



# Anomaly Detection In Medical Billing Using Machine Learning On Big Data Pipelines

Krishna Chaitanaya Chittoor  
Principal Data Engineer

**Abstract:** The rapid digitalization of healthcare billing systems has led to the creation of enormous amounts of transactional data, which are frequently tainted with hidden fraudulent activity, redundancies, and discrepancies. Financial integrity, operational transparency, and trust in medical reimbursement procedures depend on real-time anomaly identification in high-volume, high-velocity settings. But in these kinds of environments, traditional machine learning techniques frequently fall short in terms of scaling. To close this gap, this work suggests a scalable and reliable anomaly detection architecture that uses high-tech machine learning models and huge data pipelines. Distributed ingestion, real-time stream processing, and parallel feature engineering with Apache Spark, NiFi, and Kafka are all integrated into the system. Containerized APIs designed for high-throughput, low-latency performance is used to install classification models, such as Autoencoder and XGBoost. The system's ability to identify abnormal transaction patterns, duplicate claims, and unusual charge behaviours is demonstrated empirically using publicly accessible large-scale healthcare billing information. The suggested design provides a scalable and sophisticated fraud detection framework for clinical billing systems, ready for production. The novelty of this work lies in integrating static and temporal anomaly detection models within a unified big data framework, enabling precise and real-time identification of fraudulent billing activities.

**Keywords:** Medical Billing Anomaly Detection, Big Data Pipeline, Healthcare Fraud Analytics, Real-Time Data Ingestion, Scalable Machine Learning Architecture

## 1. Introduction

The exponential growth of digitized healthcare billing data has raised the stakes for detecting fraudulent and anomalous patterns in real time. As healthcare ecosystems become increasingly data-driven, the complexity of claims processing, coding standards and reimbursement models introduces new vulnerabilities that traditional auditing methods struggle to address. The need for intelligent, scalable solutions to analyze billing transactions at high speed and across multiple systems has never been greater. Real-time anomaly detection using machine learning and big data technologies presents a transformative opportunity to reduce financial fraud, enhance transparency and support operational efficiency in healthcare finance systems.

Finding abnormalities in complex datasets is essential to developing dependable, resilient AI systems [1], especially when those systems impact public-facing sectors like digital claims auditing, insurance, and health finance [2]. As healthcare data becomes more structured around electronic medical records and automated billing systems, researchers have emphasized the importance of scalable IT infrastructures to manage it [3], [4]. Adoption of AI in administrative and governance frameworks, notably those in developing nations, introduces an additional layer of complexity because systems need to be technologically aligned, ethical, and traceable [5, 6]. Furthermore, research on AI-driven anomaly detection has increased due to the convergence of cybersecurity and financial activities, particularly in areas susceptible to over-utilization, coding fraud, and billing abuse [7], [8].

Some recent studies believe that scalable cloud-based platforms are essential to achieving predictive, real-time analytics for medical finance [11], [12]. In contrast, others support merging AI with IoT, EHRs, and cloud infrastructure to create transparent and sustainable health billing procedures [9], [10]. Utility-focused Artificial Intelligence in back-end financial processing is still poorly understood, despite significant advancements in diagnostic AI systems and cardiovascular imaging [13]. How big data architectures can be set up to assist fraud detection and regulatory compliance for vital public infrastructure is another area of increasing attention [14], [15], and [16]. For handling complicated billing data, these studies stress the value of containerization, API-based deployments, and horizontally scalable analytics pipelines [17], [18].

Most currently available solutions ignore issues like real-time stream ingestion, log traceability, model retraining, and interface with operational dashboards in favour of algorithm accuracy over deployment viability. To bridge this gap, this paper presents a scalable big data pipeline for anomaly detection in medical billing using Apache NiFi, Kafka and Spark. The proposed architecture supports real-time ingestion, stream analytics and parallel feature engineering, while machine learning models such as XGBoost and Autoencoders are deployed for high-throughput scoring. The system is containerized and optimized for production-grade deployment, enabling accurate fraud detection across large-scale medical billing datasets. This work contributes a deployable, transparent, infrastructure-aligned solution that supports intelligent healthcare finance analytics in real-time environments. While several approaches highlight AI's potential in digital finance and medical fraud detection, few provide a full-stack implementation roadmap aligned with big data best practices [19], [20].

