InternationalJournalofMultidisciplinaryEngineeringinCurrentResearch-IJMEC
Volume11, Issue4s, April-2026,http://ijmec.com/,ISSN:2456-4265

Full Length Article

Machine Learning based Method for Insurance Fraud Detection on Class Imbalance Datasets with Missing Values

Rehan Zakiuddin[1], Mohammed Ali[2] ,Tayyab Mohiuddin[3], Dr. Md Zainlabuddin[4]
[1,2,3]B.E Students; Department Of Computer Science And Engineering, ISL Engineering College, Hyderabad, India.
[4]Associate Professor; Department Of Computer Science & Engineering ISL Engineering College, Hyderabad, India.
Mail Id;160522733171@islec.edu.in ,160522733164@islec.edu.in, 160522733134@islec.edu.in

Accepted 24-04-2026

Abstract:
Machine Learning based insurance fraud detection plays an important role in identifying fraudulent insurance claims and reducing financial losses. In this project, machine learning techniques were implemented to detect fraudulent and genuine insurance claims on class imbalance datasets containing missing values. The dataset was preprocessed using data cleaning, missing value handling, normalization, and balancing techniques to improve model performance and reliability. Different machine learning algorithms were trained and evaluated to classify claims accurately despite the imbalance in data distribution.
The proposed system focuses on improving fraud detection accuracy while handling practical challenges such as incomplete records and highly imbalanced datasets. Performance evaluation was carried out using metrics such as accuracy, precision, recall, and F1-score to ensure reliable classification results. The experimental results demonstrate that machine learning models can effectively identify suspicious insurance claims and support insurance companies in minimizing fraud-related losses. Future enhancements may include ensemble learning, deep learning approaches, and real-time fraud detection systems for improved performance and scalability.

Keywords: *Insurance fraud detection, Machine learning, Class imbalance dataset, Missing values, Data preprocessing, Fraud classification, Precision, Recall,*

1. Introduction:

Insurance fraud has become one of the major challenges faced by insurance companies, leading to significant financial losses and affecting the efficiency of claim processing systems. Fraudulent claims not only increase operational costs but also impact genuine policyholders through higher premiums and delayed services. With the rapid growth of digital insurance records, large volumes of claim data are generated daily, making manual fraud detection difficult, time-consuming, and prone to errors. To overcome these challenges, machine learning techniques have emerged as effective solutions for automated fraud detection and classification.

Machine learning models are capable of identifying hidden patterns and suspicious activities from historical claim data, enabling accurate classification of fraudulent and genuine claims. This project focuses on developing a machine learning-based insurance fraud detection system for class imbalance datasets containing missing values. By applying preprocessing techniques such as missing value handling, normalization, and data balancing, the model aims to improve classification performance and reliability.

2.Literature

Recent advances in machine learning have significantly improved fraud detection systems through advanced data preprocessing and intelligent classification techniques. Research on fraud detection in imbalanced datasets has shown that traditional machine learning models often struggle with biased predictions due to unequal class distribution and incomplete data records. To address these challenges, several studies have focused on preprocessing methods such as missing value imputation, feature scaling, and oversampling techniques like SMOTE to enhance model reliability and performance.

Various machine learning algorithms including Decision Trees, Random Forest, Support Vector Machine (SVM), Logistic Regression, and ensemble learning methods have been widely applied for insurance fraud detection. Recent studies highlight that ensemble-based approaches and hybrid models provide better accuracy, precision, and recall in detecting fraudulent claims compared to single-model approaches. Researchers have also emphasized the importance of handling missing values effectively, as incomplete datasets can negatively impact classification performance and decision-making accuracy.

Furthermore, advancements in anomaly detection, data balancing methods, and automated feature extraction have improved the capability of machine learning systems to identify suspicious insurance claims in real-world environments. These studies collectively demonstrate that machine learning techniques can effectively support insurance companies in minimizing financial losses and improving fraud investigation processes

https://doi.org/10.63665/IJMEC.1104s.24
ISSN: 2456-4265
IJMEC 2026

3. Methodologies

The proposed insurance fraud detection system follows a structured machine learning pipeline beginning with the collection of insurance claim datasets from publicly available sources and company records, followed by preprocessing steps such as handling missing values, data cleaning, normalization, and feature encoding to ensure data consistency and quality.

