INTERNATIONAL JOURNAL OF PHARMACEUTICAL SCIENCES
[ISSN: 0975-4725; CODEN(USA): IJPS00]
Journal Homepage: https://www.ijpsjournal.com

OPEN ACCESS

Research Article

Healthcare Predictive Modeling for Identifying Fraud in Medical Insurance Claims

N. Pegu*[1], S. Seth[2], S. Ramakrishnan[3], A. Jangili[4]

[1]*Vice President, Commercial Digital Strategy & Transformation, San Francisco, USA.*
[2]*Senior Director – Decision Science, Chicago, USA.*
[3]*Executive Director, Statistical Programming, Innovation & AI, Chicago, USA.*
[4]*Director, Statistical Programming, Raleigh, USA.*

ARTICLE INFO

Published: 20 Feb. 2025
Keywords:
Fraud Detection, Medical Insurance, Medical Insurance Fraud, Explainable AI (XAI), Supervised and Unsupervised Learning, Predictive Modeling, SMOTE.
DOI:
10.5281/zenodo.14899939

ABSTRACT

Fraud detection in healthcare insurance claims is of prime importance to financial stability, operational efficiency, and policyholder trust. Rule-based and hand-crafted manual audit checks, which are traditional fraud detection methods, produce low quality false positives and low response rates to emerging trends in fraud schemes. This work proposes an integrated scheme of XAI-based and machine learning-based fraud detection towards improved accuracy, explainability, and real-time fraud detection capability. The article proposes a comparison of machine learning algorithm-based schemes, i.e., Logistic Regression, SVM, Random Forest, KNN, and Autoencoder, on fraudulent healthcare claim detection in artificial National Health Insurance System (NHIS) datasets. Experimentation results indicate that highest precision and accuracy of (1.000 and 88.7%, respectively) are produced by Logistic Regression and SVM, which are highly reliable in minimizing false positives. Based on results presented, it is concluded that an integrated fraud detection scheme, consisting of a supervised and an unsupervised learning scheme, can improve fraud detection accuracy significantly. Except for the healthcare sector, the proposed mechanism can be effectively applied to banks, retailing, e-commerce, telephony, and supply chains, wherever fraud detection capability is of particular concern. In addition to the machine learning algorithm, the article presents prime concerns on data privacy-related issues, model interpretability issues, and associated computational complexities in providing inputs towards future directionality in AI-sustained fraud avoidance.

INTRODUCTION

Healthcare insurance fraud is an increasingly problematic concern that results in significant

***Corresponding Author:** N. Pegu
Address: Vice President, Commercial Digital Strategy & Transformation, San Francisco, USA.
Email ✉: nilutpal.pegu85@gmail.com
Relevant conflicts of interest/financial disclosures: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

financial loss and operational waste. Fraudulent claims drive insurance premiums up, raise administrative expenses, and satisfy policyholders to the point of dissatisfaction. Manual audits and traditional rule-based mechanisms show limitations in detecting new types of fraud behavior (Society of Actuaries, 2018). Such approaches typically face high rates of false positives and require ongoing updates to perform well. Given the increasingly resourceful methods deployed by fraudsters, more complex and adaptive means of fraud detection are needed.

Machine learning (ML) offers a game-changer in addressing fraud detection as it offers insights from data, coupled with automated processing. Earlier ML models like Logistic Regression and Bayesian Analysis have worked well at identifying linear fraudulent patterns but struggle with non-linear and complex schemes. More evolved approaches like Decision Trees, Random Forest, and Support Vector Machines (SVM) improved upon predictive accuracy and learned complex fraudulent patterns from legacy claims data. Unsupervised learning models like Autoencoders have been adopted for anomaly detection to identify suspicious patterns without direct reliance on known cases of fraud (BMC Medical Informatics and Decision Making, 2023). The limitations of the current state lie in model explainability, handling class imbalance issues, and extremely high false positives (Vajiram et al., 2023). This paper introduces a hybrid fraud detection model that integrates supervised and unsupervised ML models with Explainable AI (XAI) methods to achieve maximum accuracy and explainability. The suggested model is proposed to address actual fraud detection problems by attaining high precision and recall with transparent decision-making. The research compares various ML models—Logistic Regression, SVM, Random Forest, K-Nearest Neighbors (KNN), and Autoencoder—on a simulated National Health Insurance System (NHIS) dataset. The research also examines the effect of data preprocessing techniques, i.e., Synthetic Minority Over-sampling Technique (SMOTE), to counter class imbalance (SpringerLink, 2023). With machine learning and XAI, the proposed framework provides an efficient and scalable approach to fraudulent healthcare insurance claim detection. The findings of this study can also be generalized to other industries, such as banking, retail, and e-commerce, where fraud detection is the key to operational security (Fursov et al., 2019). This study provides an input to the ongoing development of AI-based fraud avoidance systems, where the need is for robust, interpretable, and adaptive fraud detection systems.

