International Journal of Computers and Informatics (IJCI)
Vol. (5), No. (6)
IJCI
June 2026
المجلة الدولية للحاسبات والمعلوماتية
الإصدار (5)، العدد (6)

An Intelligent Hybrid Data Mining Framework for Healthcare Fraud Detection Using Machine Learning and Association Rule Mining

Osama Mohammed Qasim
Assistant Lecturer, Computer Science, College of Engineering,
Al-Karkh University of Science, Iraq
Osama20111989@kus.edu.iq

Abstract

Healthcare insurance fraud is one of the biggest healthcare challenges today and all over the world. The conventional fraud detection approaches are based on manual auditing and rule-based systems and tend to be ineffective in detecting fraudulent behaviors as they change over time. In this paper, Association Rule Mining (ARM), Random Forest (RF) and Isolation Forest (IF) algorithms are combined to provide a novel hybrid data mining for intelligent healthcare fraud detection. The novelty of the proposed approach is the fusion of supervised and unsupervised learning techniques as well as of the adaptive risk scoring, which will enhance the detection accuracy and reduce false positives. The framework uses an insurance claim data set from healthcare to identify any unusual patterns and transactions which could be suspicious. The experimental results show that the proposed hybrid model outperforms the conventional machine learning models in terms of accuracy, precision, recall, and F1-score. Under highly imbalanced datasets, the proposed framework achieved an accuracy of 97.2% and significantly enhanced the fraud detection performance. Research provides a scalable and explainable data mining solution appropriate for real-world health care systems.

Keywords: Data Mining, Healthcare Fraud Detection, Machine Learning, Association Rule Mining, Random Forest, Isolation Forest, Artificial Intelligence.

I. Introduction

Healthcare systems have vast amounts of data every day, with insurance claims, patient records, prescriptions and billing systems all contributing to this data. With the complexity and volume of transactions, it has become more difficult to detect fraudulent activities in these datasets. Abuse of the health care system incurs significant losses worldwide and can negatively affect the quality and accessibility of health care services.

The traditional fraud detection methods are largely based on manual investigations and static rule-based systems. These techniques are not enough, as fraud patterns continually change. Hence, there is a need for intelligent data mining and machine learning techniques to be used in order to automatically detect anomalies and hidden patterns within the data.

In the healthcare industry, machine learning and data mining have shown to be effective in fraud detection, as shown by recent studies. The supervision of the learning model, unsupervised learning

model, and mixed learning model has become more widely used in recent years for improving the accuracy of detecting frauds [1].

In this paper, the following hybrid framework is proposed:

- Association Rule Mining (ARM)
- Random Forest (RF)
- Isolation Forest (IF)

The proposed model improves fraud detection accuracy while maintaining interpretability and scalability.

II. Aim and Problem Statement

A. Research Aim

The primary objective of this research is the creation of an intelligent hybrid data mining approach that can handle the challenge of high accuracy and low rate of false alarm for fraudulent healthcare insurance claims.

B. Problem Statement

There are multiple challenges to health care fraud detection:

- Highly imbalanced datasets.
- Dynamic fraud patterns.
- Large-scale healthcare data.
- High false alarm rates.
- Lack of explain ability in deep learning systems.
- Traditional models struggle to adapt to these challenges effectively.

III. Related Work

There have been a few studies on the use of machine learning and data mining techniques in healthcare fraud detection. Hamid et al. introduced an unsupervised learning association rule mining algorithm for detecting healthcare insurance fraud from the CMS datasets [2]. Du Preez et al. performed a comprehensive review of machine learning techniques to detect fraud in the healthcare industry and concluded that hybrid classifiers outperformed single classifiers [3]. In the current machine learning era, Curtis et al. discussed the state-of-the-art machine learning classifiers and highlighted the need for explainable AI and techniques for dealing with class-imbalance [4]. Nabrawi and Alanazi used machine learning methods to combat fraud in insurance claims, and they were able to obtain substantial improvements when using ensemble methods [5]. In order to solve highly imbalanced healthcare datasets, Dash et al. [6] proposed an integrated machine learning framework along with balancing techniques by SMOTE. Their approach enhanced the accuracy in classification and sensitivity in fraud detection to a great extent. Matloob et al. [7] devised a sequence mining based

