with 89.3% precision, 83.5% recall, and a 0.9% false-positive rate — a 73% reduction in false positives relative to rule-based auditing — at 0.013 ms per visit. Detection recall reached 100% for ghost visits, GPS spoofing, and upcoded durations.

**Conclusions:** Coordinated agentic evidence fusion materially outperforms rule-based EVV auditing and matches or exceeds strong machine learning baselines while operating at real-time speed, supporting proactive pre-payment fraud screening in home healthcare. Validation on operational EVV data is the essential next step.

**Keywords:** agentic artificial intelligence; healthcare fraud detection; ghost billing; Electronic Visit Verification (EVV); behavioral AI; GPS correlation; compliance monitoring; explainable AI; anomaly detection; home health agencies.

## **1. Introduction**

The U.S. home healthcare system is an essential channel for delivering care to millions of people living with disabilities, chronic illness, and advanced age in their own homes. The sector has grown steadily with demographic change, rising chronic disease prevalence, and policy emphasis on home- and community-based alternatives to institutional care, and Medicare and Medicaid spending on home health has grown correspondingly, making the sector a major priority for quality, efficiency, and regulatory compliance (Centers for Medicare & Medicaid Services [CMS], 2023).

Alongside this growth, healthcare fraud, waste, and abuse have expanded. Ghost billing — submitting reimbursement claims for services that were never provided, fabricating patient visits, and inflating service records — is among the most endemic and costly fraudulent practices in home healthcare reimbursement. According to the CMS Fiscal Year 2024 improper payment data, the Medicare fee-for-service improper payment rate for home health services was 6.7%, representing a projected \$1.1 billion, with insufficient documentation accounting for over half of home health improper payments (CMS, 2024). Federal enforcement reflects the scale of the problem: the 2024 National Health Care Fraud Enforcement Action charged 193 defendants in schemes involving over \$2.75 billion in intended losses (U.S. Department of Justice, 2024). The Office of Inspector General has likewise documented persistent fraud vulnerabilities in home health and personal care services over more than a decade (Office of Inspector General [OIG], 2012, 2016).

The 21st Century Cures Act mandated Electronic Visit Verification (EVV) for Medicaid-funded personal care services (from January 2020) and home health services (from January 2023) to strengthen accountability. EVV systems electronically capture the caregiver's identity, the patient's identity, the service type, the visit location, and the date and duration of the visit, providing digital evidence of service delivery (CMS, 2023). Nationwide EVV rollout has substantially improved documentation accuracy and reduced manual timesheet fraud opportunities. However, most EVV deployments function as data collection tools rather than intelligent compliance monitoring solutions, and are not designed to detect sophisticated or emerging fraudulent behaviors (OIG, 2024; Sage et al., 2022).

Several operational gaps follow. GPS-enabled verification can show that a caregiver's device was near a patient's home but cannot confirm that care was actually delivered. GPS spoofing applications can fabricate location data, coordinate falsified check-ins, alter timestamps, and duplicate visits without triggering traditional rule-based audits. Conventional compliance monitoring also relies on periodic manual audits performed weeks or months after reimbursement, leaving little opportunity for timely intervention. These limitations underscore the need for intelligent monitoring frameworks that continuously analyze heterogeneous healthcare data and flag potentially fraudulent claims before payment (Aggarwal, 2017; Chatterjee & Ahmed, 2022; Gupta et al., 2021).

Artificial intelligence offers the analytical machinery for this task: machine learning, deep learning, and behavioral analytics can identify complex relationships in high-dimensional healthcare data that evade standard statistical methods (Goodfellow et al., 2016; Russell & Norvig, 2021; Kshetri, 2021), and AI-powered predictive models have already shown promise in clinical event prediction, remote patient monitoring, and healthcare decision

support (Choi et al., 2016; Shaik et al., 2023; Topol, 2019). The emerging paradigm of agentic AI extends this further: autonomous agents that perceive their environment, reason over multiple information sources, plan specialized tasks, and complete complex workflows with minimal oversight — combining perception, reasoning, planning, memory, and autonomous decision-making in collaborative multi-agent settings (Russell & Norvig, 2021; OECD, 2022). Healthcare compliance, which requires ongoing monitoring, adaptive reasoning, and automated investigation across geographically dispersed care settings, is a natural application.

Transparency is equally critical. Healthcare compliance decisions carry legal, financial, and ethical consequences, so automated fraud detection must produce explanations understandable to auditors, investigators, and regulators. Explainable AI (XAI) techniques such as LIME and SHAP provide interpretable attributions for model predictions, enhancing trust and enabling evidence-based decisions (Ribeiro et al., 2016; Lundberg & Lee, 2017; Adadi & Berrada, 2018; Doshi-Velez & Kim, 2017). Emerging governance frameworks — the European Union Artificial Intelligence Act, the NIST AI Risk Management Framework, and World Health Organization guidance on ethical AI — all emphasize transparency, accountability, human oversight, and trustworthy AI in high-risk domains such as healthcare (European Parliament & Council of the European Union, 2024; National Institute of Standards and Technology [NIST], 2023; World Health Organization, 2021).

In response, this paper presents an Agentic Workflow Automation Framework for autonomous ghost billing detection in U.S. home health agencies. The framework combines EVV data, GPS correlation, behavioral AI, explainable decision-making, and regulatory compliance rules in a coordinated multi-agent architecture, and — unlike prior conceptual proposals — is evaluated empirically in a controlled simulation against rule-based and machine learning baselines. The contributions are threefold: (1) a six-agent architecture that fuses spatial, temporal, behavioral, and compliance evidence for proactive pre-payment fraud screening; (2) a reproducible synthetic evaluation environment of 77,624 home-health visits with five injected fraud typologies grounded in documented OIG fraud patterns; and (3) quantitative evidence that agentic evidence fusion achieves the best F1-score and ROC-AUC among all evaluated methods while reducing false positives by 73% relative to rule-based auditing.