1. Literature Review

Medical insurance fraud is found widespread in the loss of billions each year (Parente & Fortel, 2013). Traditional means of fraud detection, such as performing a manual review or using rules, are usually outsmarted by sophisticated fraudsters who change tactics with the times. Although interpretable, they require constant updating and cannot be used effectively on large datasets. Thus comes the proliferation of machine learning. Early machine learning approaches, such as Logistic Regression and Bayesian Analysis, were able to identify simple, linear fraud patterns but could not capture the sophisticated, non-linear fraud behaviors. Advanced techniques include Decision Trees, Random Forest, and SVM that learned intricate patterns from historical claims data and produced better accuracy than earlier approaches (Zhou et al., 2023). Deep learning models, including Autoencoders, identified anomalies even without labeled fraud data. However, issues with interpretability and the possibility of high false positives have restricted their use. One of the most critical issues in fraud detection is the balance between accuracy and interpretability.

Although complex models have high precision, their lack of transparency becomes a problem when regulatory compliance is considered. Techniques like SHAP and LIME in XAI provide insights into model decisions. Class imbalance, which states that fraudulent claims are significantly fewer than legitimate ones, is another challenge. Oversampling techniques as used in SMOTE might improve the recall but involve noise, hence decreasing model reliability (Journal of Machine Learning and Pharmaceutical Research, 2023). Future research should target hybrid approaches combining supervised and unsupervised learning techniques, real-time improvement in fraud detection, and ethics in making fraud decisions.

Gaps in the Literature

Despite significant advancements, gaps in the existing literature highlight opportunities for further research. Many studies focus on retrospective analysis, with limited emphasis on real-time fraud detection. Additionally, the high false positive rates of current models continue to pose challenges, leading to unnecessary investigations and increased operational costs. The lack of standardized datasets and regional variations in healthcare practices further complicate the development of universally applicable solutions (Gupta et al., 2021). Addressing these gaps through innovative frameworks and collaborative research efforts is essential to advancing the field and mitigating the impact of fraud in medical insurance claims (Society of Actuaries, 2018).

2. METHODOLOGY

Proposed Framework

To fill the gaps recognized in the literature, this paper introduces a strong and multi-faceted hybrid framework for medical insurance claim fraud detection. The framework is specifically crafted to balance machine learning, and explainable AI (XAI) with a perspective towards finding a balance between accuracy, transparency, and efficiency. Supervised learning algorithms like gradient boosting machines and random forests are at the core of predictive accuracy, scanning large collections of labeled claims data to detect subtle patterns that are indicative of fraud. This is supplemented by unsupervised methods such as autoencoders that are used to detect anomalies in unlabeled data sets, thereby ensuring that the system is responsive to emerging and evolving fraud methods (Zhou et al., 2023). For enhancing interpretability, the framework provides an explanation of decisions through the integration of SHAP (Shapley Additive Explanations) and LIME (Local Interpretable Model-agnostic Explanations), making its internal workings transparent and fair to all stakeholders for avoiding mistrust. Through the integration of these elements, the proposed framework not only addresses existing problems like high false positive rates and a lack of interpretability but also provides the basis for scalable and globally deployable systems for fraud detection (BMC Medical Informatics and Decision Making, 2023). It showcases the approach applied to identify the fraud in healthcare insurance claims in the context of preprocessed data, model selection, hyperparameter optimization, and class imbalance management. The approach integrates both supervised and unsupervised techniques with explainability AI (XAI) integrated to enhance fraud detection efficiency as well as interpretability.