Machine learning algorithms are then implemented for hierarchical pattern extraction and fraud classification, along with feature selection and optimization techniques to improve stability and reduce overfitting, culminating in the classification of fraudulent and genuine insurance claims.The model is trained using suitable machine learning techniques while hyperparameters such as learning rate, batch size, and model parameters are carefully tuned. Validation techniques and performance optimization strategies are employed to improve convergence and training efficiency. Performance evaluation is conducted on a separate test dataset using metrics including accuracy, precision, recall, F1-score, ROC-AUC, and confusion matrix analysis to ensure comprehensive assessment of the system performance

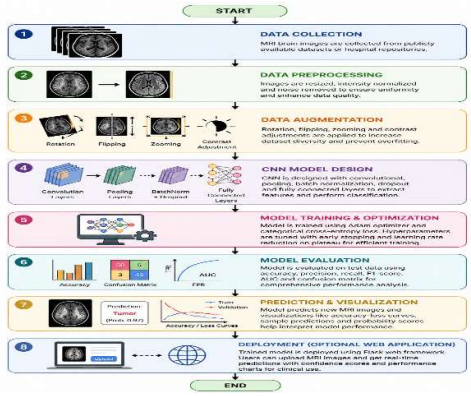

Fig3.1 Methodologies Flow Chart

.4.Implementation

The implementation of the Machine Learning based Insurance Fraud Detection System integrates software-driven machine learning techniques to enable automated fraud identification. The system begins by loading pre-processed insurance claim datasets into the computational environment, where missing values are handled, data is normalized, and features are formatted to match the machine learning model requirements.

The designed machine learning architecture, consisting of data preprocessing, feature selection, classification algorithms, and validation techniques, processes the claim records to extract meaningful patterns relevant to fraud detection. During inference, the trained model evaluates each insurance claim and computes probability scores for two classes: fraudulent and genuine claims. If the predicted probability exceeds a defined classification threshold, the claim is labelled accordingly.

The implementation also includes validation checks to ensure data compatibility and prevent processing errors. Additionally, performance logs and prediction confidence scores are generated to support reliability and transparency. The complete system can be integrated into a user interface or deployed as a web-based application, allowing insurance companies or investigators to upload claim records and obtain real-time fraud classification results efficiently and accurately.

The implemented system is designed to be scalable, efficient, and adaptable to different insurance environments. By automating fraud detection, the system reduces manual investigation effort, minimizes processing time, and improves decision-making accuracy. The modular design also allows future enhancements such as integration with deep learning models, real-time monitoring systems, and cloud-based deployment for large-scale insurance claim analysis.

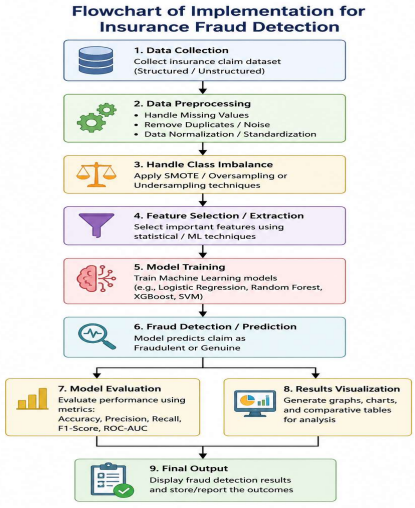

Fig4.1 Implementation Block-Diagram

5. Testing

Testing plays a crucial role in ensuring the reliability, accuracy, and robustness of the proposed Machine Learning based Insurance Fraud Detection System. The primary objective of testing is to identify faults, verify functional correctness, and ensure that the system meets specified requirements without unacceptable failures. A comprehensive test plan is developed to evaluate both general functionality and specialized features across different execution environments, following strict quality assurance procedures.