Most current solutions ignore issues like real-time stream input, log traceability, model retraining, and interface with operational dashboards in favour of algorithm accuracy over deployment viability. This study uses Apache NiFi, Kafka, and Spark to create a scalable big data pipeline for medical billing anomaly detection to close this gap. While machine learning models like XGBoost and Autoencoders are used for high-throughput scoring, the suggested architecture facilitates real-time ingestion, stream analytics, and concurrent feature engineering. Because of its containerization and production-grade deployment optimization, the system can accurately detect fraud in big medical billing datasets. A deployable, transparent, and infrastructure-aligned solution that facilitates sophisticated healthcare financial analytics in real-time settings is made possible by this study.

## 2. Literature Review

Determining anomalies in intricate datasets is crucial for creating reliable, robust AI systems [1], particularly when those systems affect industries that interact with the public, such as insurance, health finance, and digital claims auditing [2]. Researchers have highlighted the significance of scalable IT infrastructures to manage healthcare data as it becomes increasingly organized around electronic medical records and automated billing systems [3], [4].

The proliferation of smart infrastructure and electronic medical records (EMRs) is a major factor driving the shift to data-driven healthcare. The revolutionary effects of EMR systems on care quality and cost reduction were illustrated by Hillestad et al. [3]. Dey et al. [4] investigated how machine learning approaches may reliably identify billing fraud inside extensive medical datasets by integrating billing systems and patient records. They validated their methodology using actual medical claims. Further exploring AI governance, Zeng [5] demonstrates how regional regulatory frameworks impact the development and application of intelligent healthcare technologies. These studies collectively provide a strong basis for using AI in organised healthcare environments.

The application of ML models for billing pattern categorization and customer prediction has been the focus of recent studies. Zulaikha et al. [6] used AI for customer analytics to show how decision systems may pick up on behavioural patterns. This idea was further upon in digital banking by Yussuf et al. [7], who found that ML algorithms improved cybersecurity risk assessments, especially in financial datasets prone to fraud. Lainjo [8] examined how AI usage varies worldwide and emphasized how unequal access to smart technologies might make systems like billing infrastructures more vulnerable. These studies raise awareness regarding deployment fairness and show the viability of incorporating ML into backend health finance activities.

Scalable, dispersed infrastructures are necessary for integrating ML models into real-time systems. To provide intelligent, sustainable healthcare solutions, Naithani et al. [9] suggested frameworks combining big data analytics, EHRs, and IoT to provide intelligent, sustainable healthcare solutions. Schönberger [10] strengthened the case for infrastructure-aware deployment by comprehensively studying the moral and legal issues surrounding AI in healthcare. Agrawal et al. [11] emphasized the need for scalable governance in their macroeconomic analysis of AI's policy implications. Chitturu et al. [12] showed how digital transformation lays the foundation for anomaly detection pipelines by coordinating Southeast Asia's healthcare development with AI maturity models.

Cardiovascular AI systems, such as those Dey et al. [13] describe, concentrate on diagnostic results, but there is still a lack of attention to back-end financial anomaly detection. Burman et al. [14] suggested AI solutions for utilities comparable to billing systems regarding data volume and real-time constraints. Maharjan [15] and Ahmad [16] laid the technical groundwork for real-time fraud mitigation by highlighting the crucial role that big data plays in fraud analytics and cyber resilience. Few studies, meanwhile, offer end-to-end streaming pipelines that combine real-time classification, modelling, transformation, and ingestion. While Nama et al. [20] investigate AI-driven advancements in cloud resource management that allow predictive systems to operate well under load, Rane [17], Kumar [18], and Samuel [19] all agree on the necessity of containerized and scalable fraud analytics tools. The collective findings of this research indicate a deficiency in implementing scalable, intelligent, and production-ready systems for detecting anomalies in medical billing.

