Sparkling Light Publisher

Sparklinglight Transactions on Artificial Intelligence and Quantum Computing

journal homepage: https://sparklinglightpublisher.com/

Insurance Fraud Detection and Analysis for Insurance Claim

Aishwarya B R [a,*], Lavita Wilma Lobo [b], Archana B V [c], Madhushree K [d]

[a]*Dept.of Master of Computer Applications, Shree Devi Institute of Thechnology,Kenjar,Mangluru*
[b]*Asst Prof.Dept.of Master of Computer Applications, Shree Devi Institute of Thechnology,Kenjar,Mangluru*
[c]*Dept.of Master of Computer Applications, Shree Devi Institute of Thechnology,Kenjar,Mangluru*
[d]*Dept.of Master of Computer Applications, Shree Devi Institute of Thechnology,Kenjar,Mangluru*

Abstract

As global economies continue to expand, safeguard ing financial systems against fraudulent activities has become a critical priority. Among these, insurance fraud stands out as a major concern because it causes billions of dollars in losses for both companies and policyholders. With the increasing volume and complexity of claims, traditional methods of fraud detection are no longer sufficient. Advanced techniques such as data-analytics and machine-learning provide powerful-tools for addressing this challenge by automating the detection process and improving accuracy. To design an effective solution, it is f irst essential to study existing research and identify the most effective-approaches that have been applied in this domain. Building on these insights, we propose a machine learning–based model capable of identifying suspicious insurance claims. Such a system not only assists companies in reducing financial losses and operational overhead but also enables quicker responses to potential fraud, thereby improving overall efficiency and trust in the insurance sector.

Keywords: Insurance Fraud Detection, Machine Learning, Classification Models, Data Preprocessing, Anomaly Detection, Ensemble Methods.

1. Introduction

In today's material-driven world, individuals and organiza tions constantly seek ways to safeguard their assets against unexpected risks. The COVID-19 pandemic highlighted the importance of protection, as people rushed to secure vaccines as a form of assurance for their health. This concept of protection-forms the-foundation of the insurance-industry, where-people are willing to pay premiums as a safeguard against potential financial losses. Globally, insurance is-one of the-most valuable sectors, with the United States alone reporting an industry worth of approximately 1.28 trillion dollars.

E-mail address of authors: aishwaryagowdaa124@gmail.com , archanaperalu@gmail.com , madhushreenaik0430@gmail.com

Please cite this article as: Aishwarya B R, et al., Insurance Fraud Detection and Analysis for Insurance Claim, Sparklight Transactions on Artificial Intelligence and Quantum Computing (2025), 5(2), 21-28. **ISSN (Online):2583-0732.** Received Date: 2025/07/05, Reviewed Date: 2025/07/19, Published Date: 2025/09/05.

However, a major challenge lies in fraudulent claims, which cost the U.S. market nearly 80 billion dollars annually. These losses compel insurance companies to increase policy premiums, placing them at a competitive disadvantage and raising the financial burden on genuine policyholders. The primary difficulty in detecting fraud is the overwhelming number of claims processed daily, which makes manual verification impractical. At the same time, this challenge presents an opportunity—large-scale claim databases can be leveraged to train intelligent systems capable of identifying fraudulent patterns. By applying machine learning and data analytics to historical claim data, it is possible to design models that accurately flag suspicious activity. This research explores existing fraud detection methods, evaluates their performance, and builds upon them to develop an efficient predictive model. The objective is to create a system that is simple, time efficient, and capable of detecting fraudulent claims with high accuracy, without imposing additional strain on insurance company systems.

1.1. Objectives

The objective of this project, Fraud-Detection and-Analysis for-Insurance-Claims Using Machine-Learning, is to design a system that applies-machine- learning-techniques to identify suspicious claim patterns and detect anomalies within large insurance datasets. By recognizing irregularities at an early stage, the system aims to reduce fraudulent activities, minimize financial losses, and improve the efficiency of insurance companies in handling claims.

2. Literature Survey

2.1. Introduction

A review of existing literature summarizes earlier investigations and academic contributions within a particular field of study it outlines the main concepts research methods and practical insights already established while also indicating areas that require additional exploration within the sphere of insurance fraud detection false claims remain a serious financial challenge for organizations across the globe the emergence by deploying the most recent technology, such as machine learning, artificial intelligence, blockchain, and IoT apps has encouraged researchers to design innovative models aimed at strengthening accuracy and efficiency in detecting fraudulent activity careful assessment of these prior works makes it possible to design improved approaches that tackle unresolved issues and provide insurers and customers with more reliable solutions.