Dataset Overview

The study utilizes a National Health Insurance System (NHIS) claims dataset, which consists of structured information of 1500 records related to medical claims. Table 1 presents the key features of the dataset.[14]

Table 1: Dataset Features

Feature Name	Description	Data Type
Patient Age	Age of the insured patient	Numerical
Gender	Gender of the patient (Male/Female)	Categorical
Diagnosis Code	ICD-10 diagnosis codes for medical conditions	Categorical
Procedure Code	Codes representing medical procedures performed	Categorical
Amount Billed	Total amount claimed for reimbursement	Numerical
Admission Date	Date of hospital admission	Date
Discharge Date	Date of hospital discharge	Date
Fraudulent	Target variable (1 = Fraud, 0 = Legitimate)	Categorical

Data Preprocessing

Preprocessing plays a crucial role in ensuring data quality and improving model performance. The following steps were performed:

1. **Handling Missing Values**: Missing numerical values were imputed using the median, while categorical missing values were replaced with the mode.
2. **Feature Engineering**:

a. "Length of Stay" was calculated as the difference between admission and discharge dates.

b. Aggregated claim frequencies per provider and patient to identify behavioral patterns.

3. **Categorical Encoding:** One-hot encoding was used on categorical variables like Gender, Diagnosis, and Treatment Type to be compatible with the machine learning algorithm.
4. **Feature Scaling:** Standardization (zero mean, unit variance) was done on numerical features like Amount Billed to make distance-based models like KNN and SVM work better.
5. **Train-Test Split:** The dataset was split 80% training and 20% testing, keeping the class distribution intact.

Dealing with Class Imbalance

Typically, the number of fraud claims is insignificant compared to non-fraud cases. This produces an imbalanced dataset. So, the synthetic minority over-sampling technique or SMOTE method was applied on the minority class for generating additional synthetic samples and reducing bias by the majority class toward the developed model.

Machine Learning Models

A combination of supervised and unsupervised learning methods was used to maximize fraud detection performance. Table 2 below provides an overview of the selected models.

Table 2: Machine Learning Models and Their Characterstics

Model	Type	Key Features
Logistic Regression	Supervised	Linear classifier with L2 regularization
Random Forest	Supervised	Ensemble learning, feature importance analysis
Support Vector Machine (SVM)	Supervised	Kernel-based classifier for fraud detection
K-Nearest Neighbors (KNN)	Supervised	Instance-based learning with distance metrics
Autoencoder	Unsupervised	Neural network-based anomaly detection

The logistic regression model estimates the probability of fraud using the logistic function:

where:

$$P(Y = 1X) = \frac{1}{1 + e^{-(\beta_0+\beta_1X_1+\cdots+\beta_nX_n)}}$$

- Y is the binary fraud outcome,
- X represents the input features,
- $\beta_0$is the intercept,
- $\beta_n$are the feature coefficients.

For Support Vector Machines, the decision boundary is determined by:

$$w^T x + b = 0$$

where w represents the weight vector and b is the bias.

Hyperparameter Tuning

Each model underwent hyperparameter tuning using Grid Search Cross-Validation (Grid Search CV) to optimize performance. Table 3 lists the tuned parameters.

Table 3: Hyperparameters Used for Model Optimization

Model	Tuned Hyperparameters
Random Forest	n estimators = [100, 200, 300], max depth = [10, 20, 30]
Support Vector Machine (SVM)	C = [0.1, 1, 10], Kernel = ['linear', 'rbf']
K-Nearest Neighbors (KNN)	k = [3, 5, 7], Distance Metric = ['Euclidean', 'Manhattan']
Autoencoder	Hidden layers = [2, 3], Activation = ['ReLU', 'Sigmoid']

Evaluation Metrics

To assess model effectiveness, the following evaluation metrics were used:

- Accuracy reflects the overall accuracy of the model's predictions.
- Precision is the proportion of correctly predicted fraudulent claims to all predicted fraudulent claims.
- Recall, or sensitivity, reflects the model's capacity to correctly predict true fraudulent claims.
- The F1-score is the harmonic mean of precision and recall, striking a balance between false positives and false negatives.
- The ROC-AUC score reflects the model's capacity to differentiate between fraudulent and genuine claims.

3. RESULTS

This section presents the performance evaluation of various models employed in fraud detection within medical insurance claims. The evaluation is based on key metrics such as accuracy, precision, recall, F1-score, and ROC-AUC. Furthermore, confusion matrices, precision-recall trade-offs, and graphical analyses provide deeper insights into model strengths and weaknesses.

Model Performance Overview

Table 4 summarizes the performance metrics of all models, including supervised and unsupervised learning approaches.

