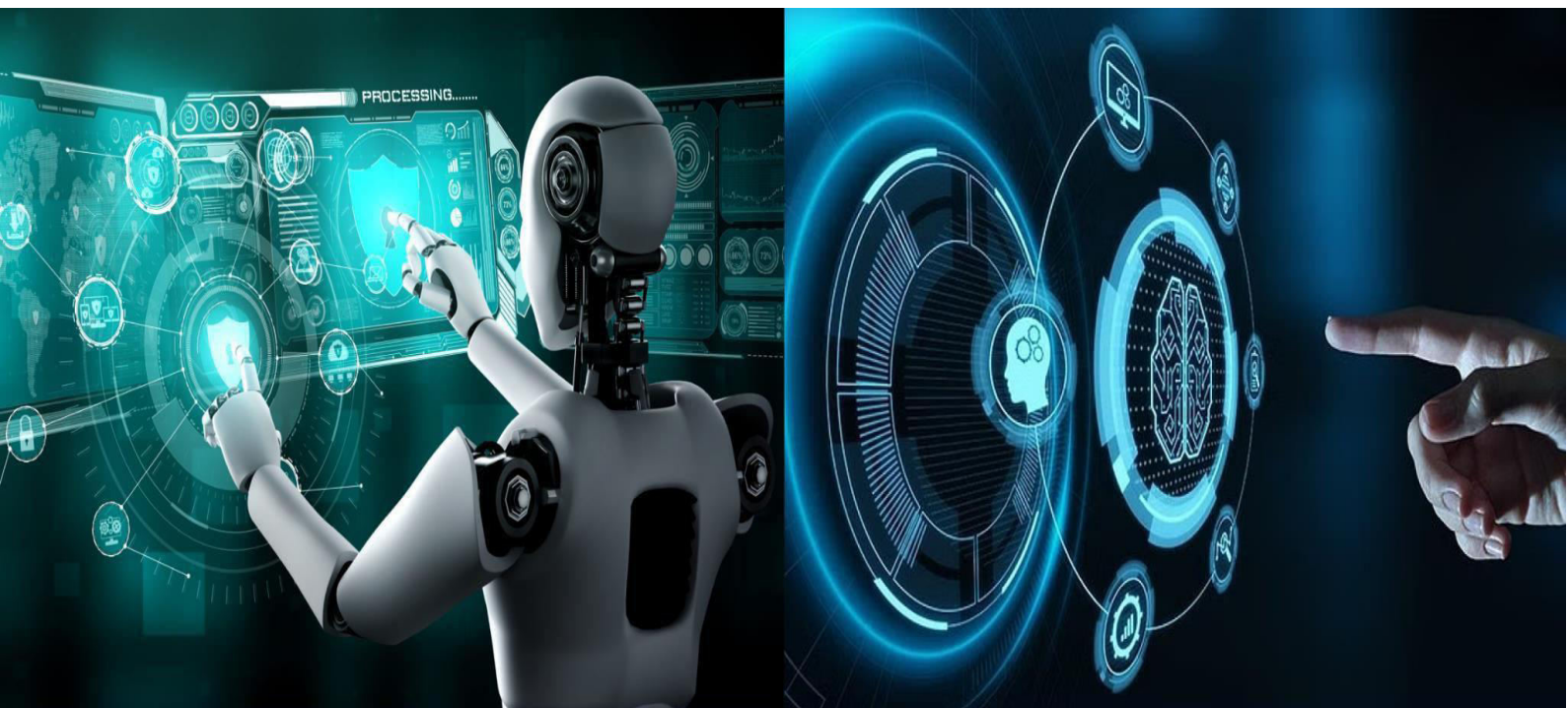




International Journal of Innovative Research in Computer and Communication Engineering

(A Monthly, Peer Reviewed, Refereed, Scholarly Indexed, Open Access Journal)





Vehicle Insurance Claims Fraud Detection using Machine Learning

Apoorva S M, Dr. P Devaki

Dept. of ISE, National Institute of Engineering, Mysore, Karnataka, India

Professor, Dept. of ISE, National Institute of Engineering, Mysore, Karnataka, India

ABSTRACT: Fraudulent practices in vehicle insurance have emerged as a significant concern, leading to considerable financial losses for insurers globally. Such fraud occurs when false or misleading information is provided to obtain benefits that are not legitimately entitled. Traditional detection approaches, often reliant on manual review or rule-based systems, struggle to process and analyse the large, complex datasets associated with modern claim records. To address this limitation, the present study applies machine learning techniques to classify claims as either genuine or fraudulent. Several algorithms—Logistic Regression, Decision Tree, Random Forest, K-Nearest Neighbours, Support Vector Machine, and Gaussian Naïve Bayes—are trained on a historical dataset of insurance claims. Model performance is assessed using precision, recall, and F1-score metrics. Results indicate that machine learning methods can significantly enhance both the accuracy and efficiency of fraud detection, enabling insurers to make faster, more informed decisions with increased confidence.

