

Machine Learning-Based Vehicle Insurance Claim Fraud Detection and Risk Assessment

Dr Chintal Kumar Patel

Associate Professor

CSE Geetanjali Institute of Technical Studies

chintal.patel@gits.ac.in

Corresponding Author: * Chintal Kumar Patel (chintal.patel@gits.ac.in)

Received: 5 September, 2026 | Revised: 10 September, 2026 | Accepted: 28 September, 2026

Abstract—Insurance fraud remains a major concern for the global insurance industry, resulting in substantial financial losses each year. Detecting fraudulent vehicle insurance claims is particularly challenging because of high dimensional data, imbalanced class distributions, and subtle behavioral patterns associated with fraud. This paper presents a unified machine learning framework designed to improve the reliability and consistency of fraud detection across vehicle insurance claims. The proposed framework is tested using the VehicleInsuranceClaim Fraud Detection Dataset from Kaggle which has 15420 samples and 33 attributes. Data preprocessing involves handling missing values, removing duplicates, performing One-Hot Encoding, Z-score normalization and class balancing using ADASYN, and splitting the data into 70:30 train-test sets with stratification. Two ensemble models, namely, Random Forest (RF) and Extreme Gradient Boosting (XGBoost), are created to classify fraudulent claims and to estimate risk. Experimental results demonstrate that the proposed RF achieves 97.95% accuracy, 97.98% precision, 97.91% recall, and 97.93% F1-score, while XGBoost achieves 97.96% accuracy, 97.94% precision, 97.89% recall, and 97.95% F1-score. The findings show that both of the suggested ensemble models provide very reliable classification performance and may facilitate effective, data-driven risk assessment and auto insurance fraud detection.

Keywords—Vehicle Insurance Fraud, Machine Learning, Insurance fraud detection, Artificial intelligence, Classification, Hyperparameter optimization.

I. INTRODUCTION

Economic uncertainty, unforeseen losses, and increasing exposure to various risks have created a growing need for reliable financial protection mechanisms. Insurance provides an organized means of managing such risks by compensating eligible losses and decreasing the financial impact of unexpected events. Insurance plays a vital role in providing financial protection against uncertain events, helping individuals and organizations manage potential economic losses [1]. Among various insurance services, vehicle insurance has become increasingly important due to the continuous growth in vehicle ownership, road accidents, and associated financial risks [2]. Vehicle insurance provides coverage for damages, liabilities, accidents, natural disasters,

and other unexpected events, thereby supporting financial stability for policyholders and insurance providers [3][4].

Vehicle insurance fraud occurs when policyholders or claimants intentionally provide false information, exaggerate damages, stage accidents, or submit illegitimate claims to obtain undue financial benefits [5]. Such fraudulent activities cause substantial financial losses to insurance companies and can contribute to increased premiums and delays for genuine policyholders [6][7]. Conventional fraud detection approaches primarily depend on manual investigations, document verification, physical inspections, and rule-based systems, which are time-consuming, labor-intensive, and often inadequate for detecting complex fraud patterns [8].

Effective risk assessment is essential for distinguishing legitimate claims from potentially fraudulent ones and prioritizing suspicious cases for further investigation [9][10]. Vehicle insurance claims contain diverse information, including policyholder details, vehicle characteristics, claim amounts, accident information, and previous claim history. Analyzing these qualities may assist uncover abnormal patterns and risk factors linked with fraudulent claims, while eliminating the constraints of manual analysis [11][12].

The capacity of ML and DL to understand complicated correlations from massive amounts of data has made them powerful tools for automated risk assessment and fraud detection in the last several years. ML algorithms can classify insurance claims, detect anomalous behavior, and identify important fraud indicators, while DL models can capture complex patterns from high-dimensional data [13][14][15]. Therefore, this study focuses on developing a Machine Learning-Based VehicleInsuranceClaim Fraud Detection and Risk Assessment framework to improve automated fraud identification and support efficient insurance claim evaluation.