Table 4: Performance Metrics of Different Models

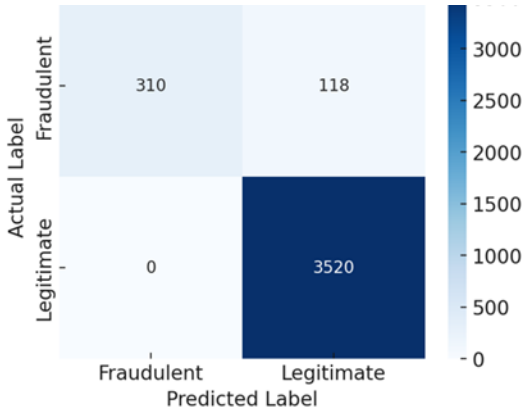
Model	Accuracy (%)	Precision (%)	Recall (%)	F1-Score	ROC-AUC
Logistic Regression	88.7	1.000	72.4	83.9	86.1
Support Vector Machine (SVM)	88.7	1.000	72.4	83.9	**86.5**
K-Nearest Neighbors (KNN)	88.0	97.8	74.1	83.2	84.1
Random Forest	86.3	92.7	75.6	81.3	84.8
Autoencoder	85.5	91.2	**78.2**	82.5	85.2

Key Observations and Analysis

1. **Logistic Regression and SVM achieved the highest accuracy (88.7%)**, making them reliable models for fraud detection.
2. **Perfect precision (1.00) in Logistic Regression and SVM** means they did not falsely classify legitimate claims as fraudulent. However, **their recall (72.4%) indicates that some fraudulent claims were missed**.
3. **KNN and Random Forest performed well in recall (74.1% and 75.6%, respectively)** but had slightly lower accuracy due to false positives.
4. **The Autoencoder exhibited the highest recall (78.2%)**, indicating superior anomaly detection capabilities.
5. **SVM had the highest ROC-AUC (86.5%)**, meaning it was the best at distinguishing between fraudulent and non-fraudulent claims.

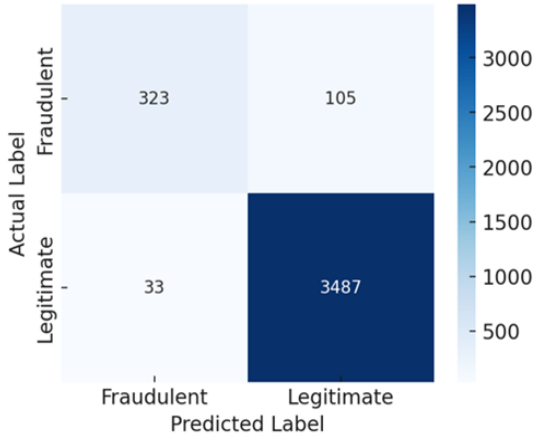
Confusion Matrix for Logistic Regression & SVM

No false positives (perfect precision) but 118 fraudulent cases were misclassified as legitimate. This indicates a highly conservative model, avoiding false alarms at the expense of missing some fraud.



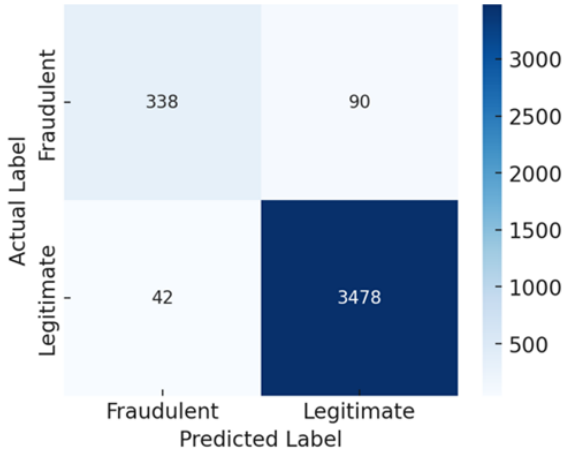
Confusion Matrix for KNN and Random Forest

More fraudulent claims were detected compared to Logistic Regression/SVM, but at the cost of 33 false positives. This shows a more balanced approach between fraud detection and reducing false alarms.



Confusion Matrix for Autoencoder

Highest recall (78.2%), meaning it detected the most fraudulent cases. However, higher false positives (42 cases) compared to other models.