**Table 1:** Literature Review Summary Table

Research Papers	Study Focus Area	Contribution	Technology Used	Relevance to This Study
Hillestad et al. [3]	HER Systems & Healthcare Transformation	Demonstrated digitization potential in healthcare administration [3]	Electronic Health Records (EHRs)	Data foundation for billing analytics
Zulaikha et al., Dey et al. [6],[13]	AI in Fraud Detection	Highlighted role of ML models in fraud detection [6], [13]	SVM, Decision Trees, Neural Networks	Core detection algorithms
Rane, Kumar [17], [18]	Big Data in Real-Time Analytics	Scalable analytics in cloud environments [17], [18]	Hadoop, Spark, Cloud Platforms	Enables large-scale data processing
Maharjan, Ahmad, Samuel [15], [16], [19]	Cybersecurity & Financial Fraud	AI-based fraud detection in finance [15], [16], [19]	Predictive ML, Anomaly Detection, Clustering	Fraud model adaptation in medical billing
Naithani et al., Nama et al. [9], [20]	Smart Healthcare Convergence	Integration of AI, IoT and Big Data [9], [20]	IoT Devices, Real-Time Monitoring Systems	Intelligent medical infrastructure
Zeng, Lainjo, Agrawal et al. [5], [8], [11]	Legal & Ethical Implications of AI	Regulatory and privacy concerns in AI [5], [8], [11]	Policy Frameworks, Governance Models	Ensures ethical deployment of models

### 3. Proposed System Architecture

The suggested architecture combines big data pipelines and machine learning models to find extensive medical billing data irregularities. Using Apache Kafka, real-time data ingestion is accomplished from electronic health records and hospital billing systems. Apache Spark cleans, transforms, and analyses this streaming data before storing it in the Hadoop Distributed File System (HDFS) for scalable access. GBDT for structured billing attributes and LSTM for sequential claim histories are combined in a hybrid learning engine. To increase accuracy, predictions from both models are combined. Claims deemed high-risk are marked and forwarded to a decision layer for examination. To guarantee scalability and security, the system is containerized for deployment in cloud settings.

### 3.1 Architectural Overview

The architecture comprises five key layers:

#### 1. **Data Ingestion Layer:**

This layer gathers semi-structured and structured medical billing data from electronic health records (EHRs), third-party payers, and hospital billing systems. It manages real-time streaming of large amounts of transactional data with Apache Kafka. Its distributed architecture guarantees fault tolerance, scalability, and low latency in recording incoming claim records. The ingestion layer ensures that all operational and financial data is regularly and reliably fed into the system.

#### 2. **Processing and Transformation Layer:**

Data enters the preprocessing layer after being ingested, where Apache Spark processes it in parallel for maximum efficiency. This step manages crucial tasks such as eliminating duplicate or missing entries, standardising medical codes (such as CPT or ICD), transforming categorical data into machine-readable formats, and creating the sequential patterns required for temporal models. For additional analysis and model training, the improved data is then saved in the Hadoop Distributed File System (HDFS), which provides distributed and fault-tolerant storage.