KEYWORDS: Vehicle insurance, fraud detection, machine learning, Logistic Regression, Decision Tree, Random Forest, K-Nearest Neighbors, Support Vector Machine, Gaussian Naïve Bayes

I. INTRODUCTION

The insurance sector has increasingly recognized the critical importance of implementing effective fraud management systems, as fraudulent activities can cause substantial financial losses and erode customer trust. Insurance fraud is generally classified into two main categories: hard fraud and soft fraud. Hard fraud occurs when an individual deliberately fabricates an accident, loss, or incident to obtain undue benefits, whereas soft fraud involves exaggerating or partially falsifying a legitimate claim to increase the payout. A well-designed fraud detection and prevention framework not only protects insurers from such activities but also enhances customer satisfaction by ensuring fairness and transparency. Higher satisfaction levels can, in turn, contribute to reducing operational expenses, including costs related to loss adjustment.

In recent years, fraudulent practices in the motor and vehicle insurance domain have emerged as a significant concern. Such activities may originate from multiple sources, including policyholders, intermediaries, and, in some cases, internal staff, with the latter two demanding closer scrutiny from a governance and compliance standpoint. Fraudulent claims can manifest in various forms, such as misrepresenting incident details, staging events, inflating reported damages, or fabricating losses. Examples include inventing scenarios not covered under a policy, altering the reported cause of an accident to avoid liability, neglecting mandatory safety measures, or artificially increasing the claimed damage beyond the actual extent of the loss.

Traditional fraud detection methods often rely on manual investigation and predefined rules, which can be time-consuming and require expertise across multiple domains. In contrast, machine learning techniques offer a more efficient and scalable solution by analysing large volumes of claim data to detect suspicious patterns and anomalies. By adopting such data-driven approaches, insurance companies can improve the speed and accuracy of fraud detection while reducing the resources needed for investigations.

II. LITERATURE REVIEW

Fraud in the insurance sector is becoming an increasingly serious issue, draining billions in revenue each year and eroding the trust that policyholders place in their providers. Because of its impact on both financial stability and



International Journal of Innovative Research in Computer and Communication Engineering (IJIRCCCE)

(A Monthly, Peer Reviewed, Refereed, Scholarly Indexed, Open Access Journal)

customer confidence, detecting and preventing fraud has become a priority for insurers and a growing focus of research.

For many years, identifying fraudulent claims was largely a manual process. Investigators would rely on fixed sets of rules, often built from patterns observed in previous fraud cases. These rule-based approaches could handle simple, recurring schemes well, but they struggled to keep pace with more sophisticated or unfamiliar tactics—especially when the data involved was vast, varied, and complex.

As technology advanced and access to large datasets became easier, fraud detection methods began to shift. Statistical analysis and data mining tools were introduced, allowing for deeper pattern recognition. Logistic regression, for instance, has been used to estimate the likelihood that a claim is fraudulent based on historical records, while decision trees have gained popularity for their straightforward logic and ability to present results in a way that is easy for non-technical teams to understand.

In recent years, the field has moved even further into the realm of machine learning. Advanced models such as Random Forest, Gradient Boosting, and XGBoost are particularly effective at uncovering subtle, non-linear relationships in data. Other algorithms, including Naïve Bayes and Support Vector Machines (SVM), remain valuable because they can process both numerical and categorical data types efficiently. Research consistently shows that these modern techniques can outperform older statistical models, particularly in areas like precision (reducing false positives) and recall (catching more actual fraud cases).

Some approaches now blend supervised learning (where the system learns from labeled examples) with unsupervised methods (which detect unusual patterns without prior labeling). Anomaly detection tools like Isolation Forest and Local Outlier Factor (LOF) are used to highlight claims that differ significantly from the norm, warranting closer investigation. At the same time, explainable AI (XAI) techniques—such as SHAP values—are being adopted to shed light on why a claim is flagged, helping insurers maintain trust in automated decision-making.

Even with these advances, the task remains challenging. Fraud cases are typically rare in comparison to legitimate claims, leading to heavily imbalanced datasets that can bias models toward predicting “non-fraud.” Some studies also rely on artificially generated data, which may not fully match the complexity of real-world cases. And in industry, there’s always the need to balance accuracy with clarity—highly accurate models are of limited use if their decisions cannot be explained to auditors or regulatory bodies.