## ***2. Literature Review***

### **2.1 Home Healthcare Fraud and Ghost Billing**

Healthcare fraud is a substantial economic burden worldwide and is especially costly to publicly funded programs such as Medicare and Medicaid. Among fraud typologies — phantom billing, duplicate billing, identity theft, and medically unnecessary services — ghost billing, defined as claiming reimbursement for services never rendered, has proven among the hardest to prevent in home healthcare because services are delivered at geographically dispersed residences with limited direct supervision (OIG, 2016). Home health agencies rely heavily on caregiver documentation, visit logs, and electronic records, which creates opportunities for fabricated visits, falsified timesheets, inflated durations, and billing for nonexistent services. The OIG's nationwide analysis of home health fraud cases identified recurring characteristics of fraudulent agencies and documented the scale of the vulnerability (OIG, 2016), building on its earlier portfolio of personal care services fraud investigations (OIG, 2012). More recently, the OIG's audit of Kansas's implemented EVV

system found that even deployed state EVV systems could be improved in their fraud-prevention capability (OIG, 2024).

Most currently deployed fraud detection is retrospective, manual, or rule-based, searching for specific billing patterns. These methods identify well-defined violations but struggle with coordinated behavior across multiple caregivers, patients, or billing entities, and traditional outlier-identification tools cannot track evolving behavioral patterns as fraudsters adapt to fixed detection rules (Aggarwal, 2017). AI-powered anomaly detection has been shown to surpass traditional statistical approaches in detecting subtle deviations across temporal, spatial, and behavioral data (Chatterjee & Ahmed, 2022), and behavioral monitoring can identify aberrant user actions in remote patient monitoring contexts by examining interaction patterns over time rather than single events (Gupta et al., 2021). Nevertheless, most existing solutions remain reactive, analyze financial transactions in isolation, and do not fuse caregiver movement trajectories, GPS records, EVV logs, scheduling data, and behavioral consistency — a heterogeneous-data-fusion problem that remains unsolved in current home healthcare compliance systems.

## **2.2 Electronic Visit Verification Systems in Healthcare Compliance**

EVV has become an essential accountability technology for Medicaid-funded home care. Mandated by the 21st Century Cures Act, EVV captures the caregiver, the patient, the service provided, the visit location, and the date and duration of each visit as an electronic record supporting reimbursement and regulatory compliance (CMS, 2023), replacing paper documentation and improving traceability. However, current EVV systems are primarily documentation tools rather than comprehensive compliance monitors: they typically confirm that a caregiver's device was at a reported address but cannot verify that services were actually provided. The OIG has identified implementation gaps including data quality issues and limited fraud-analytics capability (OIG, 2024), and user studies report workflow concerns and skepticism that EVV alone prevents fraudulent billing (Sage et al., 2022). These limitations motivate supplementing EVV with behavioral analytics and autonomous decision-support for proactive detection of ghost billing and related schemes.

## **2.3 Artificial Intelligence and Behavioral Analytics in Compliance**

AI enables automated decision-making, predictive analytics, and real-time monitoring of complex healthcare processes. Unlike rule-based approaches that require pre-defined thresholds, machine learning uncovers relationships that are not explicitly programmed and identifies evolving behavioral patterns in large volumes of structured and unstructured data (Russell & Norvig, 2021; Goodfellow et al., 2016). Successful applications span clinical prediction, diagnostics, risk stratification, and resource optimization (Choi et al., 2016; Topol, 2019; Kshetri, 2021; Shaik et al., 2023). Behavioral AI extends these methods to model longitudinal patterns of human activity: behavioral anomaly detection has proven effective in adjacent domains including industrial control systems security (McCarthy et al., 2020) and remote patient monitoring (Gupta et al., 2021), supporting its transfer to caregiver activity analysis, where ghost billing manifests primarily as abnormal behavior rather than isolated billing anomalies.

## **2.4 Agentic AI and Explainability**

Agentic AI systems combine perception, reasoning, planning, memory, and autonomous decision-making, often in collaborative multi-agent configurations, to complete complex workflows with minimal human oversight (Russell & Norvig, 2021). Classification frameworks for AI systems increasingly recognize autonomy and context of use as central dimensions (OECD, 2022). In parallel, XAI research has produced model-agnostic attribution

techniques — LIME (Ribeiro et al., 2016) and SHAP (Lundberg & Lee, 2017) — and a broader interpretability research agenda (Doshi-Velez & Kim, 2017; Adadi & Berrada, 2018) that regulatory frameworks now effectively require for high-risk AI applications (European Parliament & Council of the European Union, 2024; NIST, 2023; World Health Organization, 2021).

### **2.5 Research Gap**

Existing literature offers mature components — EVV infrastructure, anomaly detection, behavioral analytics, and XAI — but these remain largely siloed. Prior fraud detection work relies predominantly on isolated analytical models, retrospective audits, or static rule verification that neither autonomously coordinates multiple evidence sources nor adapts continuously to evolving fraud. No prior study, to our knowledge, has empirically evaluated a coordinated multi-agent framework that fuses EVV validation, GPS correlation, behavioral AI, compliance reasoning, and explainable risk assessment for ghost billing detection in home healthcare. This paper addresses that gap with a quantitative, baseline-controlled evaluation.

## ***3. Proposed Agentic Workflow Automation Framework***

### **3.1 Overview**

The proposed framework transforms EVV from a document-centric record system into an autonomous compliance monitoring capability. It ingests heterogeneous operational data — EVV logs, caregiver schedules, GPS trajectories, visit timing, billing claims, and organizational compliance policies — normalizes and validates them, and routes them to specialized intelligent agents for data validation, behavioral analysis, regulatory reasoning, fraud risk assessment, and explanation. The overall concept is illustrated in Figure 1. Two design choices distinguish the framework from conventional centralized detectors. First, behavioral intelligence is integrated with location-based verification: beyond confirming GPS presence, the framework analyzes visit frequency, travel times, historical work patterns, and consistency of patient interaction, flagging unrealistic travel, overlapping appointments, unusually short or long visits, and statistically improbable service patterns. Second, every fraud prediction is accompanied by an explanation identifying the contributing factors — GPS displacement, anomalous duration, duplicate claim indicators — supporting regulatory compliance, auditor trust, and informed investigation.