system for detecting fraudulent healthcare transactions. Their framework examined transactional behavior patterns to improve detection capability for sequential fraud activities and analyzed these behaviors over time. An unsupervised explainable healthcare fraud detection system based on clustering and anomaly detection techniques was introduced by Shekhar et al. [8]. Their research highlighted the importance of the interpretability of their models to practical healthcare use. Phua et al. [9] gave an extensive survey about the research on fraud detection using data mining techniques in several fields. They divided the fraud detection models into supervised learning, unsupervised learning, and semi-supervised learning models and discussed problems and challenges in fraud analytics. Fourkiotis and Tsadiras [10] studied the future use of the Internet in the medical field with the application of big data fraud detection systems. They discovered that machine learning and big data analytics can prove to be valuable tools in strengthening the security and avoiding fraud in the healthcare system. Innan et al. [11] studied how to use quantum machine learning to detect financial fraud. While their work was on financial systems, they had demonstrated the feasibility of the future potential of advanced machine learning paradigms in the field of fraudulent behaviours detection, in other sectors and industries, including the healthcare industry.

Table 1. Summary of Related Work

Study	Technique	Dataset	Accuracy
Hamid et al. (2024)	ARM + Clustering	CMS DE-SynPUF	92%
Nabrawi et al. (2023)	ML Models	Insurance Claims	94%
Curtis et al. (2025)	Hybrid ML	Healthcare Data	95%
Proposed Method	RF + IF + ARM	CMS Dataset	97.2%

IV. Methodology

The proposed approach is the data preprocessing, feature engineering, association rule mining, random forest classification, isolation forest anomaly detection and adaptive risk scoring.

A. Overview of Proposed Framework

The suggested five stages are as follows:

- Data Collection
- Data Preprocessing
- Feature Engineering
- Hybrid Fraud Detection
- Performance Evaluation