A. Motivation and Contribution

This research is driven by the increasing number of fraudulent automobile insurance claims and the significant financial losses they inflict on insurance companies. Traditional manual and rule-based fraud detection methods are often time-consuming, costly, and unable to identify complex fraud patterns. Therefore, there is a need for an

automated and reliable approach that can analyze claim-related attributes and accurately distinguish fraudulent from legitimate claims. This study aims to employ machine learning techniques with effective preprocessing and class-balancing methods to improve fraud detection and support efficient vehicle insurance risk assessment. This research offers several key contributions as listed below:

- Developed a ML framework for VehicleInsuranceClaim Fraud Detection and risk assessment.
- Utilized the VehicleInsuranceClaim Fraud Detection Dataset to analyze legitimate and fraudulent insurance claims.
- Applied data preprocessing techniques, including missing-value handling, duplicate removal, One-Hot Encoding, and Z-score normalization.
- Used ADASYN to address Class Imbalance and improve the representation of fraudulent claims.
- Implemented and evaluated Random Forest (RF) and XGBoost models for fraud classification.
- Compared the proposed models with conventional ML approaches such as NB, DT, and SVM.
- Evaluated model performance using accuracy, precision, recall, F1score, Confusion Matrix, and ROC-AUC.
- Provided a data-driven framework that can support efficient claim screening, fraud identification, and insurance risk assessment.

B. Justification and Novelty

The justification for this study lies in the need for an efficient, automated approach to overcome the limitations of conventional vehicle insurance fraud detection methods. The proposed framework systematically preprocesses claim data using missing-value handling, duplicate removal, One-Hot Encoding, and Z-score normalization. ADASYN is incorporated to address Class Imbalance and improve the representation of fraudulent claims during model training. The novelty lies in the combined application of RF and XGBoost for vehicle insurance fraud detection and risk assessment. The framework provides a comparative evaluation of these ensemble models using multiple classification metrics. This integrated approach supports accurate, consistent, and data-driven identification of potentially fraudulent insurance claims.

C. Organization of the Paper

There are five sections to the remainder of this paper: In Section II, review previous studies on Vehicle Insurance Fraud Detection. Section III discusses the research method employed in this study. The outcomes obtained and the model's effectiveness relative to other methods are presented in Section IV. Finally, in Section V, provide practical conclusions and implications of the findings.

II. LITERATURE REVIEW

A systematic examination of recent studies on Vehicle Insurance Fraud Detection and Risk Prediction was carried out to identify existing approaches, research trends, and potential gaps. Table I summarizes the selected studies by presenting their models, datasets, key findings, and identified limitations.

Sree et al. (2026) proposes a hybrid Evidential Reasoning (ER) and machine learning framework for automobile insurance Fraud Detection using 718 claim records with 49 expert-defined fraud indicators. SMOTE and cost-sensitive learning address class imbalance, while Accuracy, Recall, F1-Score, and AUC evaluate performance. The ER model outperforms LR, DT, RF, ANN, and Naïve Bayes, achieving Recall = 0.82, F1-Score = 0.74, and AUC = 0.91 [16].

Kalaivani et al. (2026) proposes a first hybrid framework combining Machine Learning (ML) and blockchain technologies for fraud detection and enhanced security. The novel FX-Block model combines XGBoost and blockchain, using a public automobile insurance dataset with SMOTE for preprocessing and data balancing. The model achieves 97% accuracy and a ROC-AUC score of 0.9965. It is deployed using Fast API for real-time fraud prediction, while claims are processed using smart contracts in a blockchain environment, providing a traceable and robust insurance fraud detection system [17].

Satone and Daronde (2025) examines many classification techniques for insurance fraud detection, including SVM, RF, DT, Linear Regression (LR), and Polynomial Regression. Precision, recall, and F1score are some of the most important criteria used to assess these algorithms' performance. The analysis reveals that the best-performing model is the Random Forest, with a Mean Squared Error (MSE) of 6.50, outperforming the other techniques [18].

Sargam and Kalapala (2025) a federated anomaly detection mechanism combined with cloud-based orchestration on Amazon Web Services (AWS) for safe and collaborative fraud detection among healthcare organizations. It uses unsupervised deep auto encoders and Federated Averaging (FedAvg) for decentralized training without sharing raw data, with differential privacy controls. The system uses AWS EC2 for computation, Sage Maker for model hosting, and Amazon S3 for secure data storage. Using real insurance claim data and synthetic fraud cases, it achieved 91.2% detection accuracy, 0.885 F1-score, and a 43% reduction in FPR, while decreasing data exposure risk by 78% [19].

Bolukbasi et al. (2025) insurance fraud using 10 ML models, namely RF, SVM, GB, K-nearest Neighbors, Gaussian Naïve Bayes, Extreme Gradient Boosting, Voting Classifier, AdaBoost, Decision Tree, and Logistic Regression. The Voting Classifier achieved the best performance with 87% F1-Score and classification accuracy. Explainable Artificial Intelligence techniques, LIME and SHAP, were applied to explain model decisions, identifying incident severity as the most significant feature for detecting insurance fraud [20].

