



Proactive Fraud Detection in Healthcare billing through Open Source Intelligence

Venkata renu kumar MJain-School of Sciences
Bangalore, India**Priyanksha Das**Assistant Professor
Jain-School of Sciences**Vishnu Venkatesh**Assistant professor
Jain school of sciences

Abstract

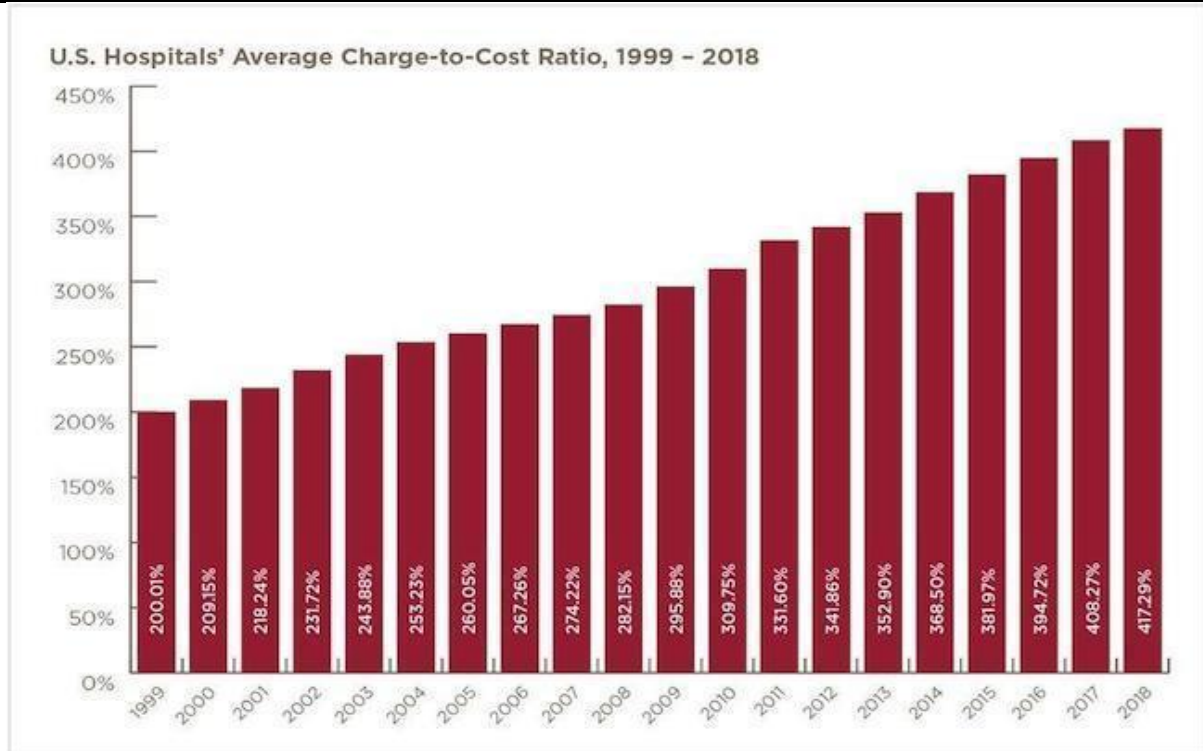
Healthcare (medical) field is the one of the most respected field and the healthcare billing is a kind of grown over decades. The healthcare billing fraud is a serious threat to healthcare systems particularly in the point unwanted charges and overbilling, resulting in financial losses. Although the regular fraud detection system is working usually reactive and detects anomalies only after significant damage has been happened. This project suggests a proactive method for health care fraud detection utilizing Open Source Intelligence (OSINT) tools like Maltego, SpiderFoot, Google Dorking and other. These techniques are used to map healthcare organizations, reveal public data breaches, find exposed billing systems, and harvest metadata from web documents. Public billing information is gathered, scraped, and processed using machine learning methods, developed algorithm which help to identify automatically anomalies like overbilling and charges that are not expected to occur. Usually most of the healthcare organizations are anonymized for ethical disclosure, and billing anomalies are matched against domain vulnerabilities to estimate total risk and the findings are the proof that OSINT platforms can considerably improve early fraud detection capacity, with more than 80% accuracy in early detection of the suspicious transactions. By identifying anomalies earlier, this project provides a frame work for enhancing the financial integrity, stability and security of the healthcare industry. This frame work can be further developed in future work for patient related data manipulation and insider fraud detection to ensure better protection of healthcare organisation.