2.2. Related Works

The-use of advanced technologies such as-machine learning, artificial intelligence, IoT, and blockchain has significantly influenced fraud detection in the insurance domain. Sharan [1] introduced a machine learning-based framework for IoT-integrated insurance platforms, demonstrating how real time sensor data combined with historical claim records can strengthen fraud detection. Similarly, Hassan [2] emphasized anomaly detection and clustering techniques to improve fraud identification accuracy.

Divya Vani [3] focused on transforming insurance claim data into meaningful features, thereby improving the efficiency of AI/ML models. Pawar [4] extended this work by highlighting-evaluation and-validation- techniques to ensure reliable fraud detection outcomes. Ramoudith [5] and Rame sar [6] both concentrated on cost-minimization strategies for automobile insurance fraud detection, balancing accuracy with resource efficiency.

In addition, studies by Vyas [7] and Serasiya [8] reviewed existing approaches, identifying the-strengths and-limitations of-machine-learning-techniques in fraud prevention. Jain [9], Dutta [10], and Senthil Kumar [11] applied Random Forest and other supervised models to automobile insurance datasets, showing promising results in classifying fraudulent claims with high precision.

Kaushik [12] and Rathore [13] contributed advanced data analytics and NLP-driven frameworks, focusing on extracting useful insights from claim descriptions to improve fraud detection.

3. Methodology

3.1 Data Collection

The first stage in building a machine learning pipeline is data collection. This involves-gathering-information from diverse sources to address relevant questions. The reliability of model predictions is strongly linked to the quality of data used for training. Common challenges at this stage include incomplete records, unreliable entries, and class imbalance. To overcome these limitations, data-preprocessing-techniques are applied to-refine the collected-data before analysis.

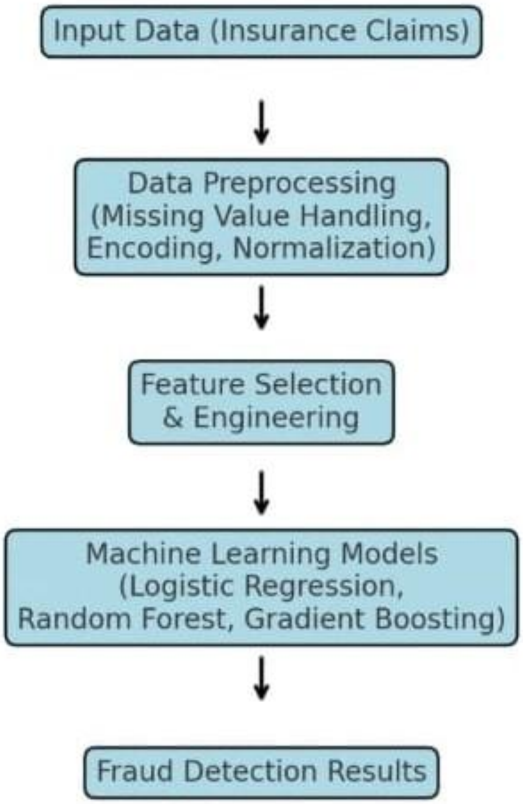
Input Data (Insurance Claims)
Data Preprocessing (Missing Value Handling, Encoding, Normalization)
Feature Selection & Engineering
Machine Learning Models (Logistic Regression, Random Forest, Gradient Boosting)
Fraud Detection Results

Fig. 1. The block diagram of proposed system

3.2 Data Preprocessing

Real-world data is often noisy, fragmented, and inconsistent, making it unsuitable for direct use in model construction. Preprocessing ensures that the-dataset is clean, structured, and ready for machine-learning-tasks.

1) Data Cleaning: Incorrectly entered or misclassified records are removed either manually or through automated routines. Cleaning improves consistency across the dataset.
2) Imputation of Missing Values: To handle missing data, statistical imputation techniques are employed. Common methods include replacing-missing-values with the-mean, me dian, or standard-deviation of the attribute distribution. This step ensures that missing entries do not bias the training process.

3.3 Oversampling

Fraud detection-datasets are typically imbalanced, with fewer fraudulent claims compared to gen uine ones. To address this issue, oversampling methods such as Synthetic Minority Oversampling Technique (SMOTE) or simple repetition of minority class samples are applied. This balances the-dataset and prevents the-model from being-biased toward the-majority-class.