Y et al. (2023) a novel hybrid approach for automated VehicleInsuranceClaim Fraud Detection by combining Federated Learning (FL), Genetic Algorithm (GA), and PSO. GA is used for optimal feature subset extraction, which is then fed into a Federated Learning with ParticleSwarmOptimization (FPSO) model. The proposed hybrid model achieves an accuracy of 94.47% and may be further improved using other nature-inspired algorithms for fraud detection [21].

TABLE I. RECENT TRENDS IN VEHICLE INSURANCE FRAUD DETECTION USING MACHINE LEARNING TECHNIQUES

Author	Proposed Work	Dataset	Key Findings	Limitations & Future Work
Sree et al. (2026)	Hybrid Evidential Reasoning (ER) + ML framework for automobile insurance fraud detection with fraud-risk scoring.	718 automobile insurance claims with 49 expert-defined fraud indicators.	ER achieved Recall 82%, F1-score 74%, and AUC 91%; SMOTE and cost-sensitive learning addressed imbalance.	Relatively small dataset; future work can investigate larger datasets and more advanced ML models.
Kalaivani et al. (2026)	FX-Block, combining XGBoost with blockchain for automobile insurance fraud detection and secure claim processing.	Public automobile insurance dataset.	Achieved 97% accuracy and 99.65% ROC-AUC; FastAPI and blockchain were used for real-time and traceable processing.	Blockchain integration may increase system complexity; broader real-world validation is required.
Satone & Daronde (2025)	Comparative evaluation of SVM, RF, DT, LR, and Polynomial Regression for insurance fraud detection.	Insurance fraud dataset.	Random Forest reported the strongest performance among the evaluated models, with reported MSE of 6.50.	Limited model integration and scope; future research can explore ensemble and advanced optimization approaches.
Sargam & Kalapala (2025)	Federated anomaly detection using deep autoencoders, FedAvg, differential privacy, and AWS for insurance fraud detection.	Real insurance claims combined with synthetic fraud cases.	Achieved 91.2% accuracy and 0.885 F1-score, with reduced false-positive rates through federated learning.	Focuses on healthcare insurance rather than vehicle insurance; future work can extend the approach to automobile claims.
Bolukbasi et al. (2025)	Comparative study of 10 ML models with LIME and SHAP for explainable insurance fraud detection.	Insurance fraud dataset.	Voting Classifier achieved 87% accuracy and F1-score; incident severity was identified as an important feature.	Performance can be improved using larger datasets, advanced ensembles, and domain-specific vehicle features.
Y et al. (2023)	Hybrid Federated Learning + Genetic Algorithm + PSO framework for vehicle insurance claim fraud detection.	Vehicle insurance claim data.	Reported 94.47% accuracy using optimized feature selection and federated learning.	Further optimization using additional nature-inspired algorithms and larger datasets was suggested.

Research Gaps: Existing studies have explored ML, DL, federated learning, blockchain, and hybrid approaches for insurance fraud detection, but several limitations remain. Many approaches rely on small or specialized datasets, while limited attention is given to robust preprocessing and severe class imbalance in vehicle insurance claims. There is still a lack of research on the efficacy of combining ADASYN with ensemble models like Random Forest and XGBoost for risk assessment and auto insurance fraud. Therefore, this study addresses these gaps by developing an integrated and data-driven framework for reliable Vehicle Insurance Fraud Detection and risk assessment.

III. RESEARCH METHODOLOGY

The proposed methodology includes data from the VehicleInsuranceClaim Fraud Detection Dataset from Kaggle, which comprises 15,420 samples with 33 features, pre-processing, categorical encoding, z-score normalization, and class balancing using ADASYN. The balanced data were divided using stratified 70:30 sampling, and RF and XGBoost models were developed for vehicle insurance fraud detection and risk assessment. For measuring an effectiveness of classification, accuracy, precision, recall, F1score, Confusion Matrix, and ROC-AUC were used with regard to model performance. Figure 1. presents the proposed workflow for VehicleInsurance Fraud Detection and Risk Prediction using ML.