This study addresses these challenges by testing and comparing several machine learning algorithms on a dataset of real vehicle insurance claims. The aim is to identify the model that not only delivers strong fraud detection accuracy but also maintains the transparency and ease of integration needed for real-world deployment in the insurance industry.

III. METHODOLOGY

The proposed approach for detecting fraudulent vehicle insurance claims follows a structured, multi-phase process consisting of data acquisition, pre-processing, feature engineering, model construction, and performance assessment. The overall workflow is depicted in Fig. 1.

1)Dataset

The dataset comprises historical automobile insurance claim records containing attributes such as policyholder demographics, incident descriptions, claim amounts, insured occupation, insured hobbies, and incident location details. The target label specifies whether each claim is classified as fraudulent or legitimate. Given that fraudulent claims form only a small proportion of the total records, the dataset exhibits a class imbalance, which must be addressed to ensure robust model performance.

2)Data Pre-processing

To prepare the dataset for model training, several pre-processing steps are applied:

- **Handling Missing Data:** Missing entries are detected and filled using appropriate imputation strategies, with categorical and numerical attributes treated separately.
- **Encoding Categorical Variables:** Nominal and ordinal features are converted into numerical form using encoding techniques such as one-hot encoding or label encoding to ensure compatibility with machine learning algorithms.



International Journal of Innovative Research in Computer and Communication Engineering (IJIRCCCE)

(A Monthly, Peer Reviewed, Refereed, Scholarly Indexed, Open Access Journal)

- **Scaling Numerical Attributes:** Normalization or standardization is performed for features where differences in scale could affect model performance, particularly for algorithms like KNN and SVM.
- **Outlier Treatment:** Extreme or anomalous values that may distort training are identified and mitigated.

3) Feature Engineering and Selection

To enhance predictive accuracy and computational efficiency, non-informative and redundant features are eliminated. Techniques employed include:

- **Pearson Correlation Analysis** to remove features with high collinearity.
- **Chi-Square Test** to evaluate the association between categorical attributes and the target variable.
- **Recursive Feature Elimination (RFE)** to identify the most relevant predictors.

4) Model Development

Several machine learning algorithms are implemented for comparison:

- **Logistic Regression:** A statistical approach for binary classification problems.
- **Decision Tree:** A rule-based model that partitions data through hierarchical decision splits.
- **Random Forest:** An ensemble of decision trees designed to minimize overfitting and improve generalization.
- **K-Nearest Neighbors (KNN):** A distance-based classification method.
- **Support Vector Machine (SVM):** A classifier that determines an optimal separating hyperplane in feature space.
- **Gaussian Naïve Bayes:** A probabilistic model grounded in Bayes' theorem with the assumption of feature independence.

5) Model Training and Validation

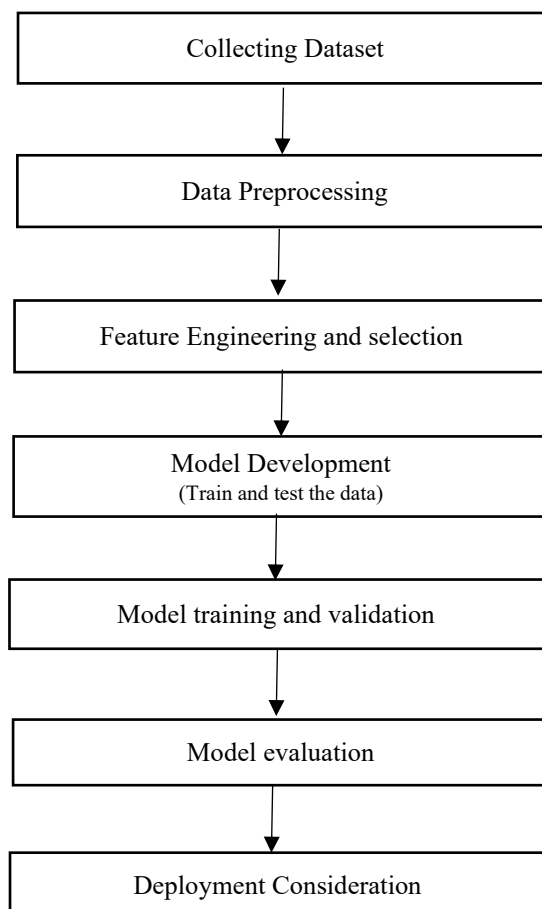


Fig 1: Proposed methodology for vehicle insurance fraud detection.