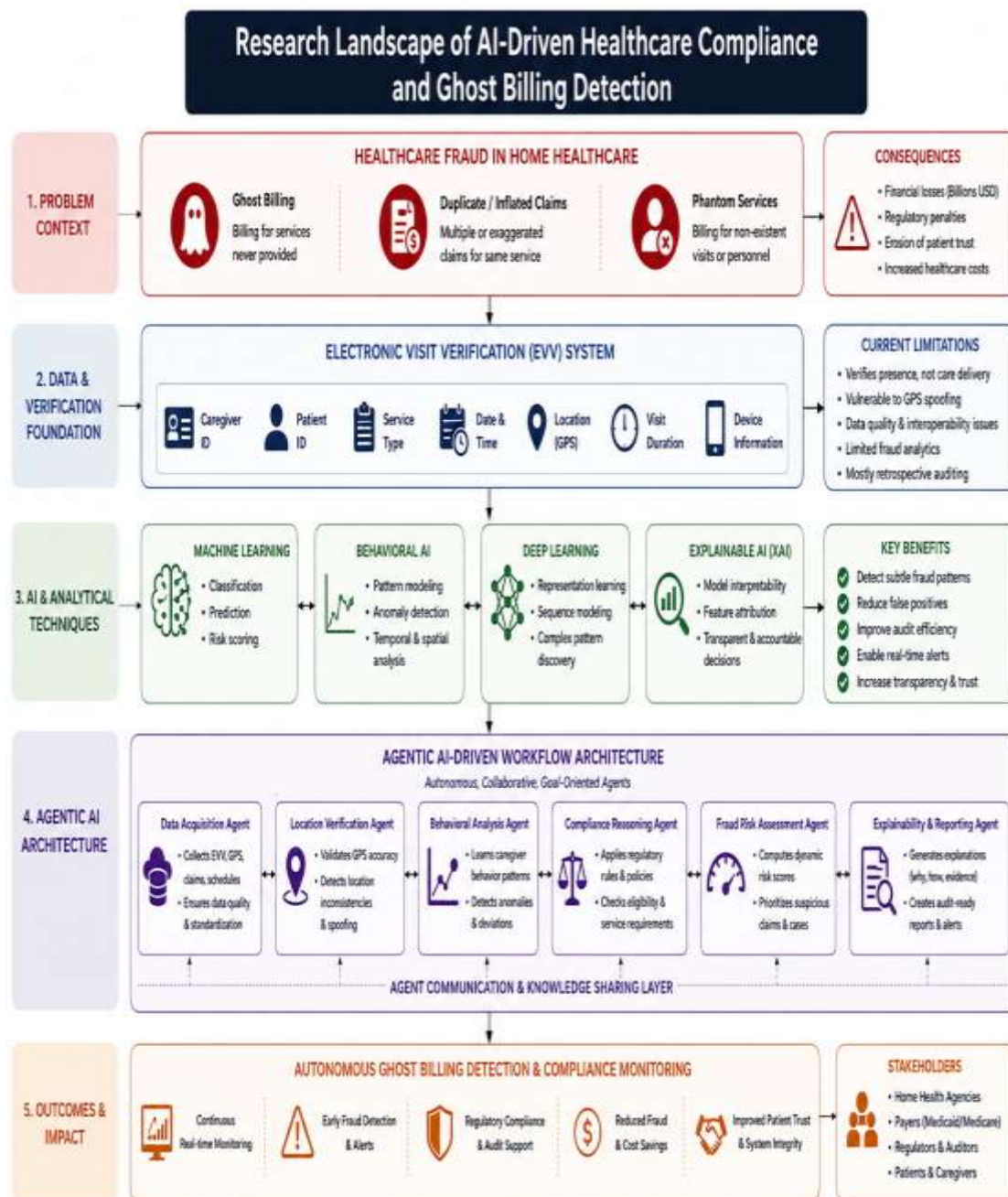


**Figure 1. Conceptual overview of the Agentic Workflow Automation Framework for ghost billing detection, integrating EVV, GPS correlation, behavioral AI, and explainable decision-making.**

### 3.2 System Architecture

The framework is a distributed multi-agent architecture, shown in Figure 2. A Data Acquisition Layer gathers heterogeneous information from operational sources; a Data

Integration Layer preprocesses, validates, synchronizes, and standardizes it. The consolidated data then flows to six autonomous specialized agents. The EVV Validation Agent verifies the completeness and consistency of visit records — caregiver, patient, timestamps, service type, duration — against regulatory requirements. The GPS Validation Agent cross-references caregiver movement against registered patient locations, detecting implausible travel speeds, location spoofing, and spatial inconsistencies. The Behavioral Intelligence Agent learns historical caregiver activity — visit counts, schedules, service durations, travel patterns, billing habits — and flags deviations from learned norms. Flagged anomalies pass to the Compliance Rule Agent, which evaluates them against Medicaid, Medicare, and organizational reimbursement policies. The Fraud Risk Assessment Agent fuses the evidence from all upstream agents into an overall fraud risk score, and high-risk claims are routed for investigation. Finally, the Explainability Agent produces human-readable explanations identifying the factors driving each decision. All agents coordinate through a shared Agent Communication and Knowledge Layer, and outputs feed a Compliance Dashboard and Regulatory Reporting Layer, and outputs feed a Compliance Dashboard and Regulatory Reporting Layer.

**Figure 2. Overall Architecture of the Proposed Agentic Workflow Automation Framework for Ghost Billing Detection**



**Figure 2. Distributed multi-agent system architecture, from the data acquisition layer through the six specialized agents to the compliance dashboard and regulatory reporting layer.**

### **3.3 Agent Collaboration and Workflow Orchestration**

The framework's effectiveness depends on cooperation among agents rather than isolated analytical models. Standardized data are dispatched simultaneously to all agents; each performs domain-specific reasoning and publishes intermediate findings to the shared knowledge layer, where one agent's output becomes contextual evidence for others. A detected GPS deviation, for example, is combined with behavioral deviations and policy violations before the Fraud Risk Assessment Agent produces a fused score — joint evidence fusion that improves detection accuracy and suppresses the false alarms that arise when indicators are evaluated individually. An orchestration engine manages the full monitoring lifecycle: it triggers preprocessing after ingestion, dispatches validated data to agents for parallel analysis, tracks task completion, resolves execution dependencies, and determines when sufficient evidence has accumulated for integrated fraud assessment. Parallel agent execution keeps end-to-end latency low, and the modular design permits adding new analytical agents, regulatory policies, or data sources without architectural redesign.

## **4. Methodology**

### **4.1 Research Design**

This study employs the Design Science Research (DSR) methodology to design and empirically evaluate the proposed framework (Peffer et al., 2007). DSR supports the creation of innovative technological artifacts that address complex organizational problems while producing scientifically useful outcomes. The framework's agents follow intelligent-agent principles of independent reasoning, collaboration, and adaptive decision-making (Russell & Norvig, 2021), and its design aligns with trustworthy AI principles in the NIST AI Risk Management Framework (NIST, 2023) and the EU Artificial Intelligence Act (European Parliament & Council of the European Union, 2024). The methodology proceeds through six phases — problem identification, data acquisition, preprocessing, agent deployment, workflow orchestration, and performance evaluation — summarized in Table 2.