The testing process includes multiple levels. Unit testing is performed to validate individual modules such as data preprocessing, missing value handling, class imbalance processing, and fraud prediction components, ensuring correct input-output behaviour and logical flow. Functional testing verifies that the system correctly accepts valid insurance claim records, rejects invalid or incomplete data, executes fraud classification functions properly, and produces accurate outputs such as fraudulent/genuine claim predictions with confidence scores.

Integration testing ensures smooth interaction between interconnected modules, including data loading, preprocessing, model training, evaluation, and deployment interfaces. System testing evaluates the fully integrated

https://doi.org/10.63665/IJMEC.1104s.24
ISSN: 2456-4265
IJMEC 2026

model to confirm that it meets overall performance and functional requirements under realistic conditions. Performance testing measures response time, prediction speed, and computational efficiency to ensure timely results suitable for practical use.

Finally, User Acceptance Testing (UAT) validates that the deployed application satisfies end-user expectations, particularly in terms of usability, prediction reliability, and interpretability of results. Overall, the structured testing strategy ensures that each component of the insurance fraud detection system operates accurately, integrates seamlessly, and delivers dependable performance in real-world insurance environments.

6. Results

The proposed insurance fraud detection system was successfully implemented and evaluated using insurance claim datasets. The preprocessing techniques, including missing value handling, normalization, feature encoding, and class balancing methods such as SMOTE, significantly improved the quality and consistency of the dataset, leading to better model performance and reduced bias in fraud prediction.

Machine learning algorithms were trained and optimized using appropriate hyperparameter tuning and validation techniques. The trained model demonstrated effective fraud classification capability by accurately distinguishing between fraudulent and genuine insurance claims. Performance evaluation metrics such as accuracy, precision, recall, F1-score, ROC-AUC, and confusion matrix analysis confirmed the reliability and effectiveness of the system in detecting suspicious claims with improved prediction accuracy.

The final system was also capable of generating fraud probability scores, visualizing prediction results, and processing new insurance claim records efficiently. Experimental results indicate that the proposed machine learning-based approach can support insurance companies in reducing fraudulent activities, minimizing financial losses, and improving the overall claim investigation process.

The results also show that data balancing and preprocessing techniques played a major role in improving fraud detection performance. By handling missing values, reducing noise, and balancing class distribution, the model achieved better generalization and reduced the risk of biased predictions toward majority classes. Feature selection and optimization further enhanced model stability and minimized overfitting, resulting in more consistent and reliable predictions across different test datasets.

Fig 6.1 Snapshot of the Result

7. Conclusion

This project demonstrates the effectiveness of machine learning techniques for automated insurance fraud detection and classification using class imbalance datasets with missing values. The implemented machine learning model successfully distinguishes between fraudulent and genuine insurance claims with high accuracy, providing a reliable, scalable, and computationally efficient solution for insurance claim analysis and fraud prevention. By leveraging preprocessing methods, missing value handling, data balancing techniques, feature selection, and optimized training methods, the system achieves robust performance across diverse insurance claim records, highlighting the growing potential of AI-assisted fraud detection tools in the insurance sector and financial security applications.

The developed framework is capable of analyzing hidden data patterns and identifying suspicious claim activities that may not be easily recognized through manual inspection, thereby improving operational efficiency and reducing human effort in fraud investigation processes. Although the current implementation primarily focuses on binary fraud classification, the framework establishes a strong foundation for future enhancements, including multimodal data integration, ensemble learning techniques, explainable AI methods, and advanced fraud category classification. Overall, this work illustrates how modern machine learning approaches can support automated insurance fraud analysis, reduce financial losses, improve decision-making efficiency, and enhance reliability in identifying fraudulent insurance claims across various real-world insurance environments and applications.

In addition, the proposed system provides a flexible and adaptable architecture that can be integrated into existing insurance management platforms with minimal modifications. Its automated prediction capability enables faster claim processing and early fraud identification, helping insurance companies improve customer service while maintaining secure and reliable claim verification procedures. The use of machine learning also reduces dependency on manual investigations and supports consistent decision-making across large volumes of insurance records.

