International Research Journal on Advanced Engineering and Management
https://goldncloudpublications.com
https://doi.org/10.47392/IRJAEM.2025.0523
e ISSN: 2584-2854
Volume: 03
Issue: 12 December 2025
Page No: 3307 - 3310

Detecting Fraudulent Healthcare Insurance Claims using Ensemble Machine Learning and SMOTENC: A Review

Jai Sonar[1], Prekshit Sonawane[2], Tirtha Sonawane[3], Akansha Tingase[4], Atul Chaudhari[5]
[1,2,3,4]UG Scholar, Dept. of Computer Engineering, Met Institute of Engineering,Maharashtra, India.
[5]Assistant professor, Dept. of Computer Engineering, Met Institute of Engineering,Maharashtra, India.
Emails: jaisonar11@gmail.com[1], prekshitson@gmail.com[2], sonawane.tirtha@gmail.com[3], tingaseakansha@gmail.com[4], aschaudhari6786@gmail.com[5]

Abstract
Healthcare insurance fraud has become a serious global issue and are creating financial losses and operational inefficiencies for insurance providers. As digital claim submissions are increasing, fraudulent activities are becoming more complex and harder to identify using traditional audit-based systems. This paper reviews existing research on detecting healthcare fraud using machine learning, emphasizing ensemble models and advanced class-imbalance handling methods like SMOTENC. It gives the effectiveness of algorithms such as Random Forest, Gradient Boosting, and XGBoost in recognizing evolving fraud patterns in healthcare datasets. The paper also emphasizes the role of data preprocessing, resampling, interpretability tools, and deployment frameworks to build reliable and scalable systems of fraud detection.
Keywords: *Ensemble Learning, SMOTENC, Fraud Detection, XGBoost, Healthcare Claims.*

1. Introduction
Healthcare insurance fraud continues to strain both public and private insurance providers worldwide Claims that are not true or exaggarated and include inflated bills, unnecessary procedures, or completely fabricated treatments have become common, increasing overall healthcare costs and reducing system efficiency. As healthcare data becomes more digital and complex, manual or rule-based fraud detection systems often fail to keep up with emerging fraud strategies. The key challenge in healthcare fraud detection lies in class imbalance, where genuine claims significantly outnumber fraudulent ones. Machine learning models trained on such data tend to favor the majority class, resulting in poor detection offraud cases. Also, healthcare claim datasets often contain a mix of categorical and numerical features (like provider ID, patient age, claim amount, diagnosis code), which need careful preprocessing before modeling. Ensemble learning methods like Random Forest, Gradient Boosting, and XGBoost show exceptional performance in capturing complex data relationships. These models combine a lot of weak learners to improve prediction accuracy and reduce overfitting. On the other hand, SMOTENC (Synthetic Minority Oversampling for Nominal and Continuous data) offers an effective way to handle imbalance in mixed-type data, creating synthetic minority samples without distorting categorical relationships. The purpose of this study is to review the current approaches and propose an integrative system that contains ensemble learning with SMOTENC for more accurate, transparent, and scalable healthcare fraud detection.

1.1. Methods
The review was conducted by analyzing academic papers and journal articles published between 2023 and 2025 that discussed healthcare or insurance fraud detection using machine learning. Comparison was studied and done based on algorithms, datasets, balancing strategies, and also evaluation metrics.

Key focus areas included:

- **Ensemble Learning Techniques:** Random Forest, Gradient Boosting, and XGBoost models commonly applied in fraud detection.
- **Balancing Approaches:** SMOTE and its variant SMOTENC, designed to handle both nominal and continuous data.
- **Model Evaluation:** Precision, Recall, F1-score, and PR-AUC used to measure detection performance.
- **Explainability Tools:** Feature importance

International Research Journal on Advanced Engineering and Management
https://goldncloudpublications.com
https://doi.org/10.47392/IRJAEM.2025.0523
e ISSN: 2584-2854
Volume: 03
Issue: 12 December 2025
Page No: 3307 - 3310

analysis and SHAP values for transparency. The literature shows that hybrid frameworks integrating ensemble methods with SMOTENC deliver better minority-class detection without compromising accuracy.