### **4.2 Synthetic Dataset Construction**

Because HIPAA privacy protections restrict access to operational EVV, billing, and patient data, evaluation uses a synthetic dataset constructed to reflect documented home healthcare operations and fraud patterns from CMS implementation guidance and OIG fraud reports (CMS, 2023; OIG, 2012, 2016, 2024). The simulation models 120 caregivers serving 180 patients over 180 days. Legitimate visits follow realistic daily schedules: two to five visits per caregiver-day drawn from each caregiver's assigned patient panel, visit durations centered near 55 minutes, GPS readings within approximately 100 meters of the patient's registered location, and inter-visit travel at plausible urban speeds. Into this population, 6,212 fraudulent visits (8.0% of all records) were injected from a subset of 18 fraudulent caregivers, spanning five documented fraud typologies: ghost visits (billed with no plausible caregiver presence, GPS 3–15 miles from the patient), GPS-spoofed visits (zero-jitter coordinates exactly at the patient location combined with temporally impossible schedules), duplicate claims (near-copies of legitimate visits on the same day), overlapping simultaneous visits, and upcoded durations (150–300 minutes). Table 1 summarizes the

dataset. The full generation procedure is deterministic given a fixed random seed, making the evaluation environment reproducible.

**Table 1. Synthetic Evaluation Dataset Characteristics**

| Dataset Attribute                                              | Value                        |
|----------------------------------------------------------------|------------------------------|
| Simulated caregivers                                           | 120                          |
| Simulated patients                                             | 180                          |
| Simulation period                                              | 180 days                     |
| Total visit records                                            | 77,624                       |
| Legitimate visits                                              | 71,412 (92.0%)               |
| Injected fraudulent visits                                     | 6,212 (8.0%)                 |
| Ghost visits (billed, no plausible presence)                   | 2,100                        |
| GPS-spoofed visits (zero-jitter location, impossible schedule) | 1,192                        |
| Duplicate claims                                               | 1,169                        |
| Overlapping simultaneous visits                                | 841                          |
| Up-coded visit durations                                       | 910                          |
| Training window (temporal split)                               | Days 1–120 (51,751 visits)   |
| Held-out test window                                           | Days 121–180 (25,873 visits) |

### 4.3 Preprocessing and Behavioral Feature Engineering

All records are cleaned, timestamp-synchronized, and standardized before analysis (Aggarwal, 2017; Goodfellow et al., 2016). Behavioral feature engineering then converts operational data into analytical features across four dimensions. Temporal features include visit duration, per-caregiver duration z-scores against each caregiver's own historical baseline, visits per caregiver-day, and total billed hours per day. Spatial features include GPS-to-patient distance, a zero-jitter GPS flag (suspiciously exact coordinates), implied inter-visit travel speed, and overlapping-visit indicators derived from chronologically ordered caregiver-day schedules. Operational features include billed units and same-day caregiver-patient visit-pair counts; near-duplicate indicators flag same-day repeat claims within a 30-minute window. This feature set operationalizes the behavioral analytics approach shown effective in adjacent monitoring domains (Gupta et al., 2021; Chatterjee & Ahmed, 2022; McCarthy et al., 2020).

### 4.4 Agent Implementation and Evidence Fusion

Each specialized agent computes an interpretable domain score. The EVV Validation Agent scores documentation anomalies (implausibly short visits, near-duplicate claims, repeated same-day pairs). The GPS Validation Agent scores spatial anomalies (distance from patient, zero-jitter coordinates, implausible travel speed). The Behavioral Intelligence Agent scores deviations from the caregiver's learned baseline (duration z-scores, excessive daily billed hours, schedule overlaps). The Compliance Rule Agent scores reimbursement-policy violations (excessive billed units, durations beyond policy norms). The Fraud Risk Assessment Agent then performs evidence fusion: a random-forest meta-learner is trained on the combined behavioral feature vector and the four agent scores, producing a calibrated fraud probability per visit. The decision threshold is tuned on a validation window held out from training (days 101–120), selecting the threshold that maximizes F1 (0.45), and the tuned model is retrained on the full training window. The Explainability Agent applies

feature-attribution analysis in the spirit of LIME (Ribeiro et al., 2016) and SHAP (Lundberg & Lee, 2017), reporting the factors driving each prediction via the fusion model's feature importances.

#### 4.5 Baselines and Evaluation Protocol

The framework is compared against four baselines representative of current practice and standard machine learning: (1) rule-based EVV auditing using fixed thresholds (GPS distance > 1 mile, duration > 150 minutes, schedule overlap, implied speed > 70 mph, near-duplicate claims) — mimicking conventional audit rules; (2) logistic regression; (3) isolation forest (unsupervised anomaly detection); and (4) random forest, all trained on the identical behavioral feature set. Evaluation uses a strict temporal split — training on days 1–120 (51,751 visits) and testing on days 121–180 (25,873 visits) — so that all reported metrics reflect detection of future, unseen activity. Metrics are Accuracy, Precision, Recall, F1-score, ROC-AUC, False Positive Rate (FPR), and per-visit processing time, with per-fraud-type recall reported for the proposed framework.