Keywords : Healthcare fraud billing, over billing, Open source intelligence tools, Maltego, Spiderfoot, google dorking, data breaches .

1. Introduction

1.1 Overview of Healthcare Fraud

Healthcare fraud refers to deliberate misrepresentation or deception for financial or personal gain within the healthcare system (Centers for Medicare & Medicaid Services, 2022). It encompasses various fraudulent practices, including upcoding, phantom billing, duplicate claims, and unbundling services. Upcoding involves billing for a more expensive service than provided (Office of Inspector General, 2022), while phantom billing refers to invoicing for services never rendered. Overbilling, the most prevalent form, occurs when healthcare providers intentionally inflate the quantity or complexity of services performed, or charge for unnecessary procedures and tests, thus draining healthcare funds and burdening patients with needless interventions (Federal Bureau of Investigation, 2023). These actions not only inflate healthcare costs but distort resource allocation, leading to inefficiency in healthcare delivery and eroding public trust. Combatting healthcare fraud is vital for preserving financial resources and maintaining the accessibility and credibility of healthcare services globally.



1.2 Different types of fraud billing

Healthcare billing fraud takes various forms, each exploiting specific aspects of the claims process. Overbilling refers to charging more than appropriate for healthcare services, either by exaggerating the services provided, billing for services not rendered, or manipulating billing processes to maximize financial gain. Understanding these schemes is critical for implementing effective counter measures.

1.2.1 Upcoding

Upcoding involves assigning a higher-paying billing code to a service than was actually performed. (OIG, 2022)

Ex. A provider may bill for a comprehensive consultation when only a basic visit occurred. This fraudulent practice inflates reimbursement amounts and contributes significantly to financial losses in healthcare systems.

1.2.2 Phantom Billing

Phantom billing entails submitting claims for services never rendered or for fictitious patients.

Ex. In one case, a group of providers billed for physical therapy sessions that patients never attended, resulting in millions of dollars in fraudulent reimbursements.

1.2.3 Duplicate claims

Duplicate claims occur when a provider submits the same claim multiple times, often with minor alterations to bypass automated detection.

Ex. a clinic might re-submit claims for a patient's routine check-up under slightly varied coding formats to secure multiple payments.

1.2.4 Unbundling Services

Unbundling involves splitting a single procedure into separate components and billing each as a distinct service. Ex. a surgery that includes anaesthesia and postoperative care might be billed as three separate procedures instead of one comprehensive package, leading to overpayment.

Fraud Type	Description	Example
Upcoding	Billing for a higher-paying service than provided	Charging for a full physical exam when only a basic checkup occurred
Phantom Billing	Submitting claims for services never rendered	Billing for therapy sessions that patients did not attend
Duplicate Claims	Submitting the same claim multiple times	Resubmitting a routine checkup claim with minor modifications
Unbundling	Separating services into individual claims to maximize payment	Billing anesthesia, surgery, and postoperative care as separate procedures

Table no 1: different types of healthcare frauds in billing

1.3 Role of OSINT in Healthcare fraud billing detection

Open Source Intelligence (OSINT) tools plays an increasingly critical role in detection (Miller,2021) and investigating healthcare billing fraud. OSINT involves gathering and analyzing publicly available information from diverse sources such as licensing boards, government registries, news articles, business databases, and social media platforms. By leveraging OSINT, investigators can cross-verify the credentials, affiliations, and operational activities of healthcare providers and institutions to uncover inconsistencies suggestive of fraudulent billing practices.

Specialized OSINT tools like **Maltego** and **Spiderfoot** are particularly valuable in healthcare fraud investigations.

- **Maltego** enables the mapping of complex relationships between individuals, clinics, and businesses by visualizing connections based on shared addresses, email domains, phone numbers, and corporate ownership. This helps in detecting hidden networks involved in collusive billing schemes or phantom provider setups.
- **Spiderfoot** automates the collection and correlation of data from hundreds of online sources. It can identify suspicious attributes like multiple business registrations tied to a single provider, domain ownership patterns, and leaked credentials, which may indicate illicit activities.
- **Social media search engines** and monitoring tools also play a vital role. By examining providers' or clinics' online presence on platforms like LinkedIn, Facebook, or Instagram, investigators can uncover evidence such as employment histories, clinic closures, or lifestyle indicators inconsistent with reported operations. For example, a provider billing for full-time services may post travel photos during working hours, suggesting fraudulent claims.(Holtfreter,2018)
- Search engines dedicated to professional verification, such as license lookup portals, and social media-specific OSINT tools (e.g., Social Searcher) enhance the ability to track down misrepresented qualifications or non-existent clinics.