International Journal of Innovative Research in Computer and Communication Engineering (IJIRCCCE)

(A Monthly, Peer Reviewed, Refereed, Scholarly Indexed, Open Access Journal)

The dataset is divided into training and testing sets, typically using an 80:20 ratio. Models are trained on the training set, and hyperparameters are optimized via cross-validation. In certain cases, clustering methods such as K-Means are applied before classification to group similar claim records, thereby improving fraud detection capability.

6) Model Evaluation

The effectiveness of each model is measured using multiple performance metrics:

- **Accuracy:** The overall percentage of claims correctly classified as either fraudulent or legitimate.
- **Precision:** The proportion of correctly identified fraudulent claims out of all claims predicted as fraudulent.
- **Recall:** The proportion of actual fraudulent claims that were correctly detected by the model.
- **F1-Score:** The harmonic mean of precision and recall, providing a balanced measure of both metrics.

7) Deployment Considerations

Following evaluation, the model demonstrating the most favourable performance across the metrics is chosen for potential integration into a real-time fraud detection framework. Key factors considered prior to deployment include:

- Periodic retraining to ensure adaptability to changing fraud tactics.
- Preserving interpretability to meet compliance and regulatory audit requirements.
- Implementing strict measures for data confidentiality and system security.

IV. RESULTS AND DISCUSSIONS

The performance of a predictive model can be examined using multiple evaluation metrics. In this project, the metrics adopted are as follows

Model	Train Accuracy	Test Accuracy	Precision	Recall	F1 Score
Logistic Regression	0.785	0.805	0.45	0.551	0.4954
Decision Tree	1	0.785	0.5424	0.6531	0.5926
Random Forest	0.993	0.760	0.6053	0.4694	0.5287
KNN	0.752	0.760	0.2893	0.7143	0.4118
SVM	0.843	0.755	0.6	0.3673	0.4557
Naive Bayes	0.718	0.720	0.4419	0.7755	0.563

Table 1: Model Performance Metrics for Fraud Detection

In this study, multiple machine learning models were evaluated using training accuracy, testing accuracy, precision, recall, and F1-score to measure their effectiveness in detecting fraudulent insurance claims. Among all tested algorithms, the Decision Tree classifier was selected for deployment in the web-based fraud detection platform due to its reliable performance, ease of integration, and transparent decision-making process.

As shown in Table 1, the Decision Tree achieved a perfect training accuracy of 1.000 and a testing accuracy of 0.785. While this reflects a strong learning capability, the gap between the two scores indicates a degree of overfitting. The model's precision score of 0.5424 means that just over half of the claims flagged as fraudulent were correctly identified. A recall score of 0.6531 shows that the model successfully detected nearly two-thirds of actual fraudulent cases. The resulting F1-score of 0.5926 demonstrates a reasonable balance between precision and recall, making the model suitable for use cases where both missed detections and false positives must be managed carefully.

The Decision Tree was chosen not only for its competitive results but also for its interpretability—an essential factor in financial and insurance applications. It's clear, rule-based structure allows predictions to be traced back to specific decision paths, enabling investigators and auditors to understand exactly why a claim was classified in a certain way.

From a fraud detection perspective, the model's higher recall rate is particularly valuable, as failing to detect fraudulent claims can lead to direct monetary losses. Although the moderate precision means that some legitimate claims may be flagged, this is considered an acceptable trade-off when the goal is to minimize undetected fraud.



International Journal of Innovative Research in Computer and Communication Engineering (IJIRCCCE)

(A Monthly, Peer Reviewed, Refereed, Scholarly Indexed, Open Access Journal)

In terms of operational benefits, the Decision Tree is lightweight and computationally efficient, producing rapid predictions that are ideal for real-time applications. It also operates without the need for extensive pre-processing, such as feature scaling or normalization, simplifying backend integration.

When integrated into the deployed web application, the model enables claim assessors to input case details and receive instant predictions on the likelihood of fraud. This real-time feedback helps prioritize high-risk claims, reduces manual investigation time, and lowers operational costs. As more claim data becomes available, the system can be retrained periodically to adapt to evolving fraud tactics, ensuring it remains effective and up to date.

By combining speed, transparency, and dependable performance, the Decision Tree provides a well-rounded and practical solution for incorporating machine learning into the insurance claims review process.