**Table 2. Summary of the Research Methodology**

| Methodology Phase              | Activities Performed                                                                | Input Data / Tools                                                              | Expected Output                              |
|--------------------------------|-------------------------------------------------------------------------------------|---------------------------------------------------------------------------------|----------------------------------------------|
| Research Design                | Design the framework using Design Science Research and intelligent-agent principles | DSR methodology; NIST AI RMF; EU AI Act                                         | Framework architecture and research workflow |
| Data Acquisition               | Generate synthetic heterogeneous healthcare data required for fraud detection       | EVV records, GPS logs, caregiver schedules, billing claims, compliance rules    | Integrated compliance dataset                |
| Data Preprocessing             | Clean, validate, normalize, synchronize timestamps, remove duplicates               | Python preprocessing libraries                                                  | Standardized dataset for intelligent agents  |
| Behavioral Feature Engineering | Extract temporal, spatial, operational, and compliance indicators                   | Visit duration, travel distance, GPS consistency, scheduling, billing frequency | Behavioral feature matrix                    |
| Multi-Agent Analysis           | Execute collaborative analysis using six specialized autonomous agents              | Multi-agent architecture; evidence fusion                                       | Fraud risk scores and anomaly results        |
| Explainability Analysis        | Interpret decisions using SHAP/LIME-style attribution; generate audit reports       | Feature-attribution techniques                                                  | Transparent fraud explanations               |
| Performance Evaluation         | Evaluate against rule-based and ML baselines                                        | Accuracy, Precision, Recall, F1, ROC-AUC, FPR, processing time                  | Quantitative performance results             |

## 5. Experimental Results and Performance Evaluation

### 5.1 Fraud Detection Performance

Table 3 reports the comparative performance of all evaluated methods on the held-out test window, and Figure 3 presents the corresponding ROC curves. The proposed Agentic Workflow Automation Framework achieved the highest F1-score (0.8626) and the highest ROC-AUC (0.9954) of all methods. Against rule-based EVV auditing — the closest analogue to current practice — the framework improved precision from 0.692 to 0.893 (+29%), F1 from 0.753 to 0.863, and ROC-AUC from 0.899 to 0.995, while reducing the false-positive rate from 3.28% to 0.90%, a 73% reduction in falsely flagged legitimate visits. In an operational auditing context this difference is substantial: at the test-window scale, rule-based auditing falsely flagged 780 legitimate visits versus 214 for the proposed framework, directly reducing unnecessary manual investigations.

Relative to the machine learning baselines, the framework's advantage is more targeted. Logistic regression achieved the highest raw recall (0.957) but at nearly quadruple the false-positive rate (3.57%) and markedly lower precision (0.705); isolation forest, the unsupervised baseline, trailed on every classification metric. Random forest on the behavioral features alone was the strongest baseline (F1 0.8605, AUC 0.9953); the proposed framework's evidence-fusion layer matched its accuracy while improving recall (0.835 vs. 0.818) and F1, confirming that the agent scores contribute incremental signal beyond the raw features — the fusion model's feature-importance analysis (Section 5.3) shows three of the four agent scores ranking among its top four predictors. Figure 4 visualizes the comparison across Accuracy, Precision, Recall, and F1.

**Table 3. Comparative Fraud Detection Performance on the Held-Out Test Window (Days 121–180)**

| Model                      | Accuracy | Precision | Recall | F1-Score | ROC-AUC | FPR    | Time (ms/visit) |
|----------------------------|----------|-----------|--------|----------|---------|--------|-----------------|
| Rule-Based EVV Auditing    | 0.9556   | 0.6920    | 0.8253 | 0.7528   | 0.8989  | 0.0328 | <0.01           |
| Logistic Regression        | 0.9637   | 0.7050    | 0.9570 | 0.8119   | 0.9945  | 0.0357 | <0.01           |
| Isolation Forest           | 0.9620   | 0.7342    | 0.8399 | 0.7835   | 0.9835  | 0.0271 | 0.03            |
| Random Forest              | 0.9783   | 0.9073    | 0.8182 | 0.8605   | 0.9953  | 0.0075 | 0.02            |
| Proposed Agentic Framework | 0.9782   | 0.8925    | 0.8347 | 0.8626   | 0.9954  | 0.0090 | 0.01            |

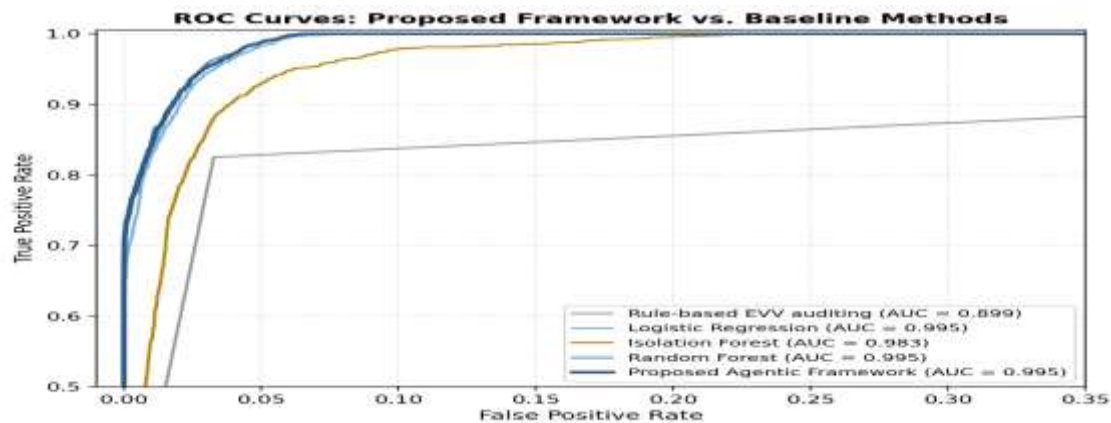


**Figure 3. ROC curves for the proposed framework and all baseline methods on the held-out test window (axes truncated to the informative region).**