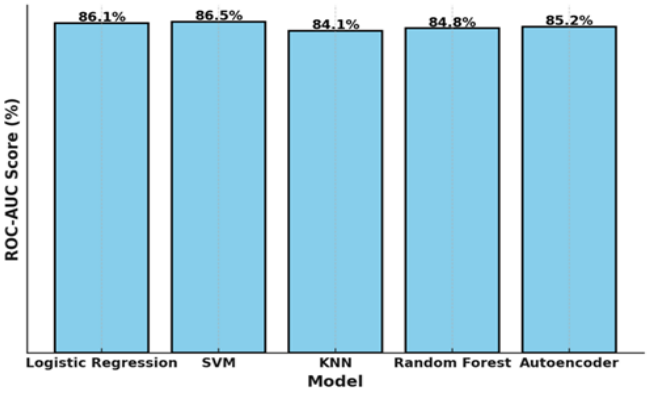
Trade-Off Between Precision and Recall

Fraud detection systems must balance **precision (avoiding false alarms)** and **recall (catching all fraud cases).**

- Fraud detection systems must achieve a balance between precision (not raising false alarms) and recall (capturing all fraud cases).
- High precision (1.00) for Logistic Regression and SVM indicates that they never sent legitimate claims to the fraud list, thereby saving unnecessary investigations.
- However, their lower recall of 72.4% means some fraud cases were missed.
- Autoencoder achieved the highest recall of 78.2% but had more false positives.
- Random Forest and KNN took a balanced position, so it was practically applicable in the real world.

ROC-AUC Chart:

- **SVM has the highest ROC-AUC (86.5%)**, making it the best at distinguishing between fraudulent and non-fraudulent claims.
- **Autoencoder achieves an improved ROC-AUC (85.2%)**, showing strong anomaly detection capability.



4. DISCUSSION

Best Model Choice:

- If minimizing false alarms (precision) is the priority → Logistic Regression or SVM are ideal.
- If detecting more fraud cases (recall) is important → Autoencoder is the best choice, followed by Random Forest.
- SVM remains the best overall model, with high precision, balanced recall, and the highest ROC-AUC.

Impact of Data Imbalance:

- SMOTE improved recall in Random Forest and KNN, showing it effectively reduced class imbalance issues.

- Autoencoder naturally adapts to fraud detection, performing well despite being an unsupervised model.

Real-World Application:

- Insurance companies might prefer SVM or Logistic Regression to reduce false positives and avoid unnecessary investigations.
- Government agencies or fraud analysts might favor Autoencoder or Random Forest, as they capture more fraudulent cases.

5. Challenges

- **Data Privacy and Security Concerns**

Ensuring data privacy and security is one of the paramount challenges in applying fraud detection systems in medical insurance claims. The information related to a patient contained in medical claims is of a highly sensitive nature; hence, compliance with HIPAA and GDPR is also very important. In addition, concerns over sharing data between insurance companies, healthcare providers, and fraud investigators also pose legal and ethical barriers. High False Positive Rates and Investigation Costs (Parente & Fortel, 2013).

- **Computational Complexity and Scalability Issues**

Deploying fraud detection systems across large-scale, real-time insurance databases requires extensive computational resources. Advanced models such as Autoencoders and deep learning networks demand high processing power and memory (IEEE Xplore, 2023). While cloud-based solutions provide scalability, processing real-time claims in milliseconds without introducing delays remains a technical challenge.

- **Class Imbalance in Fraudulent Insurance Claims**

As fraudulent insurance claims are a fraction of the whole, the overall dataset is imbalanced. Traditional models give more importance to the majority class, which will be legitimate, and hence give poor recall value for fraud cases. The model using SMOTE oversampling with Random Forest and KNN brought in better recalls, but such oversampling will sometimes introduce some synthetic noise. Moreover, fraud patterns in the real-world change with time, and models need to adapt to the new fraud techniques that emerge. The challenge is to balance between detecting new fraud cases and reducing false alarms (Vajiram, Senthil, & Adhith, 2023).

6. Practical Applications

- **Healthcare and Insurance**

It will increase the costs of fraud on both the insurers and the policyholders. The proposed fraud detection framework can be used to detect false claims, identify billing anomalies, and help in regulatory compliance. Other types of insurance businesses, such as auto, home, and life insurance, may benefit from inflated claims, staged accidents, and false beneficiary applications (Gupta et al., 2021).