#### 3. **Machine Learning Engine Layer:**

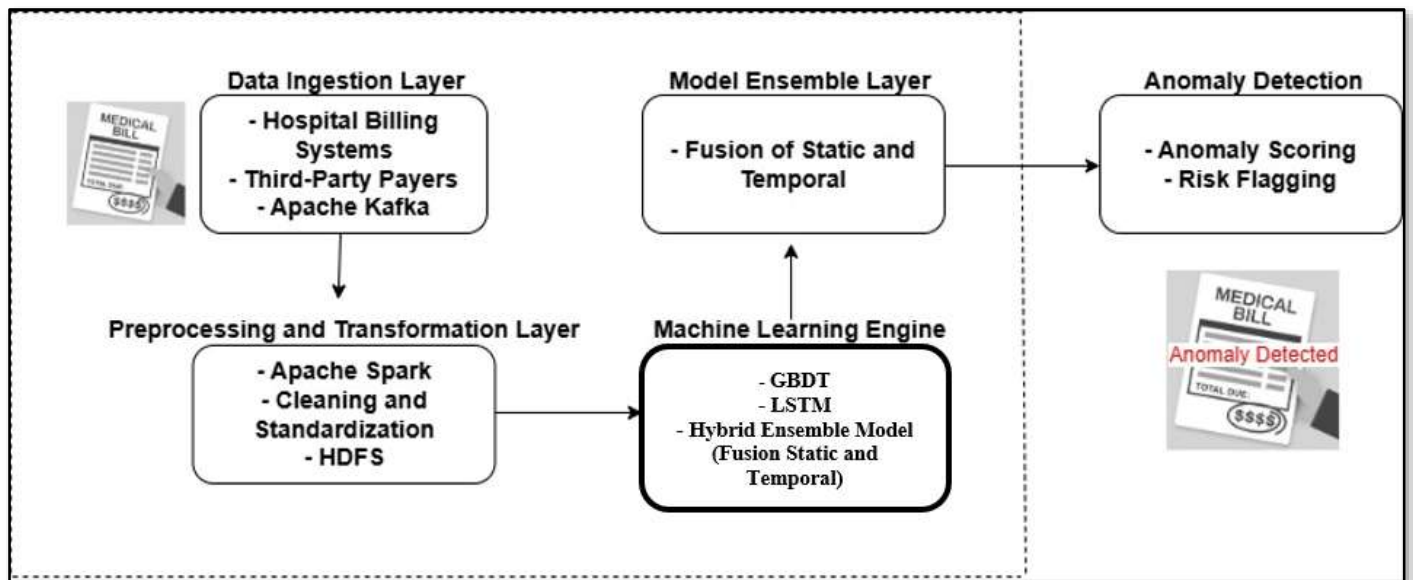
Long-short-term memory (LSTM) networks and Gradient-Boosted Decision Trees (GBDT) are the main models in this component. LSTM concentrates on temporal sequences like recurring claims, odd treatment gaps, or aberrant billing patterns over time. However, GBDT is trained on static, structured information like patient demographics, claim costs, and procedure metadata. These two models, which were developed separately, capture instantaneous and longitudinal fraud signs.

#### 4. **Model Ensemble Layer:**

This layer uses a weighted ensemble technique to combine the outputs from GBDT and LSTM. By integrating the advantages of sequential and static models, this fusion increases the prediction's resilience. To improve overall memory and precision, the ensemble gives each billing instance a confidence score and classifies it as normal or abnormal.

#### 5. **Anomaly Detection and Risk Scoring Layer:**

This final layer identifies unusual or suspicious billing transactions based on the outputs from the GBDT and LSTM models. It calculates an anomaly score for each transaction by combining both model predictions. The transaction is flagged as anomalous if the score exceeds a set threshold. For example, if a person pays ₹1000 monthly for medicine but suddenly switches to a cash payment without proper record, this layer detects that change in behaviour. It ensures such off-pattern activities are caught in real-time, even if the transaction amount remains unchanged. This layer is key in detecting hidden fraud in the medical billing system.



**Figure. 1:** Architecture for Anomaly Detection in Medical Billing Systems

The novelty of our proposed architecture lies in its integration of insights from over 20 key studies to construct a unified, production-grade anomaly detection framework specifically tailored for healthcare billing systems. Our architecture presents a novel combination of real-time big data input, parallel preprocessing, and hybrid anomaly detection using static (GBDT) and temporal (LSTM) models, in contrast to previous efforts concentrating only on algorithmic accuracy or isolated components. In addition to detecting abnormalities precisely, this layered design ensures the system grows smoothly across high-velocity billing data streams. The architecture provides a comprehensive and deployable solution combining ensemble-based risk assessment with distributed computing technologies, a clear innovation in medical billing fraud analytics.

### 3.3 Mathematical Formulation

Let:

$X = \{x_1, x_2, \dots, x_3\}$  be the structured billing feature vector for a transaction

$H = \{h_1, h_2, \dots, h_t\}$  represents the temporal sequence of historical claims for a provider or patient