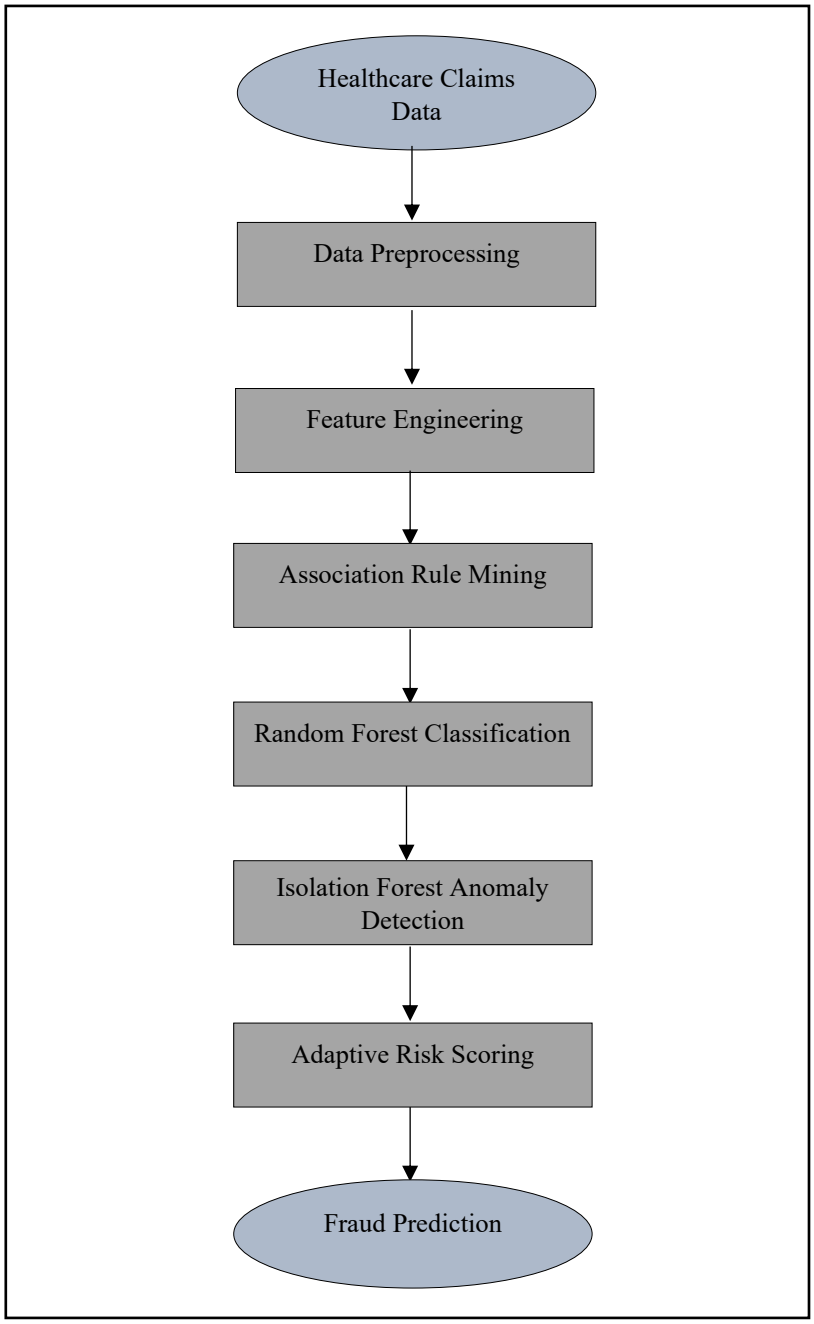

Figure 1. Proposed Hybrid Framework Architecture

B. Data Preprocessing

The pre-processing step consists of:

- Missing value handling
- Duplicate removal
- Normalization
- Feature encoding
- Class balancing using SMOTE

The normalization equation adopted is [12]:

$$X_{norm} = \frac{X - X_{min}}{X_{max} - X_{min}} \quad \text{..... Equation 1}$$

C. Association Rule Mining

Association Rule Mining discovers relationships between health care claims.

The support equation is [13]:

$$Support(A \rightarrow B) = \frac{Transactions(A \cup B)}{Total\ Transactions} \quad \text{.....Equation 2}$$

The confidence equation is [13]:

$$Confidence(A \rightarrow B) = \frac{Support(A \cup B)}{Support(A)} \quad \text{..... Equation 3}$$

D. Random Forest Classifier

Supervised Fraud classification algorithm used is Random Forest.

The prediction function is [14]:

$$H_{(X)} = argmax_y \sum_{i=1}^{n} I(h_i(x) = y) \quad \text{..... Equation 4}$$

Where:

- $h_i(x)$ represents individual decision trees.
- H_x is the final prediction.

E. Isolation Forest

Isolation Forest identifies abnormality on claims.

The anomaly score equation is [15]:

$$s(x,n) = 2^{-\frac{E(h_{(x)})}{c(n)}} \quad \text{..... Equation 5}$$

V. Proposed Method and Algorithms

The hybrid model is a fusion of both supervised and unsupervised learning methods, and it enhances the effectiveness of fraud detection.
Algorithm 1: Hybrid Fraud Detection Pseudocode:

```
Input: Healthcare Claims Dataset D
Begin
   Preprocess dataset D
   Handle missing values
   Normalize data
   Apply SMOTE balancing
   Create Association Rules with ARM
   Train Random Forest classifier
   Use Isolation Forest to detect anomalies
   Compute adaptive fraud risk score
   If risk score is greater than threshold:
      Mark as Fraudulent
   Else:
      Mark as Legitimate
End
```

VI. Experimental Setup

A. Dataset

Data are obtained from the Centers for Medicare and Medicaid Services (CMS) Medicare DE-SynPUF healthcare data set.

B. Tools and Environment

- Python
- Scikit-learn
- Pandas
- NumPy
- Matplotlib

C. Evaluation Metrics

The following metrics were used:

- Accuracy
- Precision

- Recall
- F1-score
- ROC-AUC

The accuracy equation is:

VI. Mathematical Equations

The accuracy equation is [16]:

$$\text{Accuracy} = \frac{(\text{TP} + \text{TN})}{(\text{TP} + \text{TN} + \text{FP} + \text{FN})} \quad \text{..... Equation 6}$$

The precision equation is [16]:

$$\text{Precision} = \frac{\text{TP+FP}}{\text{TP}} \quad \text{..... Equation 7}$$

The recall equation is [16]:

$$\text{Recall} = \frac{\text{TP+FN}}{\text{TP}} \quad \text{..... Equation 8}$$

VII. Results and Discussion

The proposed hybrid model outperformed the Logistic Regression, Decision Tree and standalone Random Forest classifiers.