The project further emphasizes the importance of data quality and preprocessing in developing accurate fraud detection systems. Techniques such as normalization, feature engineering, and class balancing significantly contribute to improving model learning and prediction capability. Proper handling of incomplete and imbalanced datasets ensures that the system can perform effectively even in complex real-world insurance environments where fraudulent cases are relatively rare compared to genuine claims.

Another important outcome of this work is the improvement in interpretability and transparency of fraud prediction results. By generating fraud probability scores, performance metrics, and visualization outputs, the system assists investigators and decision-makers in understanding

https://doi.org/10.63665/IJMEC.1104s.24
ISSN: 2456-4265
IJMEC 2026

model behaviour and identifying high-risk claims efficiently. This improves trust in automated fraud detection systems and supports informed decision-making during insurance claim evaluation processes.

Future research can further enhance the proposed framework by incorporating deep learning architectures, real-time streaming data analysis, cloud-based deployment, and advanced anomaly detection methods. Integration with explainable AI techniques and hybrid ensemble models may also improve prediction accuracy and interpretability. With continuous advancements in artificial intelligence and big data analytics, machine learning-based fraud detection systems are expected to play an increasingly significant role in strengthening financial security, minimizing fraudulent activities, and supporting intelligent insurance management solutions.

8. Future Enhancements

While the current project successfully demonstrates insurance fraud detection using machine learning techniques on class imbalance datasets with missing values, there are several opportunities to further improve its performance, scalability, and practical applicability. Future enhancements could include the integration of additional data sources such as customer transaction history, behavioural patterns, policy details, and claim records to achieve more accurate and comprehensive fraud analysis. Implementing ensemble learning methods or advanced machine learning algorithms could further improve model robustness, prediction capability, and generalization across different insurance datasets and fraud scenarios.

Additionally, expanding the project to classify multiple categories of insurance fraud, rather than only binary fraud detection, would provide more detailed and practically valuable insights for insurance companies and fraud investigation teams. Real-time deployment with optimized processing speed, cloud-based scalability, mobile compatibility, and integration into insurance management systems could also significantly enhance practical usability and accessibility.

Furthermore, incorporating advanced evaluation metrics, explainable AI techniques, visualization methods, and uncertainty analysis would make the system more transparent, interpretable, and reliable for users, analysts, and organizations. These improvements would help investigators better understand model predictions and support more informed decision-making during claim verification processes.

Future versions of the project could also focus on improving adaptability against evolving fraudulent strategies and complex claim manipulation techniques by using continuous learning approaches and anomaly detection systems. Integration with big data technologies and real-time monitoring frameworks can further strengthen system efficiency and scalability in large insurance environments.

Moreover, the implementation of deep learning models and hybrid intelligent systems could enhance feature extraction and fraud identification accuracy for complex datasets. Combining machine learning with natural language processing techniques may also help analyze textual claim descriptions, customer communications, and investigation reports to identify hidden fraud indicators more effectively. Overall, these future enhancements would contribute toward building a more secure, intelligent, efficient, and trustworthy machine learning-assisted framework for automated insurance fraud detection and claim verification in real-world insurance application

9. References

[1] A. A. Khalil, Z. Liu, and A. A. Ali, “Using an adaptive network-based fuzzy inference system model to predict the loss ratio of petroleum insurance in Egypt,” *Risk Management and Insurance Review*, vol. 25, no. 1, pp. 5–18, 2022, doi: 10.1111/rmir.12200.

[2] C. Bockel-Rickermann, T. Verdonck, and W. Verbeke, “Fraud analytics: A decade of research: Organizing challenges and solutions in the field,” *Expert Systems with Applications*, vol. 232, p. 120605, 2023, doi: https://doi.org/10.1016/j.eswa.2023.120605.

[3] Y. Wang and W. Xu, “Leveraging deep learning with LDA-based text analytics to detect automobile insurance fraud,” *Decision Support Systems*, vol. 105, pp. 87–95, 2018, doi: https://doi.org/10.1016/j.dss.2017.11.001.