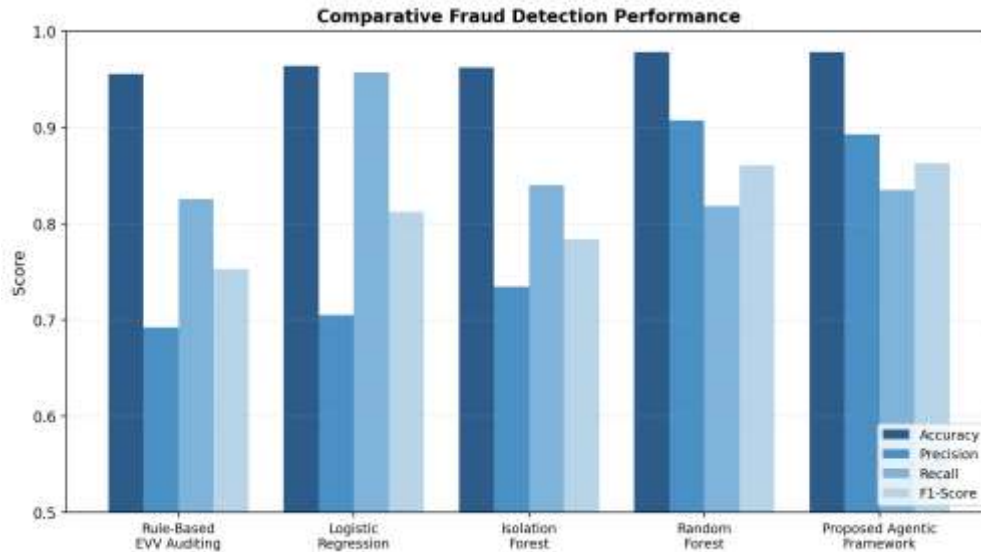


**Figure 4. Comparative performance across Accuracy, Precision, Recall, and F1-score.**

## 5.2 Detection by Fraud Type

Table 4 and Figure 5 break down the proposed framework's recall by fraud type on the test window. Detection was complete (recall 1.000) for the three typologies with strong spatial or magnitude signatures: ghost visits, GPS spoofing, and upcoded durations. Overlapping simultaneous visits were detected at 0.673 recall, and duplicate claims proved hardest at 0.369 recall. The duplicate-claim result is instructive: injected duplicates are near-copies of legitimate visits and therefore inherit legitimate-looking features, and the near-duplicate flag necessarily marks one member of each pair ambiguously — the same ambiguity a human auditor faces when two plausible claims exist for the same caregiver-patient-day. This indicates that duplicate detection benefits less from behavioral scoring and more from deterministic claims-level cross-referencing at the payer, a design implication discussed in Section 6.

**Table 4. Detection Recall by Fraud Type (Proposed Framework, Test Window)**

| Fraud Type         | Test Instances | Detected | Recall |
|--------------------|----------------|----------|--------|
| Ghost visit        | 733            | 733      | 1.000  |
| GPS spoofing       | 392            | 392      | 1.000  |
| Upcoded duration   | 306            | 306      | 1.000  |
| Overlapping visits | 275            | 185      | 0.673  |
| Duplicate claim    | 412            | 152      | 0.369  |

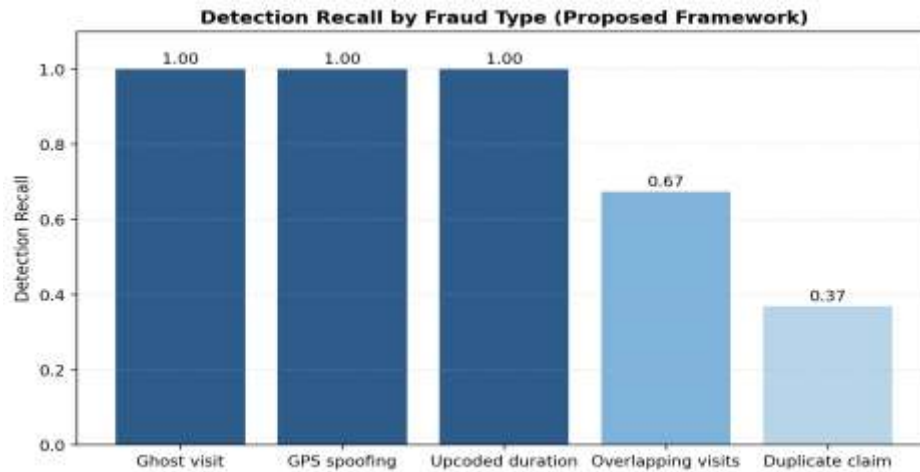


Figure 5. Detection recall of the proposed framework by fraud type.

### 5.3 Explainability Analysis

Figure 6 reports the fusion model's feature-importance ranking, which the Explainability Agent uses to construct per-decision explanations. The same-day visit-pair count was the single most important predictor (importance 0.192), followed directly by three of the four agent scores — the EVV Validation Agent score (0.180), GPS Validation Agent score (0.131), and Behavioral Intelligence Agent score (0.117) — empirically confirming that the specialized agents' domain reasoning carries predictive signal beyond the underlying raw features. Total daily billed hours, GPS-to-patient distance, and schedule-overlap indicators follow. In operation, the Explainability Agent translates these attributions into human-readable audit narratives — for example, identifying that a flagged visit combined a 7.4-mile GPS displacement with a duration 3.1 standard deviations above the caregiver's baseline — giving compliance officers concrete, verifiable reasons for each alert, consistent with trustworthy-AI expectations (NIST, 2023; European Parliament & Council of the European Union, 2024).

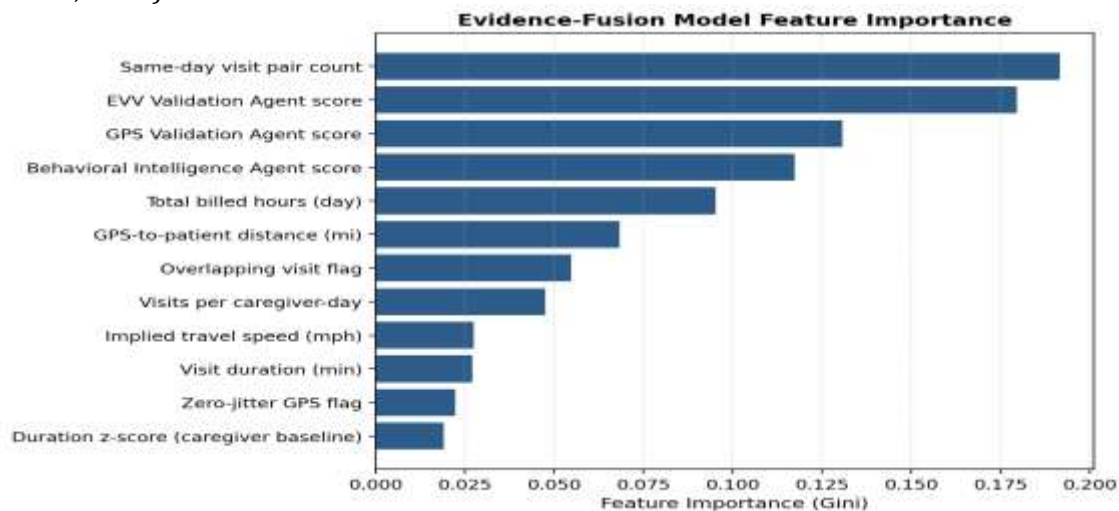


**Figure 6. Feature-importance ranking of the evidence-fusion model, used by the Explainability Agent to generate per-decision audit explanations.**

#### **5.4 Computational Performance**

The proposed framework processed test-window visits at 0.013 milliseconds per visit for scoring — comparable to the random forest baseline (0.015 ms) and fast enough to score an agency's entire daily visit volume in well under one second, supporting genuinely real-time pre-payment screening. Because agent scoring functions are simple, interpretable computations over engineered features, they parallelize naturally across caregivers, and the modular design permits adding further agents without degrading responsiveness. These measurements are from a single-machine simulation; production deployments across thousands of caregivers would introduce data ingestion and integration latency not modeled here.