The following section gives a full description of each stage in the suggested methodology:

A. Data Gathering and Analysis

This study utilized the VehicleInsuranceClaim Fraud Detection Dataset from Kaggle, consisting of 15,420 records and 33 attributes, to analyze fraudulent insurance claims and assess associated risk factors. The dataset includes relevant claimant, policy, vehicle, and claim-related attributes for developing fraud detection and risk assessment models. Data visualizations such as bar plots and heatmaps were used to examine claim distribution, feature correlations, and other relevant patterns, as presented below:

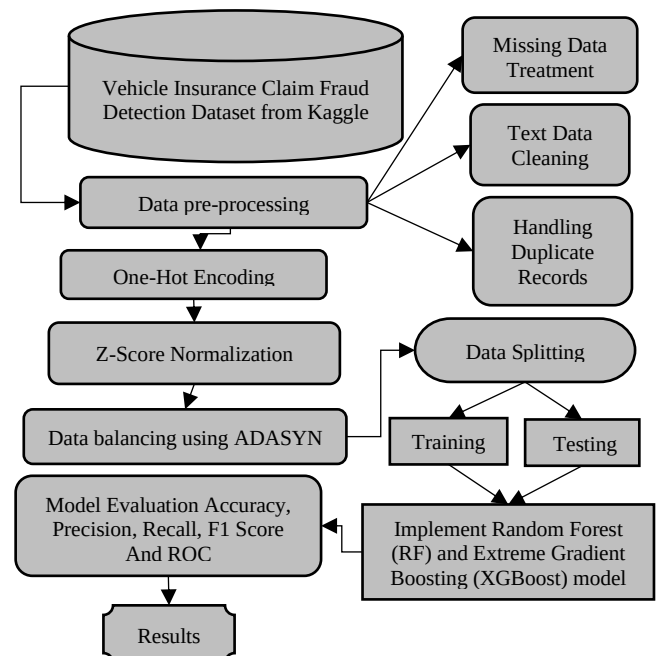


Fig. 1. Proposed Machine Learning Framework for Vehicle Insurance Fraud Detection And Risk Prediction

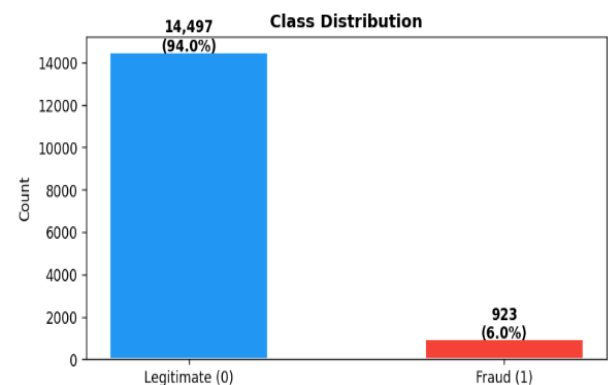


Fig. 2. Bar Graph of Class Distribution

Figure 2. illustrates that the dataset contains 14,497 legitimate claims (94.0%) and 923 fraudulent claims (6.0%). The distribution is heavily skewed and is heavily weighted in favor of legitimate claims over fraudulent claims. This imbalance should be considered during model training and evaluation for effective fraud detection.

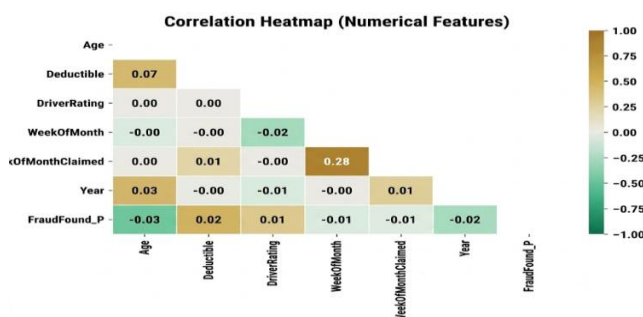


Fig. 3. Correlation Matrix Heatmap

Figure 3. shows that most numerical combination pairs of features are not correlated, with the majority having values close to zero, implying independent feature behaviors. There is a mild linear connection between WeekOfMonth and WeekOfMonthClaimed with a coefficient of 0.28, the strongest positive connection. In addition, other features have very low correlation with Freedsound, with the correlation remaining within the range of -0.03 to 0.02.

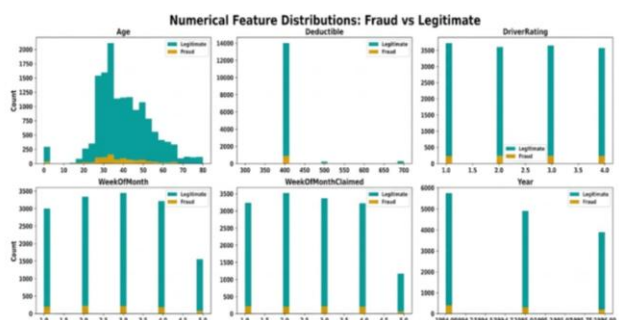


Fig. 4. Numerical Feature Distributions: Fraud vs Legitimate

The class distributions for the numerical variables are shown in Figure 4. and there is a pronounced class imbalance for each feature, with legitimate transactions much more frequent than fraudulent ones. The distributions of features like Age are right-skewed, with a normal shape around the median of 30 to 40 years, whereas those of discrete features such as Deductible, Driver Rating, WeekOfMonth, WeekOfMonthClaimed, and Year are concentrated at specific categorical values. Fraudulent cases are evenly distributed across these ranges, even though legitimate claims generally dominate in volume.