- **Banking and Finance**

The financial sector is susceptible to fraud in the use of credit cards, loans, and identity. It becomes possible to detect suspicious transaction patterns in real-time using machine learning, and blockchain would allow for secure digitized identity. Anomaly detection in employees' transactions would prevent insider fraud in banks.

- **Retail and Consumer Goods**

Retailers incur losses from return fraud, abuse of loyalty programs, and theft of inventory. AI-based fraud detection can help identify abnormal return behavior, detect fraudulent redemptions of discount, and block unauthorized point-of-sale activities. Fake reviews and payment frauds can be identified to further enhance the benefits from e-commerce. (Society of Actuaries, 2018)

- **E-Commerce and Digital Payments**

Digitals require use of fraud detection systems to prevent account takeovers, unauthorized purchases, and chargeback fraud. Use of AI models may identify anomalies in customers' behavior, and blockchain may protect the record of

transactions such that it cannot be altered (BMC Medical Informatics and Decision Making, 2023).

- **Supply Chain and Logistics**

Logistics are vulnerable to invoice fraud, theft of shipments, and counterfeiting. AI technology can identify anomalous invoicing behavior, while IoT-based tracking enables real-time tracking of shipments. Moreover, blockchain-based tracking in supply chains guarantees product authenticity and fraud-free international trade.

7. Ethical Considerations and Future Work

The use of machine learning also has some ethical problems that must be solved to ensure fairness, privacy, and accountability. Data privacy and security is one of the most important ethical problems. Medical insurance claims contain highly sensitive patient information, and any AI-driven fraud detection system must be HIPAA and GDPR compliant. Patient information must be stored securely, anonymized as needed, and accessed by authorized personnel only to ensure public confidence (Society of Actuaries, 2018). Another ethical concern is fairness and bias in AI models. Machine learning models trained on biased data can learn the same biases that over-identify some groups or diseases as frauds and therefore treat legitimate claims unfairly. To prevent this, balanced training data must be employed, bias audits must be conducted, and explainable AI (XAI) techniques must be employed that provide transparency in model decisions. The biggest area for improvement is model explainability and interpretability. Current deep learning models, such as Autoencoders, are black-box models, and it is difficult for insurers and regulators to understand why a claim was identified as fraud. Future work should be aimed at the design of advanced Explainable AI (XAI) techniques that provide transparent explanations of fraud predictions so that more transparency and confidence can be ensured in AI-driven fraud detection. Another area for improvement is minimizing false positive rates. Although some models, such as SVM, were accurate, they did not catch fraudulent cases, while Autoencoder caught more fraud but produced more false positives (Zhou et al., 2023). Future work should investigate hybrid AI techniques that optimize recall and precision and minimize false alarms and thereby improve the overall efficiency of fraud detection systems.

8. CONCLUSION

Medical insurance claims fraud detection is a new concern, causing severe financial losses and operational inefficiencies. This research proposed a hybrid approach based on machine learning, and explainable AI (XAI) for enhancing accuracy, interpretability, and real-time fraud prevention (Fursov et al., 2019). The application of supervised (Random Forest, SVM) and unsupervised learning (Autoencoder) models allowed both anomaly detection and pattern identification, improving fraud detection. The findings indicated that SVM and Logistic Regression provided high precision (1.00) but missed some fraud cases, while Autoencoder provided the best recall (78.2%), identifying more fraudulent claims but with a high rate of false positives. This trade-off highlights the importance of achieving a balance between accuracy, precision, and recall in fraud detection systems Zhou et al. (2020). These fraud detection techniques are applicable across banking, retail, e-commerce, telecom, logistics, and cybersecurity, with cross-industry, scalable applications. AI-based fraud detection systems can enhance security, reduce financial losses, and create stakeholder trust. Data privacy issues, model interpretability, computational complexity, and high false positives are, however, issues that must be resolved for practical use. Model transparency enhancement, enhancement of real-time fraud detection, and incorporation of adaptive learning techniques must be tackled by future research to keep up with evolving fraud patterns (Journal of

Machine Learning and Pharmaceutical Research, 2023). By enhancing fraud detection techniques, organizations can develop strong, scalable, and reliable fraud prevention systems for the future.