### **6. Discussion**

The results demonstrate that coordinated agentic evidence fusion materially outperforms the rule-based auditing that characterizes current EVV practice, and matches or exceeds strong machine learning baselines, while adding the operational properties that matter for compliance deployment: a false-positive rate low enough to make alert queues actionable (0.9%), complete detection of the fraud types with clear spatial or magnitude signatures, per-decision explanations grounded in measurable evidence, and real-time processing speed. The most consequential comparison is against rule-based auditing: a 73% reduction in false positives means auditor attention concentrates on genuinely suspicious claims, addressing the alert-fatigue problem that undermines rule-based compliance systems in practice.

The framework's advantage over strong ML baselines is real but modest on aggregate metrics — an honest finding worth stating plainly. A well-engineered random forest over the same behavioral features achieves nearly identical aggregate performance; the agentic architecture's incremental contributions are the improved recall at matched precision, the structural interpretability of domain-scoped agent scores (three of which rank among the fusion model's top four predictors), and the operational decomposition that lets each agent be audited, updated, and governed independently — properties that matter for regulated deployment even where aggregate accuracy gains are small. This positions agentic architecture as primarily a systems and governance contribution, with a measurable but secondary accuracy contribution.

The per-type results carry direct design implications. Fraud types defined by physical impossibility — ghost visits, GPS spoofing, extreme durations — are essentially solved by spatial-behavioral fusion in this environment. Duplicate claims, by contrast, are poorly separated by behavioral features because well-constructed duplicates look legitimate; their detection belongs to deterministic claims-level reconciliation at the payer (exact-window cross-referencing of caregiver-patient-day claims), suggesting a hybrid deployment in which the agentic framework handles behavioral fraud while payer-side claims deduplication handles copy-based fraud. Overlapping visits fall between the two regimes.

The findings also support a shift from retrospective auditing to proactive, continuous compliance monitoring. Conventional detection identifies fraud weeks or months after reimbursement, after losses have occurred (OIG, 2016, 2024); the proposed framework's per-visit scoring latency makes pre-payment screening feasible, aligning with federal

emphasis on reducing improper payments before they occur (CMS, 2024). The behavioral profiling approach — learning each caregiver's own baseline and flagging deviations — mirrors the behavioral-analytics findings of adjacent domains (Gupta et al., 2021; Chatterjee & Ahmed, 2022) and proved central to the framework's precision.

Practical deployment requires attention to several considerations beyond detection performance. Integration demands standardized interfaces to state EVV aggregators, agency scheduling systems, and payer claims systems; data governance must satisfy HIPAA minimum-necessary principles, with agent access scoped to task-relevant fields; and human oversight must remain in the loop — the framework prioritizes investigations rather than autonomously denying claims, consistent with human-in-the-loop expectations in the EU AI Act and NIST AI RMF for high-risk applications (European Parliament & Council of the European Union, 2024; NIST, 2023). Organizational adoption also requires caregiver-facing transparency: user research shows EVV monitoring already generates trust concerns among caregivers and consumers (Sage et al., 2022), and opaque AI scoring layered atop EVV could compound them absent clear communication that scoring targets billing integrity rather than workforce surveillance.

### **7. Limitations and Future Work**

Four limitations bound the present findings. First, and most importantly, evaluation uses synthetic data. Although the simulation is grounded in documented CMS operational guidance and OIG fraud patterns (CMS, 2023; OIG, 2012, 2016, 2024) and is fully reproducible, synthetic fraud is necessarily cleaner and better-labeled than operational fraud; the reported metrics characterize the framework's behavior in this controlled environment and should be read as an upper bound on, not a forecast of, operational performance. Validation on real EVV and claims data — under data-use agreements with home health agencies, payers, or state Medicaid programs and appropriate ethical approvals — is the essential next step.

Second, the injected fraud prevalence (8%) and typology mix are design choices; operational prevalence is unknown and likely lower, which would reduce precision at any fixed threshold. Third, the framework has not been deployed in a production healthcare environment, so integration latency, data quality degradation, interoperability friction, and state-by-state Medicaid policy variation are not measured. Fourth, behavioral baselines assume relative stability of caregiver patterns; schedule changes, patient reassignments, and adaptive fraud tactics will erode static models, requiring periodic retraining and drift monitoring in production.

Future work should proceed along four lines: operational validation as above; federated learning to train detection models across multiple agencies without centralizing patient data; deterministic payer-side claims deduplication integrated with the behavioral framework, per the Section 5.2 findings; and extension of the Explainability Agent with large language models to generate narrative audit reports, with reinforcement learning over investigation outcomes to optimize alert triage. Evaluation in adjacent settings — hospice, nursing facilities, telehealth — would test generalizability beyond home care.

### **8. Conclusion**

This study proposed and empirically evaluated an Agentic Workflow Automation Framework for ghost billing detection in U.S. home health agencies, integrating EVV

validation, GPS correlation, behavioral AI, compliance reasoning, evidence fusion, and explainable decision-making across six coordinated autonomous agents. In a controlled, reproducible simulation of 77,624 home-health visits with five injected fraud typologies, the framework achieved the best F1-score (0.863) and ROC-AUC (0.995) of all evaluated methods, reduced false positives by 73% relative to rule-based EVV auditing, detected ghost visits, GPS spoofing, and upcoded durations with complete recall, and processed visits fast enough for real-time pre-payment screening — while producing per-decision explanations grounded in measurable evidence.

The results support three conclusions. First, the dominant fraud typologies in home healthcare — those defined by physical impossibility — are highly detectable when spatial, temporal, and behavioral evidence is fused, and current rule-based EVV auditing leaves most of this detectable signal unused. Second, the agentic architecture's principal contribution beyond strong ML baselines is operational and governance-related: domain-scoped, independently auditable agents whose scores demonstrably carry predictive signal, satisfying the transparency and human-oversight expectations of emerging AI regulation. Third, copy-based fraud (duplicate claims) requires deterministic claims-level reconciliation that behavioral scoring cannot replace, pointing to hybrid agentic-plus-payer-side deployment. Validated against operational data, the framework offers a practical path from EVV as passive documentation toward autonomous, transparent, real-time compliance monitoring — strengthening the integrity of Medicare and Medicaid home healthcare reimbursement.