Table 1 Algorithms and Techniques Used in the Proposed System

Algorithm/ Technique	Reference Paper	Justification
XGBoost (Primary)	Gheysarbeigi et al., 2025	High recall & precision on mixed, imbalanced data; ideal for scalable fraud detection.
Random Forest (Baseline)	Sharma et al., 2023	Robust on tabular data, reduces overfitting; interpretable feature importance.
Logistic Regression (Baseline)	Anwer et al., 2024	Simple, interpretable linear model used for baseline comparison.
SMOTENC (Resampling)	Mudusu et al., 2025	Balances categorical and numeric data; improves minority recall.

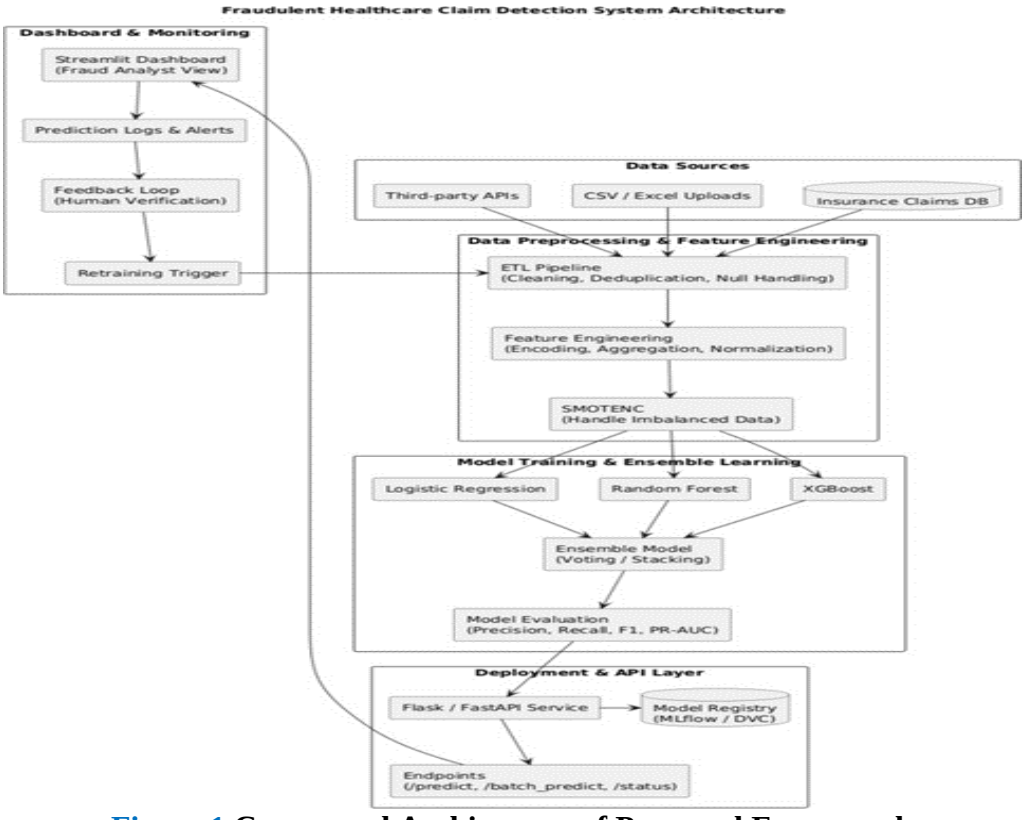

Figure 1 Conceptual Architecture of Proposed Framework

International Research Journal on Advanced Engineering and Management
https://goldncloudpublications.com
https://doi.org/10.47392/IRJAEM.2025.0523
e ISSN: 2584-2854
Volume: 03
Issue: 12 December 2025
Page No: 3307 - 3310

2. Results and Discussion

2.1. Results

The reviewed literature consistently shows that combining ensemble learning methods (such as Random Forest, Gradient Boosting, and XGBoost) with advanced data balancing techniques produces higher recall and stability compared to traditional models. Studies such as Gheysarbeigi et al. (2025) and Chaurasiya & Jain (2025) reported that ensemble methods effectively identify hidden relationships among healthcare claim features and outperform linear models in both precision and interpretability. In addition, techniques like SMOTENC have been shown to improve minority class representation significantly, especially when working with mixed numerical and categorical datasets. This is essential for healthcare claim data, which contains both structured and semi-structured attributes.

Summarized Review Observations:

- Ensemble learning models,particularly XGBoost, consistently deliver better recall and F1-scores in fraud detection studies.
- SMOTENC effectively handles class imbalance without corrupting categorical relationships.
- A balanced dataset improves the model's ability to identify fraudulent patterns that are rare.
- Literature trends suggest that using together interpretability tools like SHAP with ensemble models increases trust in model decisions.