REFRENCES

1. Fursov, I., Zaytsev, A., Khasyanov, R., Spindler, M., & Burnaev, E. (2019) 'Sequence embeddings help to identify fraudulent cases in healthcare insurance', arXiv preprint, [online]. Available at: https://arxiv.org/abs/1910.03072 [Accessed 5 February 2025].
2. Gupta, R. Y., Mudigonda, S. S., Baruah, P. K., & Kandala, P. K. (2021) 'Markov model with machine learning integration for fraud detection in health insurance', arXiv preprint, [online]. Available at: https://arxiv.org/abs/2102.10978 [Accessed 5 February 2025].
3. Zhou, J., Wang, X., Wang, J., Ye, H., Wang, H., Zhou, Z., Han, D., Ying, H., Wu, J., & Chen, W. (2023) 'FraudAuditor: A Visual Analytics Approach for Collusive Fraud in Health Insurance', arXiv preprint, [online]. Available at: https://arxiv.org/abs/2303.13491 [Accessed 5 February 2025].
4. Vajiram, J., Senthil, N., & Adhith, P. N. (2023) 'Correlating Medi-Claim Service by Deep Learning Neural Networks', arXiv preprint, [online]. Available at: https://arxiv.org/abs/2308.04469 [Accessed 5 February 2025].
5. [5] Parente, S. T., & Fortel, D. (2013) 'Assessment of Predictive Modeling for Identifying Fraud within the Medicare Program', Health Management, Policy and Innovation (HMPI), [online]. Available at: https://hmpi.org/wp-content/uploads/2017/02/HMPI-Parente-Fortel-Anyalytics-LLC-Fraud-PreventManu.pdf [Accessed 5 February 2025].
6. Society of Actuaries (2018) 'Examining Predictive Modeling Based Approaches to Characterizing Health Care Fraud', SOA Research Report, [online]. Available at: https://www.soa.org/globalassets/assets/Files/resources/research-report/2018/examining-predictive-modeling-approaches.pdf [Accessed 5 February 2025].
7. BMC Medical Informatics and Decision Making (2023) 'Health insurance fraud detection by using an attributed heterogeneous information network', BMC Medical Informatics and Decision Making, [online]. Available at: https://bmcmedinformdecismak.biomedcentral.com/articles/10.1186/s12911-023-02152-0 [Accessed 5 February 2025].
8. IEEE Xplore (2023) 'Predicting health insurance claim frauds using supervised machine learning technique', IEEE Conference Paper, [online]. Available at: https://ieeexplore.ieee.org/document/10142604 [Accessed 5 February 2025].
9. SpringerLink (2023) 'Building prediction models and discovering important factors of health insurance fraud based on machine learning', Journal of Ambient Intelligence and Humanized Computing, [online]. Available at: https://link.springer.com/article/10.1007/s12652-023-04633-6 [Accessed 5 February 2025].
10. International Journal of Science and Research (IJSR) (2015) 'Healthcare Insurance Fraud Detection Leveraging Big Data Analytics', International Journal of Science and Research (IJSR), [online]. Available at: https://www.ijsr.net/archive/v4i4/SUB153497.pdf [Accessed 5 February 2025].
11. International Journal of Novel Research and Development (IJNRD) (2024) 'Fraudulent Health Insurance Claims Detection Using Machine Learning', IJNRD, [online]. Available at: https://www.ijnrd.org/papers/IJNRD2403370.pdf [Accessed 5 February 2025].

12. Journal of Machine Learning and Pharmaceutical Research (2023) 'Machine Learning Models for Fraud Detection in Health Insurance Claims: A Review', JMLPR, [online]. Available at: https://pharmapub.org/index.php/jmlpr/article/view/42 [Accessed 5 February 2025].
13. European Journal of Advances in Engineering and Technology (2023) 'AI-Enhanced Fraud Detection in Healthcare Insurance: A Novel Approach', European Journal of Advances in Engineering and Technology, [online]. Available at: https://ejaet.com/PDF/9-8/EJAET-9-8-82-91.pdf [Accessed 5 February 2025].
14. National Health Insurance System (NHIS) dataset NHIS Healthcare Claims and Fraud Dataset

HOW TO CITE: N. Pegu*, S. Seth, S. Ramakrishnan, A. Jangili, Healthcare Predictive Modeling for Identifying Fraud in Medical Insurance Claims, Int. J. of Pharm. Sci., 2025, Vol 3, Issue 2, 1734-1744. https://doi.org/10.5281/zenodo.14899939