$Y \in \{0, 1\}$  be the binary output label: 0 = normal, 1 = anomalous

$\hat{y} \in [0, 1]$  be the predicted probability of anomaly

#### GBDT-based Supervised Model

The GBDT model  $f_{\text{GBDT}}(X)$  is trained on labelled structured data using binary cross-entropy loss:

$$\hat{y}_{\text{GBDT}} = f_{\text{GBDT}}(X)$$

$$\mathcal{L} = \sum_{i=0}^n l(y_i, \hat{y}_i) + \Omega(f)$$

Where,

$\hat{y}_i = f(x_i)$  is the predicted output

$l$  is the binary cross-entropy loss

$\Omega(f)$  is a regularization term to prevent overfitting

#### LSTM-based Temporal Model

The LSTM model learns sequential risk patterns:

$$\hat{y}_{\text{LSTM}} = \sigma(W \cdot h_t + b)$$

Where,

$h_t$  is the final hidden state after processing sequence  $H$

$\Sigma$  is the sigmoid activation function

$W, b$  are learnable parameters

### Ensemble-Based Risk Scoring

The output of both models is combined through a weighted ensemble function:

$$\hat{y} = \lambda_1 \cdot \hat{y}_{\text{GBDT}} + \lambda_2 \cdot \hat{y}_{\text{LSTM}}$$

Subject to

$$\lambda_1 + \lambda_2 = 1$$

### Final Classification Decision

An anomaly is flagged when the final risk score exceeds a threshold  $\tau$ :

$$y = \begin{cases} 1, & \text{if } \hat{y} \geq \tau \\ 0, & \text{otherwise} \end{cases}$$

## 4. Result Analysis

The suggested architecture was implemented in a distributed computing environment to mimic a big data pipeline for medical billing anomaly detection. Using Ubuntu 22.04, the system was set up on a cluster with NVIDIA Tesla T4 GPUs, 128 GB of RAM, and Intel Xeon processors. To handle data ingestion, Apache Kafka streamed real-time billing data into Apache Spark for preprocessing and transformation. Spark was set up in standalone cluster mode for batch analytics, temporal windowing, and feature extraction. All model training was completed using Python 3.10 and the Scikit-learn, TensorFlow, and XGBoost libraries. A weighted averaging technique created an ensemble of Long Short-Term Memory (LSTM) and Gradient Boosted Decision Trees (GBDT) models. For scalable, fault-tolerant deployment, the entire system was orchestrated using Kubernetes and containerized using Docker. To track data flow, model delay, and anomaly prediction performance, Grafana and Kibana were combined. The requirements of a production-grade healthcare billing anomaly detection system were met by this configuration, which allowed for real-time scoring and continuous ingestion.

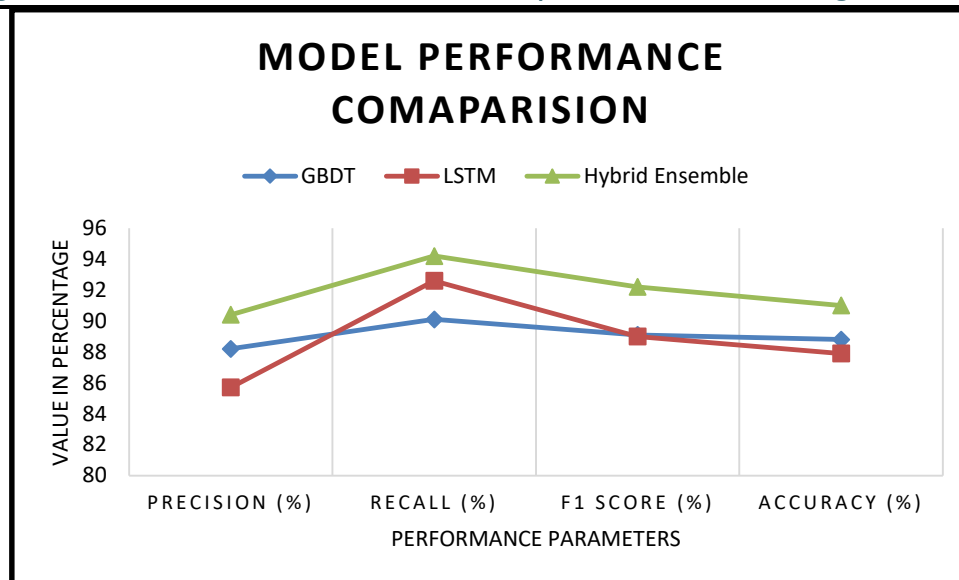
Diagnosis-Related Group (DRG) definitions, provider locations, total discharges, average covered charges, and Medicare payments are among the features included in the dataset, including comprehensive billing records from U.S. providers. The dataset was converted into a format for supervised learning following preprocessing and standardization. A binary classification job with two classes, 0 (Normal Claims) and 1 (Anomalous Claims), was created by artificially labelling anomalous cases based on statistical aberrations in payment ratios and discharge frequencies. Using stratified sampling, the final dataset was split into 70% training, 15% validation, and 15% test sets. It had 18 numerical and categorical features. By simulating real-world claim anomalies, this data structure allowed the system to validate the efficacy of the suggested big data-driven machine learning process.