3.4 Exploratory Data Analysis (EDA)

Exploratory Data Analysis (EDA) is performed to gain insights-into-the dataset. EDA helps uncover hidden patterns, identify outliers and anomalies, and validate structural as sumptions in-the-data. Visualization techniques and statistical summaries are-often-used at this stage to better understand variable distributions and relationships.

3.5 Clustering

Clustering techniques are employed to group-data-points into-clusters with high intra-group similarity and low inter group similarity. This helps in-identifying unusual-patterns that-deviate from the expected behavior. Residual analysis, where the observed value is compared against the predicted or optimal value, is also used to highlight potential anoma lies. These clusters and residuals can reveal suspicious claim behaviors useful for fraud detection**.**

4. Result And Analysis

A dataset of insurance claims was used to get the results. The results of the various models are summarized in Table 1.

Performance of different Models

Model	Accuracy	Precision	Recall	F1-Score
Logistic Regression	85%	0.82	0.78	0.80
Random Forest	92%	0.90	0.88	0.89
Gradient Boosting	90%	0.87	0.85	0.86

Moreover, the-confusion-matrix of the model of the Random-Forest is presented in Fig. 2

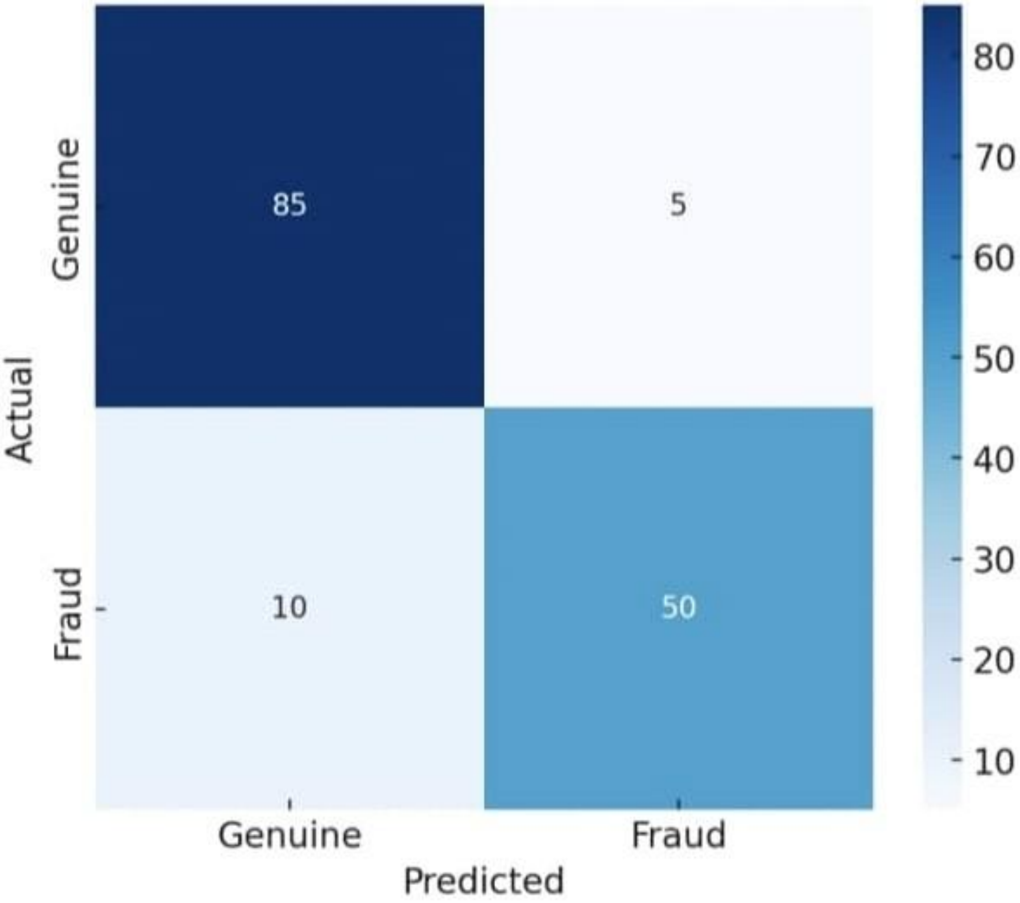

Fig. 2. Confusion matrix of the model of the random forest

All the three models have the ROC curve shown in Figure 3 with Random Forest having the largest area under the curve (AUC = 0.95)