[4] B. Itri, Y. Mohamed, Q. Mohammed, and B. Omar, “Performance comparative study of machine learning algorithms for automobile insurance fraud detection,” in *2019 Third International Conference on Intelligent Computing in Data Sciences (ICDS)*, 2019, pp. 1–4, doi: 10.1109/ICDS47004.2019.8942277.

[5] R. P. B. Piovezan, P. P. de Andrade Junior, and S. L. Ávila, “Machine Learning Method for Return Direction Forecast of Exchange Traded Funds (ETFs) Using Classification and Regression Models,” *Computational Economics*, 2023, doi: 10.1007/s10614-023-10385-4.

[6] A. A. Khalil, Z. Liu, A. Salah, A. Fathalla, and A. Ali, “Predicting Insolvency of Insurance Companies in Egyptian Market Using Bagging and Boosting Ensemble Techniques,” *IEEE Access*, vol. 10, pp. 117304–117314, 2022, doi: 10.1109/ACCESS.2022.3210032.

[7] N. Boodhun and M. Jayabalan, “Risk prediction in life insurance industry using supervised learning algorithms,” *Complex & Intelligent Systems*, vol. 4, no. 2, pp. 145–154, 2018, doi: 10.1007/s40747-018-0072-1.

[8] D. Tiwari, B. Nagpal, B. S. Bhati, A. Mishra, and M. Kumar, “A systematic review of social network sentiment analysis with comparative study of ensemble-based techniques,” *Artificial Intelligence Review*, vol. 56, no. 11, pp. 13407–13461, 2023, doi: 10.1007/s10462-023-10472-w.

[9] M. Liao, S. Tian, Y. Zhang, G. Hua, W. Zou, and X. Li, “PDA: Progressive Domain Adaptation for Semantic Segmentation,” *Knowledge-Based Systems*, vol. 284, p. 111179, 2024, doi: https://doi.org/10.1016/j.knosys.2023.111179.

[10] A. Khalil, Z. Liu, and A. Ali, “Precision in Insurance Forecasting: Enhancing Potential with Ensemble and Combination Models based on the Adaptive Neuro Fuzzy