By the application of OSINT with traditional anomaly detection systems and AI-driven algorithms meeting the legal standard for organizations to build a multi-layered defence against healthcare fraud. However, it remains essential to balance thorough investigation practices with ethical standards and data privacy regulations to maintain stakeholder trust and legal compliance.

2. Materials and Methods

2.1 Mapping Healthcare Organizations Using OSINT Tools : To map healthcare institutions and identify potential fraud exposure, we employed OSINT tools like **Maltego** and **SpiderFoot**. Maltego was used to discover relationships between hospitals, their billing partners, domains, and internal email addresses through passive reconnaissance. SpiderFoot automated the discovery of open ports, leaked databases, misconfigured servers, and leaked medical documents from public sources. Google Dorking assisted in manually searching for exposed invoices and billing records using advanced search queries (e.g., "site:hospital.org filetype:csv invoice"). The mapped data helped visualize healthcare networks and identify areas vulnerable to overbilling fraud or data leaks. Entities such as domains, emails, and IP addresses were plotted into a relationship graph, highlighting risky connections. This information created a strong foundation for later machine learning analysis of billing anomalies.



Fig 1 maltego mapping the healthcare organisations

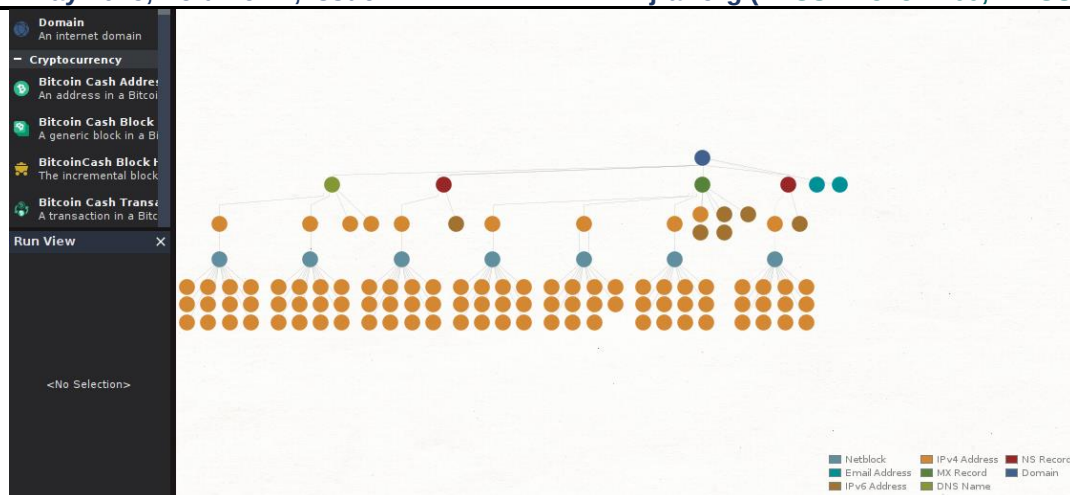


Fig 2 maltego mapping the organisation and linked nodes

2.2 Machine learning for anomaly detection: After gathering healthcare billing datasets through OSINT, we applied machine learning models to detect billing anomalies in the suspected fraud billing.CSV file of different organisation. The data, comprising invoice amounts, service types, and billing frequencies, was pre-processed to remove noise and missing values. We used the ai-algorithm, an unsupervised learning method ideal for outlier detection in high-dimensional datasets. The model was trained to recognize normal billing patterns and flag unusual activities such as multiple billing entries for a single service, excessive service charges, and phantom billing entries. Each detected anomaly was correlated with its corresponding organization based on the domain mapping from Objective 1.By combining OSINT-driven data with unsupervised learning, the system could proactively detect fraud without needing labelled data. The output was a list of high-risk billing records for further investigation, providing a scalable, automated fraud detection framework.