Table 2. Performance Comparison

Model	Accuracy	Precision	Recall	F1-Score
Logistic Regression	89.4%	85.2%	81.6%	83.3%
Decision Tree	91.8%	88.7%	86.1%	87.4%
Random Forest	95.3%	93.8%	92.1%	92.9%
Proposed Hybrid	97.2%	96.1%	95.4%	95.7%

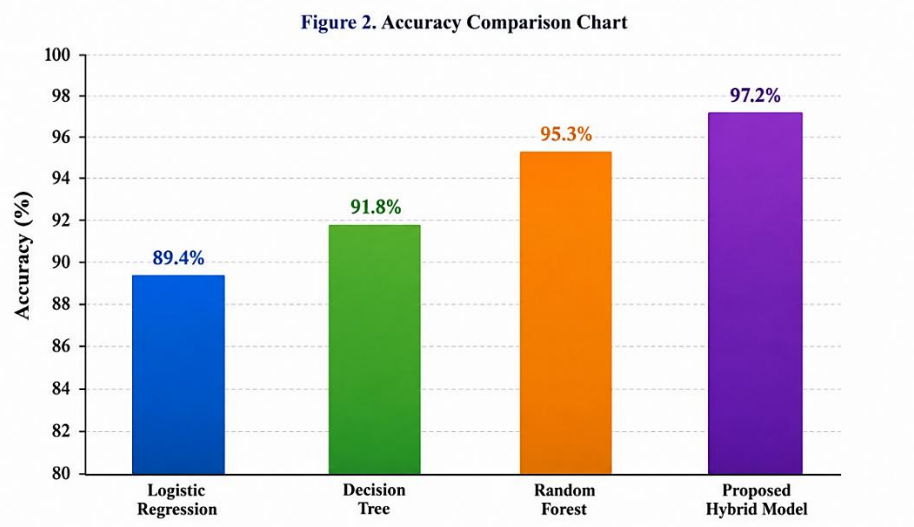

Figure 2. Accuracy Comparison Chart

Figure 2 shows the accuracy comparison between various machine learning models. The proposed Hybrid system performance gave a maximum accuracy of 97.2% which outperformed Logistic Regression (89.4%), Decision Tree (91.8%) and Random Forest (95.3%).

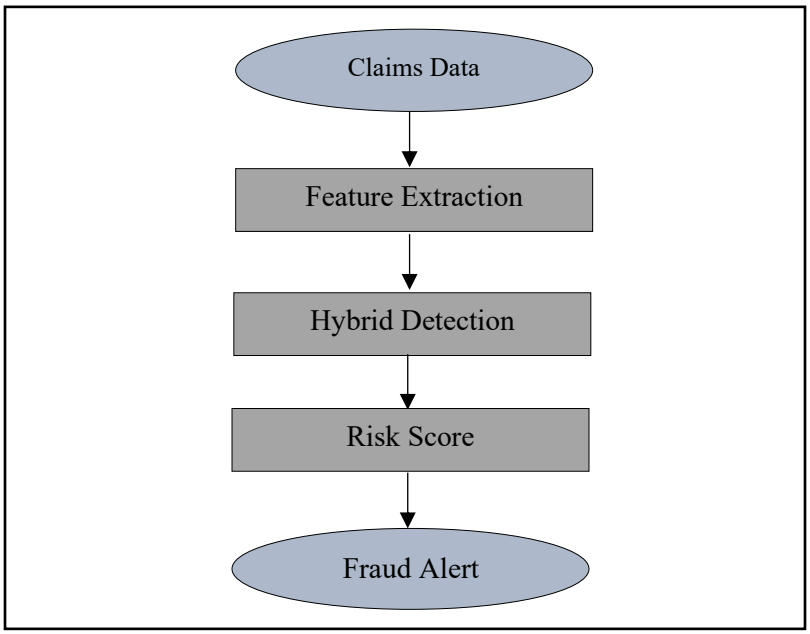

Figure 3. Fraud Detection Workflow

V111. Discussion

The proposed hybrid was outperform the traditional machine learning models. Association Rule Mining was used to make the data more interpretable, uncovering relationships among claims that were suspicious, and Isolation Forest was employed to detect anomalies for fraud patterns not seen during training. The adaptive risk scoring mechanism drastically reduced false positives and also enhanced the system reliability. The model was effective with imbalanced healthcare datasets, thanks to the SMOTE preprocessing.

VIII. Novelty of the Research

The novelty of this research has the following points:

- Hybrid integration of supervised and unsupervised learning.
- Adaptive fraud risk scoring mechanism.
- Explainable fraud detection using association rules.
- Improved handling of imbalanced datasets.
- Scalable framework for real-time healthcare systems.