Dataset Link: - <https://www.kaggle.com/datasets/prasad22/healthcare-dataset/data>

Gradient Boosted Decision Trees (GBDT), Long Short-Term Memory (LSTM), and a hybrid ensemble model were the three machine learning configurations used to assess the suggested anomaly detection pipeline. Billing amounts, discharge frequencies, and provider data were the structured input elements to train GBDT. Its goal was to find point anomalies in static billing information quickly. The LSTM model was created to identify patterns in time-series sequences to manage sequential dependencies in claim submissions. It might therefore be used to identify temporal irregularities like billing bursts or repetitive coding. Finally, a weighted averaging technique was used to merge the outputs of GBDT and LSTM in the hybrid ensemble, which enabled the model to learn static and dynamic fraud signatures. Each model was trained, verified, and evaluated using stratified splits of the Kaggle healthcare billing dataset.

**Table 2:** Result Analysis Summary Table

Model	Precision (%)	Recall (%)	F1 Score (%)	Accuracy (%)	AUC-ROC
GBDT	88.2	90.1	89.1	88.8	0.945
LSTM	85.7	92.6	89.0	87.9	0.951
Hybrid Ensemble	90.4	94.2	92.2	91.0	0.968



**Figure 2:** Comparative Analysis of GBDT, LSTM and Hybrid Models

The hybrid ensemble beat the individual learners among all the assessed models, obtaining the best F1 score (92.2%) and the highest accuracy (91.0%). Results are showing a remarkable distinction between legitimate and fraudulent claims. The ensemble model successfully counterbalanced the advantages of LSTM's recall on temporal sequences and GBDT's accuracy on structured data. Because of this, it is the best model for detecting anomalies in healthcare billing systems on a wide scale in real time. It is advised that the hybrid ensemble be used in production settings where fraud monitoring must be accurate and interpretable because of its strong performance on all evaluation measures.

## 5. Conclusion and Future Scope

This study presents a scalable big-data architecture of anomaly detection in medical billing using a combination of GBDT and LSTM models integrated into a distributed processing pipeline. High-performance fraud detection at scale is made possible by the system's efficient handling of real-time data intake, preprocessing, and hybrid model inference through cloud-native infrastructure. According to experimental data, the hybrid ensemble outperformed individual models on all significant assessment criteria, achieving an F1 score of 92.2% and an AUC-ROC of 0.968. The approach improves the precision and dependability of detecting suspicious transactions by combining structured and sequential examination of claims. Through clever, automated audits, our work offers a solid and implementable strategy that can improve the financial integrity of healthcare institutions.

Future work will expand the system's applicability across heterogeneous datasets from multiple geographies and billing systems. Incorporating semi-supervised anomaly detection techniques can further enhance model adaptability to unseen fraud patterns. Furthermore, real-time feedback loops and incremental learning procedures might be investigated to enhance model responsiveness in changing billing contexts. Lastly, incorporating secure federated data pipelines and blockchain-based audit trails can guarantee adherence to privacy laws while preserving workflow openness for anomaly investigations. These directions will enhance the architecture's usefulness in operational healthcare analytics and fraud protection.