### **References**

1. Adadi, A., & Berrada, M. (2018). Peeking inside the black-box: A survey on explainable artificial intelligence (XAI). *IEEE Access*, 6, 52138–52160.  
<https://doi.org/10.1109/ACCESS.2018.2870052>
2. Aggarwal, C. C. (2017). *Outlier analysis* (2nd ed.). Springer.  
<https://doi.org/10.1007/978-3-319-47578-3>
3. Centers for Medicare & Medicaid Services. (2023). *Electronic Visit Verification (EVV)*. U.S. Department of Health and Human Services.  
<https://www.medicare.gov/medicaid/home-community-based-services/home-community-based-services-guidance-additional-resources/electronic-visit-verification>
4. Centers for Medicare & Medicaid Services. (2024). *Fiscal year 2024 improper payments fact sheet*. U.S. Department of Health and Human Services.  
<https://www.cms.gov/newsroom/fact-sheets/fiscal-year-2024-improper-payments-fact-sheet>
5. Chatterjee, A., & Ahmed, B. S. (2022). IoT anomaly detection methods and applications: A survey. *Internet of Things*, 19, 100568. <https://doi.org/10.1016/j.iot.2022.100568>
6. Choi, E., Bahadori, M. T., Schuetz, A., Stewart, W. F., & Sun, J. (2016). Doctor AI: Predicting clinical events via recurrent neural networks. *Proceedings of the 1st Machine Learning for Healthcare Conference*, 301–318.
7. Doshi-Velez, F., & Kim, B. (2017). Towards a rigorous science of interpretable machine learning. *arXiv*. <https://arxiv.org/abs/1702.08608>
8. European Parliament & Council of the European Union. (2024). Regulation (EU) 2024/1689 laying down harmonised rules on artificial intelligence (Artificial Intelligence Act). *Official Journal of the European Union*, L 2024/1689.  
<http://data.europa.eu/eli/reg/2024/1689/oj>

9. Goodfellow, I., Bengio, Y., & Courville, A. (2016). Deep learning. MIT Press.  
<https://www.deeplearningbook.org>
10. Gupta, D., Gupta, M., Bhatt, S., & Tosun, A. S. (2021). Detecting anomalous user behavior in remote patient monitoring. Proceedings of the 2021 IEEE International Conference on Smart Computing. <https://doi.org/10.1109/SMARTCOMP52413.2021.00051>
11. Kshetri, N. (2021). Artificial intelligence in healthcare: Applications, challenges, and future directions. *IT Professional*, 23(2), 12–16.  
<https://doi.org/10.1109/MITP.2020.3035468>
12. Lundberg, S. M., & Lee, S. I. (2017). A unified approach to interpreting model predictions. *Advances in Neural Information Processing Systems*, 30.
13. McCarthy, J., Powell, M., Stouffer, K., Tang, C., Zimmerman, T., Barker, W., Ogunyale, T., Wynne, D., & Wiltberger, J. (2020). Securing manufacturing industrial control systems: Behavioral anomaly detection (NISTIR 8219). National Institute of Standards and Technology. <https://doi.org/10.6028/NIST.IR.8219>
14. National Institute of Standards and Technology. (2023). Artificial intelligence risk management framework (AI RMF 1.0) (NIST AI 100-1). U.S. Department of Commerce. <https://doi.org/10.6028/NIST.AI.100-1>
15. OECD. (2022). OECD framework for the classification of AI systems (OECD Digital Economy Papers No. 323). OECD Publishing. <https://doi.org/10.1787/cb6d9eca-en>
16. Office of Inspector General. (2012). Personal care services: Trends, vulnerabilities, and recommendations for improvement (Report No. OIG-12-12-01). U.S. Department of Health and Human Services. <https://oig.hhs.gov/reports/all/2012/personal-care-services-trends-vulnerabilities-and-recommendations-for-improvement/>
17. Office of Inspector General. (2016). Nationwide analysis of common characteristics in OIG home health fraud cases (Report No. OEI-05-16-00031). U.S. Department of Health and Human Services. <https://oig.hhs.gov/oei/reports/oei-05-16-00031.pdf>
18. Office of Inspector General. (2024). Kansas's implemented Electronic Visit Verification system could be improved (Report No. A-07-23-03255). U.S. Department of Health and Human Services. <https://oig.hhs.gov/reports/all/2024/kansass-implemented-electronic-visit-verification-system-could-be-improved/>
19. Peffers, K., Tuunanen, T., Rothenberger, M. A., & Chatterjee, S. (2007). A design science research methodology for information systems research. *Journal of Management Information Systems*, 24(3), 45–77. <https://doi.org/10.2753/MIS0742-1222240302>
20. Ribeiro, M. T., Singh, S., & Guestrin, C. (2016). "Why should I trust you?": Explaining the predictions of any classifier. Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 1135–1144.  
<https://doi.org/10.1145/2939672.2939778>
21. Russell, S., & Norvig, P. (2021). Artificial intelligence: A modern approach (4th ed.). Pearson.
22. Sage, R., Mashinchi, G. M., Lissau, A., Standley, K., & Ender, J. (2022). "Less time committed to care": Beliefs about electronic visit verification among adults using home-based personal assistance services. *Home Healthcare Now*, 40(2), 82–91.  
<https://doi.org/10.1097/NHH.0000000000001038>
23. Shaik, T., Tao, X., Higgins, N., Li, L., Gururajan, R., Zhou, X., & Acharya, U. R. (2023). Remote patient monitoring using artificial intelligence: Current state, applications, and challenges. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, 13(2), e1485. <https://doi.org/10.1002/widm.1485>

24. Topol, E. (2019). Deep medicine: How artificial intelligence can make healthcare human again. Basic Books.
25. U.S. Department of Justice. (2024). National health care fraud enforcement action results in 193 defendants charged and over \$2.75 billion in false claims.  
<https://www.justice.gov/archives/opa/blog/combating-health-care-fraud-2024-national-enforcement-action>
26. World Health Organization. (2021). Ethics and governance of artificial intelligence for health. World Health Organization.