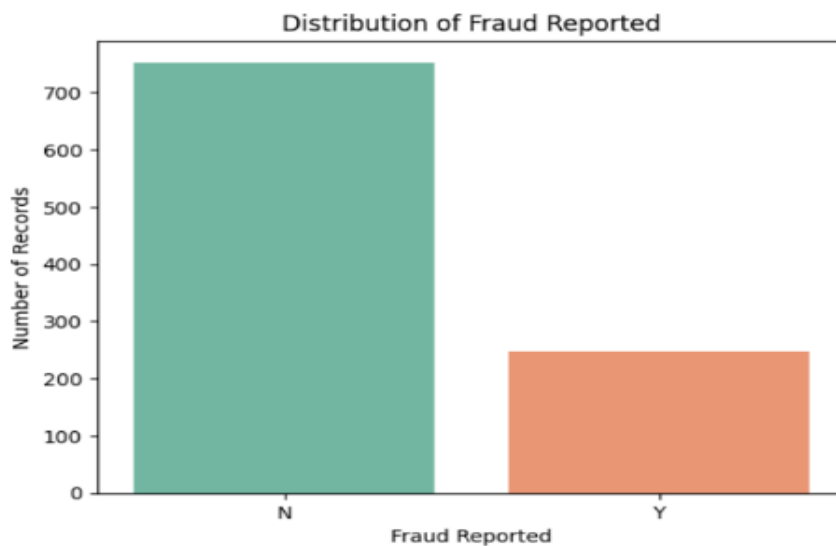


Fig 2: Distribution of fraud-reported claims in the dataset

Figure 2 illustrates the distribution of the target variable, showing a clear imbalance between non-fraudulent claims (“N”) and fraudulent claims (“Y”). The dataset contains far more genuine claims than fraudulent ones, which can cause classification models to lean toward predicting the majority class. This skewed distribution makes it harder for models to detect the minority class—in this case, fraudulent claims—accurately. Because of this imbalance, it becomes essential to focus on evaluation metrics such as precision, recall, and F1-score, rather than relying solely on accuracy. Additionally, techniques like resampling or applying class weights can be used to help the model perform better in identifying rare fraud cases.



International Journal of Innovative Research in Computer and Communication Engineering (IJIRCCE)

(A Monthly, Peer Reviewed, Refereed, Scholarly Indexed, Open Access Journal)

(XAI) tools, like SHAP values, to ensure that the decision-making process remains transparent and interpretable for auditors and investigators. Addressing the imbalance between fraudulent and legitimate claims could be achieved through techniques such as SMOTE or cost-sensitive learning, which would help the model better detect rare fraud cases. Additional improvements could involve incorporating natural language processing (NLP) to analyse textual information in claim reports, enabling the system to detect hidden patterns that may not be evident in structured data. Connecting the platform directly to live insurance databases would allow for real-time claim evaluation, while automated retraining schedules would keep the model up to date with emerging fraud trends.

From a usability standpoint, implementing secure authentication, interactive analytics dashboards, and multilingual support would make the system more user-friendly and adaptable to different regions and operational environments. These upgrades would not only enhance system accuracy and efficiency but also improve its adoption and impact in real-world insurance operations.

REFERENCES

- [1] “Detecting insurance claims fraud using machine learning techniques” published by Riya Roy and Thomas George K in 2017
- [2] “Fraud Detection and Analysis for Insurance Claim using Machine Learning” published by Abhijeet Urunkar, Rashmi Bhat and Nandinee Mudgol in 2022
- [3] “Enhancing Auto Insurance Fraud Detection Using Convolutional Neural Networks” published by Ratchanon Wongpanti and Sirion Vittayakorn in 2024
- [4] “The Identification of Insurance Fraud – an Empirical Analysis Working papers on Risk Management and Insurance” published by Kajia muller in 2013
- [5] “A Model for the Detection of Insurance Fraud, Geneva Papers on Risk and Insurance Theory” published by Belhadji, E., G. Dionne, and F. Tarkhani in 2012
- [6] “A Survey on Fraud Analytics Using Predictive Model in Insurance Claims” published by K. Ulaga Priya and S. Pushpa in 2017
- [7] “A Model for the Detection of Insurance Fraud” published by E. B. Belhadji, G. Dionne, and F. Tarkhani
- [8] “Performance comparative study of machine learning algorithms for automobile insurance fraud detection” published by B. Itri, Y. Mohamed, Q. Mohammed, and B. Omar in 2019
- [9] “Detecting Auto Insurance Fraud by Data Mining Techniques” published by Bhowmik R in 2019
- [10] Kaggle, “Insurance Claim Fraud Detection Dataset,” Kaggle.com, [Online]. Available: <https://www.kaggle.com/> [Accessed: 02-Aug-2025].



INTERNATIONAL
STANDARD
SERIAL
NUMBER
INDIA



INTERNATIONAL JOURNAL OF INNOVATIVE RESEARCH

IN COMPUTER & COMMUNICATION ENGINEERING

 9940 572 462  6381 907 438  ijircce@gmail.com



www.ijircce.com

Scan to save the contact details