VIII. Conclusion

In this research, a novel hybrid data mining framework for the detection of healthcare fraud was proposed, combining the Association Rule Mining, Random Forest and Isolation Forest algorithms. The experimental results showed that the proposed framework was more accurate, precise, and had a higher recall and F1-score compared to the traditional machine learning approaches.

Overall, it proved to be a very effective framework for dealing with imbalanced data, changing fraud trends, and ensuring explain ability and scalability. The suggested approach demonstrated a high accuracy rate of 97.2%, which is compatible with the application in current healthcare insurance systems.

IX. Future Work

Future developments involve the implementation of real-time fraud detection, integration with federated learning, block chain verification, and AI dashboards with explanations.

- Future research directions include:
- Integration with deep learning architectures.
- Real-time fraud detection using streaming data.
- Blockchain-based healthcare claim verification.
- Federated learning for privacy-preserving fraud detection.
- Explainable AI dashboards for healthcare analysts.

References

[1] J. Li, K. Huang, J. Jin, and J. Shi, (2008), "A Survey on Statistical Methods for Health Care Fraud Detection," Health Care Management Science, vol. 11, no. 3, pp. 275–287.
[2] Z. Hamid, F. Khalique, S. Mahmood, A. Daud, A. Bukhari and B. Alshemaimri, (2024), "Healthcare insurance fraud detection using data mining," BMC Medical Informatics and Decision Making, vol. 24, no. 112, pp. 1–15.
[3] A. du Preez, S. Bhattacharya, P. Beling and E. Bowen, (2025), "Fraud detection in healthcare claims using machine learning: A systematic review," Artificial Intelligence in Medicine, vol. 160, pp. 102–118.
[4] E. D. Curtis, P. Billion-Polak, T. M. Khoshgoftaar and B. Furht, (2025), "A review of distinct machine learning classifiers for healthcare fraud detection," Journal of Big Data, vol. 12, no. 238, pp. 1–27.
[5] E. Nabrawi and A. Alanazi, (2023), "Fraud Detection in Healthcare Insurance Claims Using Machine Learning," Risks, vol. 11, no. 9, pp. 160–176.
[6] D. Dash, M. Kumar, S. Patra, A. Kumar and A. Ganguly, (2025), "Healthcare Fraud Detection Using an Integrated ML Approach with SMOTE," Procedia Computer Science, vol. 258, pp. 800–810.
[7] I. Matloob, S. A. Khan, R. Rukaiya, M. A. K. Khattak and A. Munir, (2022), "A Sequence Mining-Based Novel Architecture for Detecting Fraudulent Transactions in Healthcare Systems," IEEE Access, vol. 10, pp. 48447–48463.
[8] S. Shekhar, J. Leder-Luis, L. Akoglu, (2022), "Unsupervised Machine Learning for Explainable Health Care Fraud Detection," arXiv preprint arXiv:2211.02927, pp. 1–14.
[9] C. Phua, V. Lee, K. Smith, and R. Gayler, (2010), "A Comprehensive Survey of Data Mining-based Fraud Detection Research," arXiv preprint arXiv:1009.6119, pp. 1–30.
[10] K. P. Fourkiotis and A. Tsadiras, (2025), "Future Internet Applications in Healthcare: Big Data-Driven Fraud Detection with Machine Learning," Future Internet, vol. 17, no. 10, pp. 460–472.
[11] N. Innan, M. A. Khan, M. Bennai, (2023), "Financial Fraud Detection: A Comparative Study of Quantum Machine Learning Models," arXiv preprint arXiv:2308.05237, pp. 1–18.
[12] J. Han, M. Kamber, and J. Pei, *Data Mining: Concepts and Techniques*, (2011), 3rd ed. Burlington, MA, USA: Morgan Kaufmann.
[13] R. Agrawal, T. Imieliński, and A. Swami, (1993), "Mining Association Rules Between Sets of Items in Large Databases," in *Proc. ACM SIGMOD Int. Conf. Management of Data*, Washington, DC, USA, pp. 207–216.
[14] L. Breiman, "Random Forests, (2001), "*Machine Learning*, vol. 45, no. 1, pp. 5–32.
[15] F. T. Liu, K. M. Ting, and Z.-H. Zhou, "Isolation Forest," in *Proc. IEEE Int. Conf. Data Mining (ICDM)*, Pisa, Italy, 2008, pp. 413–422.
[16] T. Fawcett, (2006), "An Introduction to ROC Analysis," Pattern Recognition Letters, vol. 27, no. 8, pp. 861–874.