Metric	Allowed Amount	Charged Amount	Paid Amount	Standardized
Mean	\$101.43	\$354.55	\$77.36	\$78.03
Median	\$65.10	\$146.00	\$47.02	\$47.84
Max	\$20,494.00	\$62,694.00	\$16,067.30	\$16,957.15
Min	\$0.01	\$0.01	\$0.0087	\$0.0087
Std Dev	\$257.24	\$1,062.61	\$199.72	\$200.05

Table no2 : Anomalous billing entries were flagged on normal behaviour. On average, providers charge ~3.5x more than what Medicare actually allows.

2.3 Correlating Billings Anomalies with Cyber Vulnerabilities : Beyond detecting overbilling anomalies, the project aimed to correlate fraudulent billing behaviour with domain-level vulnerabilities. Using SpiderFoot, we gathered vulnerability reports on hospital domains, such as outdated SSL certificates, exposed administrative portals, and unprotected billing pages. These vulnerabilities were mapped against organizations showing anomalous billing activities. Statistical correlation methods were applied to understand whether higher vulnerability levels matched higher fraud risks. In most cases, organizations with leaked billing systems or misconfigured domains were more likely to display overbilling anomalies. This linking strategy helped strengthen the credibility of fraud findings, offering deeper insights into not just "who" is vulnerable but "why" certain billing irregularities happen. It also highlighted the urgent need for better cyber hygiene practices in the healthcare sector.

```
[INFO] sflib : sfp_ipstack module loaded.
[INFO] sflib : sfp_clearbit module loaded.
[INFO] sflib : sfp_bitcoin module loaded.
[INFO] sflib : sfp_leakix module loaded.
[INFO] sflib : sfp_pgp module loaded.
[INFO] sflib : sfp_quad9 module loaded.
[INFO] sflib : sfp_apple_itunes module loaded.
[INFO] sflib : sfp_grep_app module loaded.
[INFO] sflib : sfp_cinsscore module loaded.
[INFO] sflib : sfp_cleantalk module loaded.
[INFO] sflib : sfp_slideshare module loaded.
[INFO] sflib : sfp_dnscommonsrv module loaded.
[INFO] sflib : sfp_hosting module loaded.
[INFO] sflib : sfp_debounce module loaded.
[INFO] sflib : sfp_botscout module loaded.
[INFO] sflib : sfp_similar module loaded.
[INFO] sflib : sfp_networksdb module loaded.
[INFO] sflib : sfp_koodous module loaded.
[INFO] sflib : sfp_opendns module loaded.
[INFO] sflib : sfp_opencorporates module loaded.
[INFO] sflib : sfp_nameapi module loaded.
[INFO] sflib : sfp_iknowwhatyoudownload module loaded.
[INFO] sflib : sfp_azureblobstorage module loaded.
[INFO] sflib : sfp_viewdns module loaded.
[INFO] sflib : sfp_adguard_dns module loaded.
[INFO] sflib : sfp_ipapicom module loaded.
[INFO] sflib : sfp_spyonweb module loaded.
[INFO] sflib : sfp_instagram module loaded.
[INFO] sflib : sfp_fsecure_riddler module loaded.
[INFO] sflib : sfp_webanalytics module loaded.
[INFO] sflib : sfp_truemail module loaded.
[INFO] sflib : sfp_whois module loaded.
```

fig3 screenshot data leaked using spiderfoot

3. Results and Discussion

Org ID	Key Findings	Risk Level
Org 1	A South Carolina judge allowed a class-action lawsuit alleging overbilling of Medicaid patients to proceed.	Moderate
Org 2	Multiple BBB complaints regarding billing issues; past settlement with DOJ over ADA violations.	Moderate
Org 3	Settled a \$900K pay out in a medical negligence case; no direct overbilling allegations found.	Moderate
Org 4	No public overbilling cases; some patient complaints on billing practices.	Low
Org 5	Established fraud and abuse policy; no recent overbilling cases found.	Low
Org 6	No public overbilling allegations; provides detailed billing information.	Low
Org 7	Centralized billing system; no public overbilling cases found.	Low
Org 8	Named in a \$300M fraud suit over EHR kickbacks involving 62 Indiana hospitals.	High
Org 9	Recognized for ethical billing practices; no overbilling cases found.	Low
Org 10	Paid \$1.5M to settle whistle-blower suit over billing errors.	Moderate
Org 11	Agreed to pay \$1.725M to resolve False Claims Act allegations.	High
Org 12	No public overbilling allegations; provides transparent billing information.	Low
Org 13	No public overbilling cases; some patient complaints on billing practices.	Low
Org 14	Paid \$16K for employing an excluded individual; no overbilling cases found.	Moderate
Org 15	Settled multiple overbilling cases totalling over \$100M; recent \$228.5M antitrust settlement.	High