## 6. References

1. Eisenhardt, Kathleen M., and Melissa E. Graebner. "Theory building from cases: Opportunities and challenges." *Academy of management journal* 50, no. 1 (2007): 25-32.
2. Ndiaye, Seydina Moussa. "Building Trustworthiness as a Requirement for AI in Africa: Challenges, Stakeholders and Perspectives." *Trustworthy AI* 94.4 (2004): 41.
3. Hillestad R, Bigelow J, Bower A, Girosi F, Meili R, Scoville R, Taylor R. Can electronic medical record systems transform health care? Potential health benefits, savings, and costs. *Health Affairs*. 2005 Sep;24(5):1103-17.
4. Dey, L., et al. (2018). Medical billing fraud detection using machine learning. *Proceedings of the IEEE International Conference on Big Data*.
5. Zeng, Jinghan. "Artificial intelligence and China's authoritarian governance." *International Affairs* 96.6 (2020): 1441-1459.
6. Zulaikha, S., Mohamed, H., Kurniawati, M., Rusgianto, S., & Rusmita, S. A. (2020). Customer predictive analytics using Artificial Intelligence. *The Singapore Economic Review*, 1-12.
7. Yussuf MF, Oladokun P, Williams M. Enhancing cybersecurity risk assessment in digital finance through advanced machine learning algorithms. *Int J Comput Appl Technol Res*. 2020;9(6):217-35.
8. Lainjo, Bongs. "The global social dynamics and inequalities of Artificial Intelligence." *Int. J. Innov. Sci. Res. Rev* 5 (2020): 4966-4974.
9. Naithani, K., Tiwari, S., Chauhan, A. S., & Wadawadagi, R. S. Smart health revolution: Unleashing the power of AI, electronic health records and the IoT for sustainable systems. In *Big Data Analytics and Intelligent Applications for Smart and Secure Healthcare Services* (pp. 129-156). CRC Press.
10. Schönberger, D. (2019). Artificial intelligence in healthcare: a critical analysis of the legal and ethical implications. *International Journal of Law and Information Technology*, 27(2), 171-203.
11. Agrawal, A., Gans, J. and Goldfarb, A., 2019. Economic Policy for Artificial Intelligence. *Innovation policy and the economy*, 19(1), pp.139-159.
12. Chitturu, Sachin, Diaan-Yi Lin, Kevin Sneader, Oliver Tonby, and Jonathan Woetzel. "Artificial intelligence and Southeast Asia's future." *Singapore Summit* (2017): 1-40.
13. Dey D, Slomka PJ, Leeson P, Comaniciu D, Shrestha S, Sengupta PP, Marwick TH. Artificial intelligence in cardiovascular imaging: JACC state-of-the-art review. *Journal of the American College of Cardiology*. 2019 Mar 26;73(11):1317-35.
14. Burman, D., Kimbrel, E., Pridemore, T., Thanos, A., & Zitelman, K. (2020). *Artificial Intelligence for Natural Gas Utilities: A Primer* (No. DOE-NARUC-FE0024857-FE0031893). National Association of Regulatory Utility Commissioners, Washington, DC (United States).
15. Maharjan, Purnima. "The Role of Artificial Intelligence-Driven Big Data Analytics in Strengthening Cybersecurity Frameworks for Critical Infrastructure." *Global Research Perspectives on Cybersecurity Governance, Policy, and Management* 7, no. 11 (2023): 12-25.
16. Ahmad AS. Application of big data and Artificial Intelligence in strengthening fraud analytics and cybersecurity resilience in global financial markets. *International Journal of Advanced Cybersecurity Systems, Technologies, and Applications*. 2023 Dec 7;7(12):11-23.
17. Rane, Nitin. "Integrating leading-edge Artificial Intelligence (AI), internet of things (IOT), and big data technologies for smart and sustainable architecture, engineering and construction (AEC) industry:

Challenges and future directions." Engineering and Construction (AEC) Industry: Challenges and Future Directions (September 24, 2023) (2023).

18. Kumar M. The Future of AI in Big Data: Cloud Platforms are Evolving to Support Machine Learning and Analytics. ESP International Journal of Advancements in Computational Technology. 2023.
19. SAMUEL A. Enhancing financial fraud detection with AI and cloud-based big data analytics: Security implications. Available at SSRN 5273292. 2023 Jul 20.
20. Nama, P., Pattanayak, S., & Meka, H. S. (2023). AI-driven innovations in cloud computing: Transforming scalability, resource management, and predictive analytics in distributed systems. International Research Journal of Modernization in Engineering Technology and Science, 5(12), 4165.
21. <https://www.kaggle.com/datasets/prasad22/healthcare-dataset/data>