https://doi.org/10.63665/IJMEC.1104s.24
ISSN: 2456-4265
IJMEC 2026

Inference System in the Egyptian Insurance Industry," *Applied Artificial Intelligence*, vol. 38, no. 1, p. 2348413, 2024, doi: 10.1080/08839514.2024.2348413.
[11] A. K. I. Hassan and A. Abraham, "Modeling insurance fraud detection using ensemble combining classification," *International Journal of Computer Information Systems and Industrial Management Applications*, vol. 8, pp. 257–265, 2016.
[12] V. R. Shetty and R. L. Malghan, "Safeguarding against Cyber Threats: Machine Learning-Based Approaches for Real-Time Fraud Detection and Prevention," *Engineering Proceedings*, vol. 59, no. 1, p. 111, 2023.
[13] A. R. Khalid, N. Owoh, O. Uthmani, M. Ashawa, J. Osamor, and J. Adejoh, "Enhancing Credit Card Fraud Detection: An Ensemble Machine Learning Approach," *Big Data and Cognitive Computing*, vol. 8, no. 1, p. 6, 2024.
[14] A. A. Khalil, Z. Liu, and A. Ali, "Enhancing operational efficiency of insurance companies: a fuzzy time series approach to loss ratio forecasting in the Egyptian market," *Journal of Business Analytics*, pp. 1–19, 2024, doi: 10.1080/2573234X.2024.2393609.
[15] M. Hanafy and R. Ming, "Improving imbalanced data classification in auto insurance by the data level approaches," *International Journal of Advanced Computer Science and Applications*, vol. 12, no. 6, 2021.
[16] B. Baesens, S. Höppner, I. Ortner, and T. Verdonck, "robROSE: A robust approach for dealing with imbalanced data in fraud detection," *Statistical Methods & Applications*, vol. 30, no. 3, pp. 841–861, 2021, doi: 10.1007/s10260-021-00573-7.
[17] S. Subudhi and S. Panigrahi, "Effect of Class Imbalanceness in Detecting Automobile Insurance Fraud," in *2018 2nd International Conference on Data Science and Business Analytics (ICDSBA)*, 2018, pp. 528–531, doi: 10.1109/ICDSBA.2018.00104.
[18] T. Olalekan Yusuf and A. Rasheed Babalola, "Control of insurance fraud in Nigeria: an exploratory study (case study)," *Journal of Financial Crime*, vol. 16, no. 4, pp. 418–435, 2009, doi: 10.1108/13590790910993744.
[19] R. Bhowmik, "Detecting auto insurance fraud by data mining techniques," *Journal of Emerging Trends in Computing and Information Sciences*, vol. 2, no. 4, pp. 156–162, 2011.
[20] K. Nian, H. Zhang, A. Tayal, T. Coleman, and Y. Li, "Auto insurance fraud detection using unsupervised spectral ranking for anomaly," *The Journal of Finance and Data Science*, vol. 2, no. 1, pp. 58–75, 2016, doi: https://doi.org/10.1016/j.jfds.2016.03.001.
[21] G. Kowshalya and M. Nandhini, "Predicting Fraudulent Claims in Automobile Insurance," in *2018 Second International Conference on Inventive Communication and Computational Technologies (ICICCT)*, 2018, pp. 1338–1343, doi: 10.1109/ICICCT.2018.8473034.
[22] L. Goleiji and M. Tarokh, "Identification of influential features and fraud detection in the Insurance Industry using the data mining techniques (Case study: automobile's body insurance)," *Majlesi Journal of Multimedia Processing*, vol. 4, pp. 1–5, 2015.
[23] S. Goundar, S. Prakash, P. Sadal, and A. Bhardwaj, "Health Insurance Claim Prediction Using Artificial Neural Networks," *International Journal of System Dynamics Applications (IJSDA)*, vol. 9, no. 3, pp. 40–57, 2020.
[24] J. Debener, V. Heinke, and J. Kriebel, "Detecting insurance fraud using supervised and unsupervised machine learning," *Journal of Risk and Insurance*, vol. 90, no. 3, pp. 743–768, 2023, doi: https://doi.org/10.1111/jori.12427.
[25] A. Urunkar, A. Khot, R. Bhat, and N. Mudegol, "Fraud Detection and Analysis for Insurance Claim using Machine Learning," in *2022 IEEE International Conference on Signal Processing, Informatics, Communication and Energy Systems (SPICES)*, 2022, pp. 406–411, doi: 10.1109/SPICES52834.2022.9774071.
[26] Y. Abakarim, M. Lahby, and A. Attioui, "A Bagged Ensemble Convolutional Neural Networks Approach to Recognize Insurance Claim Frauds," *Applied System Innovation*, vol. 6, no. 1, 2023, doi: 10.3390/asi6010020.
[27] B. Xu, Y. Wang, X. Liao, and K. Wang, "Efficient fraud detection using deep boosting decision trees," *Decision Support Systems*, vol. 175, p. 114037, 2023, doi: https://doi.org/10.1016/j.dss.2023.114037.
[28] S. Subudhi and S. Panigrahi, "Use of optimized Fuzzy C-Means clustering and supervised classifiers for automobile insurance fraud detection," *Journal of King Saud University - Computer and Information Sciences*, vol. 32, no. 5, pp. 568–575, 2020, doi: https://doi.org/10.1016/j.jksuci.2017.09.010.
[29] A. Jadhav, D. Pramod, and K. Ramanathan, "Comparison of Performance of Data Imputation Methods for Numeric Dataset," *Applied Artificial Intelligence*, vol. 33, no. 10, pp. 913–933, 2019, doi: 10.1080/08839514.2019.1637138.
[30] G. G. Sundarkumar, V. Ravi, and V. Siddeshwar, "One-class support vector machine based undersampling: Application to churn prediction and insurance fraud detection," in *2015 IEEE International Conference on Computational Intelligence and Computing Research (ICCIC)*, 2015, pp. 1–7, doi: 10.1109/ICCIC.2015.7435726.

https://doi.org/10.63665/IJMEC.1104s.24
ISSN: 2456-4265
IJMEC 2026