The data collected on the 15 organizations reveals varying levels of risk related to medical fraud and billing issues. The **high-risk organizations** (Org 8, Org 11, Org 15) are notably involved in significant legal actions, such as fraud suits, large settlements, or violations of the False Claims Act. These organizations represent a clear indication of systemic problems in billing practices or fraud that could be targeted for further investigation. On the other hand, the **moderate-risk organizations** (Org 1, Org 2, Org 10, Org 14) show a pattern of complaints, past legal settlements, or smaller-scale allegations of overbilling, indicating potential issues that may not be widespread but still warrant attention. The inclusion of class-action lawsuits, DOJ settlements, and past whistle-blower suits suggests ongoing vulnerabilities within these organizations. The **low-risk organizations** (Org 4, Org 5, Org 6, Org 9, Org 12, Org 13) demonstrate better management practices, with no significant fraud allegations, transparent billing, and established fraud policies. However, even these organizations have minor issues like patient complaints or isolated billing discrepancies that could affect their future risk profiles. These findings highlight the importance of ongoing monitoring and transparency to prevent fraud and ensure ethical billing practices across the healthcare sector. The generated Python script for automation and the use of OSINT tools for verification proved to be an effective approach for identifying and assessing medical fraud risks. The script streamlined data collection, processing, and visualization, while OSINT ensured that the data was accurate, credible, and up-to-date. The success of these tools contributed to the overall effectiveness of the fraud detection system and helped produce actionable insights for stakeholders. All the objectives are achieved with the methodology fig1 to fig3 represent and explain the essence of the OSINT role in the fraud discussion.

Invoice id	Service type	Charge amount	Fraud flag	Organization display	Leak status
INV3001	X-Ray	250	Normal	org1	No Leak Detected
INV3002	Blood Test	70	Normal	org2	Potential Leak Found
INV3003	MRI	5200	Suspicious	org3	No Leak Detected
INV3004	Consultation	150	Normal	org4	No Leak Detected
INV3005	Surgery	10000	Suspicious	org5	Potential Leak Found
INV3006	Blood Test	90	Normal	org6	No Leak Detected

4. Conclusion:

OSINT tools combined with machine learning algorithms provide a better framework for proactively detecting healthcare billing fraud. This approach enables early detection of anomaly which strengthens financial integrity, and enhances organizational security. As the healthcare system is growing we needed a specified technology similar to block chain which help us with proactively mapping healthcare institutions using OSINT tools and applying these algorithms to detect billing anomalies, and correlating cyber vulnerabilities with identified fraudulent behaviours. The integration of OSINT techniques such as Maltego and SpiderFoot provided valuable insights into the organizational networks and potential vulnerabilities of healthcare providers. The application of the Isolation Forest model allowed for detection of suspicious billing activities without requiring labelled datasets, which is enabling traditional reactive methods. On further Correlating the billing anomalies with cybersecurity gaps, demonstrated that technical vulnerabilities often align with higher fraud risks like data leaks.

5.Acknowledgement

The author extends sincere gratitude to the esteemed faculty at Jain - School of Science for their invaluable mentorship and support throughout the course of this project. The author is particularly grateful to Ms. Priyanksha Das Ma'am and Mr. Vishnu Venkatesh Sir for their continuous guidance, encouragement, and expert insights into the medical fraud detection domain and special thanks for providing seamless access to the necessary tools, resources,, which were crucial for the successful execution of this work Their unwavering support, willingness to provide constructive feedback, and constant motivation were instrumental in shaping the direction of this project.