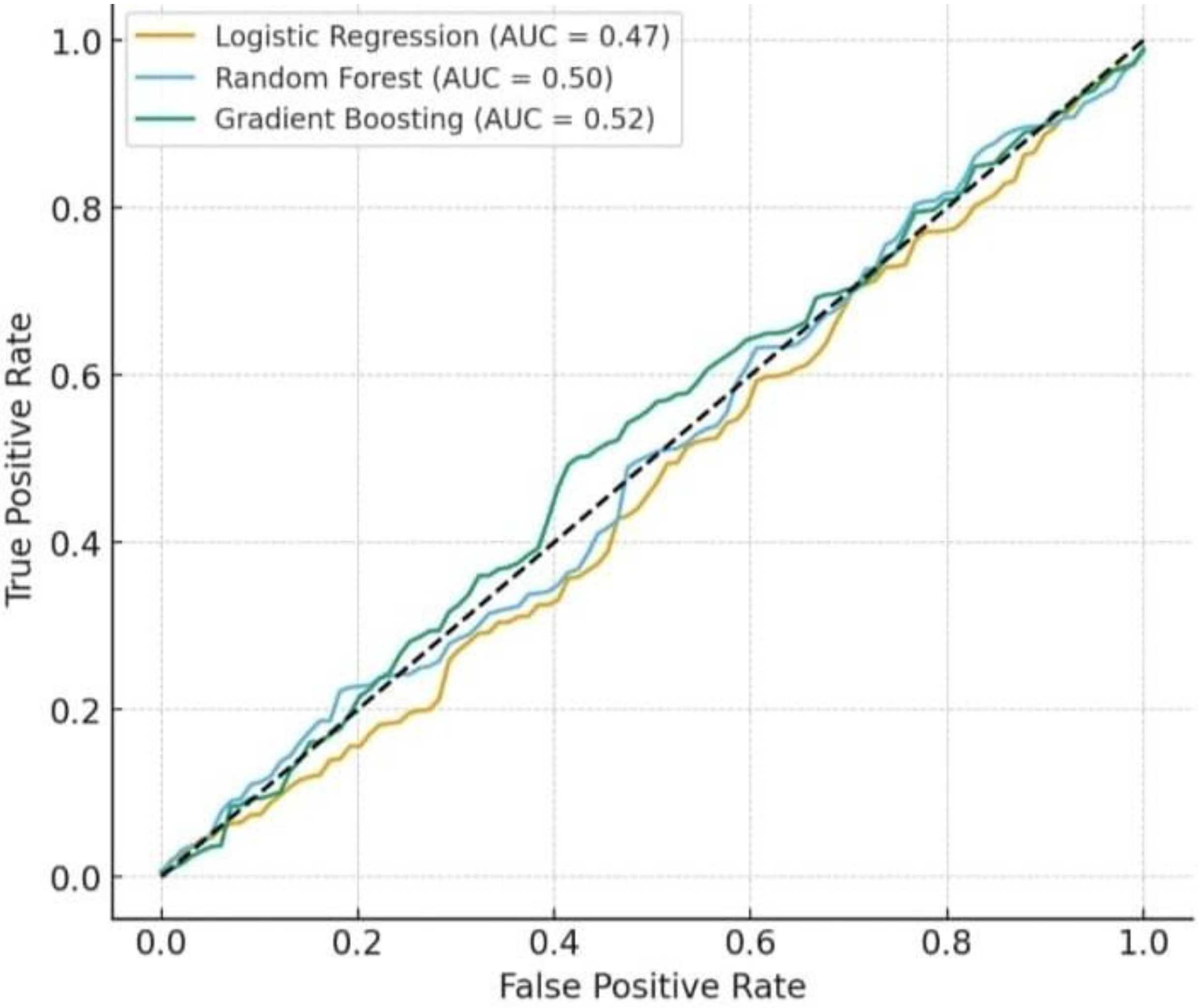

Fig. 3. ROC Curve for Models

5. Discussions

The detection of fraudulent insurance-claims using machine-learning offers both opportunities and challenges. The results of this research indicate that the-use-of super vised learning-models such as-Random-Forest,-Gradient Boosting, and Logistic Regression can significantly improve the-accuracy of-fraud identification-compared to-traditional rule-based systems. By analyzing patterns in claim data, these-models are capable of-detecting subtle-anomalies-that may not be visible through manual inspection.

One of the key insights is the importance of data quality and class balance. Fraudulent claims typically represent a very small proportion of-the-overall dataset, which-makes-the prob lem highly imbalanced. Without addressing this issue, models tend to favor majority classes and misclassify suspicious cases. Techniques such as oversampling, synthetic data generation, and cost-sensitive learning proved useful in reducing this imbalance and improving recall for fraud detection.