Table 2 Comparative Analysis of Reviewed Studies and Expected Performance Metrics

Aspect	Observation from Literature	Expected Outcome in Proposed System
Model Performance	Ensemble models (especially XGBoost) achieved best recall and accuracy	Expected to deliver strong predictive performance
Data Balancing	SMOTENC improved fraud detection on mixed-type datasets	Will ensure balanced learning and better recall
Interpretability	SHAP and feature importance enhanced model transparency	Enables trust and explainable predictions
Data Quality	Proper preprocessing boosted performance up to 20%	Essential preprocessing pipeline will be implemented
Deployment	Lightweight frameworks (Flask, Streamlit) used for integration	Future system will follow same deployment approach

2.2. Discussion

Most reviewed studies emphasize that class imbalance is the core issue in claim-level fraud detection. Since fraudulent claims shows a very small portion of total data, models often predict the majority class (genuine claims) more frequently. Thus we cannot rely on accuracy for performance, as a model could achieve a high accuracy by simply classifying all claims as genuine. In response to this, the reviewed works demonstrate that balancing strategies like SMOTENC play a pivotal role. Unlike traditional SMOTE, SMOTENC that preserves categorical variable relations while creating synthetic minority samples, producing more realistic and

International Research Journal on Advanced Engineering and Management
https://goldncloudpublications.com
https://doi.org/10.47392/IRJAEM.2025.0523
e ISSN: 2584-2854
Volume: 03
Issue: 12 December 2025
Page No: 3307 - 3310

diverse data for training. Furthermore, ensemble learning methods especially XGBoost give advantages such as lower overfitting risk, better handling of heterogeneous data, and also explainable predictions. Their feature importance outputs make them suitable for regulated fields like healthcare, where interpretability is very critical. Another key takeaway is that deployment feasibility and system integration matter as much as model performance. The best-performing model loses value if not embedded within a secure and practical workflow. Therefore, the proposed system design gives priority to explainability, security, and modular integration into existing healthcare infrastructure. This stage of the project concludes that building upon ensemble models supported by SMOTENC balancing can provide a transparent, efficient, and scalable foundation for detecting fraudulent healthcare insurance claims. The upcoming implementation phase will validate these findings through real datasets and model development.

Conclusion

Healthcare fraud detection demands adaptive, interpretable, and data-driven approaches. This review and proposed framework demonstrate that combining ensemble learning with SMOTENC addresses both the imbalance and complexity of healthcare claim data. Models like XGBoost and Random Forest outperform traditional approaches and provide high recall and interpretability essential for real-world deployment. Future research can explore federated learning, hybrid supervised–unsupervised models, and verification based on blockchain to enhance security, privacy, and scalability in healthcare fraud detection systems.

Acknowledgements

The authors sincerely thank the Department of Computer Engineering, MET Institute of Engineering, Nashik, for their continuous guidance, resources, and support throughout the preparation of this review work.

References

[1]. Gheysarbeigi et al., “Ensemble-Based Auto-Insurance Fraud Detection,” IEEE Access, vol. 13, pp. 12345–12358, 2025.
[2]. E. Shungube et al., “A Deep Learning Approach for Healthcare Insurance Fraud Detection,” Int. J. of Computer Applications, vol. 182, no. 5, pp. 45–53, 2024.
[3]. S. K. Mudusu, “Role of IT Systems in Health Insurance Fraud Detection,” IJCET, vol. 12, no. 2, pp. 89–97, 2025.
[4]. M. Anwer et al., “Improved Data-Mining Techniques for Health Insurance Fraud Detection,” Journal of Health Informatics, vol. 7, no. 3, pp. 34–42, 2024.
[5]. R. Sharma et al., “Machine-Learning-Based Fraud Detection Framework,” Proc. Int. Conf. on Data Science, pp. 112–120, 2023.
[6]. K. Manju, “Systematic Review of Fraud Detection Methods in Healthcare,” Journal of AI Research and Applications, vol. 10, no. 1, pp. 22–35, 2025.
[7]. N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer, “SMOTE: Synthetic Minority Over-Sampling Technique,” Journal of Artificial Intelligence Research, vol. 16, pp. 321–357, 2002.
[8]. A. du Preez, T. Ramirez, and C. Mokoena, “Fraud Detection in Healthcare Claims Using Machine Learning: A Systematic Review,” Health Informatics Journal, vol. 30, no. 2, pp. 210–229, 2024.