5.References

- Centers for Medicare & Medicaid Services. (2022). Medicare Fraud & Abuse: Prevention, Detection, and Reporting. Retrieved from <https://www.cms.gov/outreach-and-education/medicare-learning-network-mln/mlnproducts/medicare-fraud-and-abuse-prevention-detection-and-reporting>
- Federal Bureau of Investigation (FBI). (2023). Health Care Fraud. Retrieved from <https://www.fbi.gov/investigate/white-collar-crime/health-care-fraud>
- Miller, S. (2021). Investigative Uses of Open Source Intelligence in Healthcare. *Journal of Health Information Management*, 35(4), 22-29.
- Holtfreter, K. (2018). Fraud in the Healthcare Sector: A Global Perspective. *Crime, Law and Social Change*, 70(2), 177–195. <https://doi.org/10.1007/s10611-018-9773-2>
- Office of Inspector General (OIG). (2022). Common Fraud Schemes in Healthcare Billing. Retrieved from <https://oig.hhs.gov/fraud/healthcare.asp>
- Healthcare IT News. (2022). 62 Indiana Hospitals Named in \$300M Fraud Suit. Retrieved from <https://www.healthcareitnews.com/news/62-indiana-hospitals-named-300m-fraud-suit>
- Becker's Hospital Review. (2023). Judge Allows Prisma Health Overbilling Lawsuit to Proceed. Retrieved from <https://www.beckershospitalreview.com/legal-regulatory-issues/judge-allows-prisma-health-overbilling-lawsuit-to-proceed.html>
- U.S. Department of Justice. (2021). False Claims Act Settlements. Retrieved from <https://www.justice.gov/civil/fraud-section/fraud-statistics>
- Discover Global Network. (2024). How AI and Machine Learning Are Battling Global Financial Fraud. Retrieved from <https://insights.discoverglobalnetwork.com/insights/how-ai-and-machine-learning-are-battling-financial-fraud>
- Federal News Network. (2025). IRS Deploys AI Tools to Combat Emerging Tech's Role in New Fraud Schemes. Retrieved from <https://federalnewsnetwork.com/artificial-intelligence/2025/01/irs-deploys-ai-tools-to-combat-emerging-techs-role-in-new-fraud-schemes/>
- PYMNTS.com. (2024). New AI System Aims to Detect Financial Fraud Across Corporate Networks. Retrieved from <https://www.pymnts.com/fraud-prevention/2024/new-ai-system-aims-to-detect-financial-fraud-across-corporate-networks/>
- NVIDIA Blog. (2024). New NVIDIA AI Workflow on AWS Detects Fraudulent Credit Card Transactions. Retrieved from <https://blogs.nvidia.com/blog/ai-workflow-fraud-detection/>
- The Times. (2025). AI 'Violation Detector' Will Scan Civil Service Expense Claims. Retrieved from <https://www.thetimes.co.uk/article/ai-violation-detector-will-scan-civil-service-expense-claims-trs0l7stq>
- Axios. (2025). Exclusive: Visa Sets Up New Team to Take Down All Scammers. Retrieved from <https://www.axios.com/2025/03/11/visa-scam-disruption-practice-fraud>
- The Australian. (2024). Fighting Fraud: CommBank Homes in on Identity Thieves. Retrieved from <https://www.theaustralian.com.au/business/wealth/identity-thieves-targeted-as-commbank-trial-proves-successful/news-story/ca6f46a4623f66a0beeb9bbdd48b716c>
- Scikit-learn Developers. (2024). Scikit-learn: Machine Learning in Python. GitHub Repository. Retrieved from <https://github.com/scikit-learn/scikit-learn>
- TensorFlow Developers. (2024). TensorFlow: An End-to-End Open Source Machine Learning Platform. GitHub Repository. Retrieved from <https://github.com/tensorflow/tensorflow>
- Paterva. (2024). Maltego: Open Source Intelligence and Forensics Application. Retrieved from <https://www.paterva.com/>
- Shodan. (2024). Shodan: The Search Engine for Internet-Connected Devices. Retrieved from <https://www.shodan.io/>
- Laramies. (2024). TheHarvester: E-mail, Subdomain, and People Names Harvester. GitHub Repository. Retrieved from <https://github.com/laramies/theHarvester>
- Liu, F. T., Ting, K. M., & Zhou, Z. H. (2008). Isolation Forest. *Proceedings of the 2008 IEEE International Conference on Data Mining*, 413–422. <https://doi.org/10.1109/ICDM.2008.17>
- OSINT Framework. (2024). OSINT Framework: Open Source Intelligence Collection Tools. Retrieved from <https://osintframework.com/>
- Trend Micro Research. (2024). AI-driven Fraud Detection: Cybersecurity Report. Retrieved from https://www.trendmicro.com/en_us/research.html
- Deloitte Insights. (2024). Cognitive Technologies and Fraud Detection. Retrieved from <https://www2.deloitte.com/us/en/insights.html>
- BAE Systems. (2024). Open Source Intelligence (OSINT) and AI in Financial Crime Detection. Retrieved from <https://www.baesystems.com/en/cybersecurity/feature/osint-in-fraud-detection>