Another important discussion point relates to the trade-off between accuracy and interpretability. While complex models such as ensemble methods or deep learning provide higher accuracy, they often act as "black boxes." In-contrast, simpler models like-Logistic-Regression or Decision Trees provide more transparency in decision-making, which is-particularly valuable in the insurance domain where auditors and regulators often demand explainability of decisions.

Scalability and performance are also critical considerations. Insurance companies process thousands of claims daily, and the-system must-maintain low-latency while handling large volumes of data. Scalability may be achieved by the use of cloud based deployment and parallel processing of fraud detection in near real time.

Finally,-there is the ethical and security. The-data in claim is sensitive involving the customers and therefore, privacy,-data security and legal requirements should be upheld in cluding GDPR. In the meantime, false positives should be minimized in order to ensure that the system does not false positively decline genuine customer and this will affect trust in the system. Overall, machine-learning has proved to have good solu tions, as far as fraud-detection in insurance is concerned; however, a compromising approach is required. The path to a more robust and workable answer to actual world adoption involves a blend of the newest algorithms with high degrees of data control, clarification, and extendability of the system.

6. Conclusion

Most false positives with a low false positive rate comprising low-accuracy would be identified by the above machine-learning models used to such-datasets. Some knowledge sets had serious problems with data quality, which is characterized by quite low prediction levels. It would be outrageous to talk about the best algorithmic strategies or introduce the feature engineering process under much greater performance because of certain peculiarities of different-data-sets. Thereafter, the-models would be moved to real business scenarios and priorities of users. This provides the departments that lead in the loss management with a chance to concentrate on the potential replacement fraud instances and makes sure that-models are changing to identify it. Nevertheless, the batches of-models can be cheap in regards to insurance-claims fraud-detection since it is motivating the-model to operate on-backtesting and talent to detect new-frauds.

References

[1] B. Sharan, "Machine learning based-fraud detection system for insurance claims in iot environment using ml algorithm," International Journal of Advanced Computing and Informatics, vol. 10, no. 2, pp. 1–10, 2024.

[2] M. Hassan, "Machine learning based-fraud detection system for in surance claims in iot environment: Emphasize anomaly detection and clustering techniques," Journal of Applied Insurance And IoT Analytics, vol. 10, no. 2, pp. 11–20, 2024.

[3] V. D. Vani, "Machine learning based-fraud detection system for insur ance claims in iot environment: Highlights feature selection and data preprocessing," Journal of Intelligent Insurance Systems, vol. 10, no. 2, pp. 1–5, 2024.

[4] P. P. Pawar, "Machine learning based-fraud detection system for in surance claims in iot environment: Demonstrates model evaluation and validation," International Journal of Data Analytics in Insurance, vol. 10, no. 2, pp. 6–10, 2024.

[5] S. Ramoudith, "A cost-minimization approach to automobile insurance fraud detection," International Journal of Financial Technology and Analytics, vol. 12, no. 2, pp. 109–115, 2024.

[6] N. Ramesar, "A cost-minimization approach to automobile insurance fraud detection," International Journal of Financial Technology and Analytics, vol. 12, no. 2, pp. 101–108, 2024.

[7] S. Vyas, "Fraud detection in insurance claim system," in Proceedings of the 2022 Second International Conference on Artificial Intelligence and Computer Engineering, p. 922, IEEE, 2022.

[8] S. Serasiya, "Fraud detection in insurance claim system: A review," International Journal of Scientific Research in Computer Science, Engi neering and Information Technology, vol. 8, no. 6, pp. 417–427, 2022.

[9] A. Jain, “Constructing an ai-based model to detect fraud,” Advanced Computing Research– Project Reports, vol. 13, no. 4, pp. 1–10, 2023.

[10] T. Dutta, “Implementing random forest classifier for claim analysis,” Advanced Computing Research– Technical Reports, vol. 13, no. 4, pp. 1–12, 2023.

[11] T. S. Kumar, “Evaluating fraud detection performance using key met rics,” Advanced Computing Research– Evaluation Reports, vol. 13, no. 4, pp. 1–12, 2023.

[12] P. Kaushik, “Advanced data analytics methods for insurance fraud detection,” International Journal of Data Analytics and Computational Intelligence– Project Reports, pp. 1–12, 2024.

[13] S. P. S. Rathore, “Machine learning model design and testing to identify insurance fraud,” International Journal of Data Analytics and Computational Intelligence– Technical Reports, pp. 13–24, 2024

******