B. Data Pre-Processing

The VehicleInsuranceClaim Fraud Detection Dataset was pre-processed by handling missing values, removing duplicates, and correcting inconsistent records. Categorical features were encoded using One-Hot Encoding, while numerical features were standardized for effective model training. The main preprocessing procedures are summarized as follows:

- **Missing Data Treatment:** Missing values in structured insurance variables were identified and

appropriately handled to maintain data consistency. Claims with insufficient information were excluded where reliable completion was not possible.

- **Text Data Cleaning:** Claim descriptions and damage-related text were cleaned by removing unnecessary characters, noise, and irrelevant information. The cleaned text was normalized and converted into numerical word-vector representations for model development.
- **Handling Duplicate Records:** Duplicate claim records were identified and removed to avoid redundancy in the dataset. This ensured that each insurance claim contributed only once to the analysis and model training.

C. Encoding Categorical Variables Using One-Hot Encoding

Categorical variables such as accident type and damage category were transformed into numerical representations using One-Hot Encoding. The target variable was encoded in binary form, where 0 represents “No Fraud Found” and 1 represents “Fraud Found.”

D. Z-Score Normalization

Data Normalization is a technique used to transform or standardize the data to have a similar distribution. The most widely used method for data normalization is rescaling or min-max normalization and z-score normalization. In this study, have applied z-score normalization, a standardization technique with a mean of 0 and having a standard deviation of 1. This scaling technique transforms the values centered around the average value with unit standard deviation. The z-score normalization is defined as given in Equation (1).

$$E' = \frac{E - \bar{M}}{\sigma_M} \quad (1)$$

Where:

E' and E are new and old for each data entry, $\bar{M}$ is the mean, and σ_M is the standard deviation.

E. Data Balancing Using ADASYN

Data balancing is a preprocessing technique used to reduce class imbalance by adjusting the distribution of samples among different classes. It helps prevent the model from being biased toward the Majority Class and improves its ability to identify minority-class instances. In this study, ADASYN was used to generate synthetic minority samples based on their classification difficulty. More synthetic samples are created for challenging or frequently misclassified instances, allowing the model to learn complex minority-class patterns and achieve more balanced predictive performance.

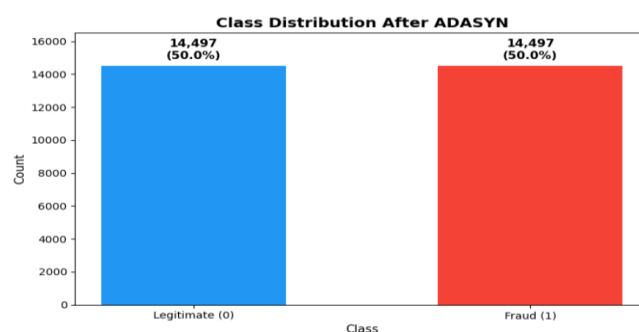


Fig. 5. Bar Graph of Class Distribution After ADASYN

Figure 5. illustrates that the dataset exhibits a balanced distribution between the two categories. Both Legitimate (0) and Fraud (1) contain 14,497 samples (50.0%), indicating an equal class distribution after ADASYN. This balancing helps provide a more uniform representation of legitimate and fraudulent claims for model training.

F. Data Splitting

The dataset was split into 70% training and 30% testing sets using stratified sampling. This approach maintained the original proportion of fraudulent and non-fraudulent claims across both subsets, ensuring a consistent class distribution during model evaluation.

G. Implement of Proposed Model

The proposed RF and XGBoost models were implemented for VehicleInsuranceClaim Fraud Detection and risk assessment.

1) Proposed Random Forest (RF) Model

RF is an ML ensemble approach that uses a combination of decision trees to enhance classification accuracy while decreasing the likelihood of overfitting. The model generates several independent decision trees using randomly selected subsets of training samples and features. Each tree provides an individual prediction, and the final classification outcome is determined through majority voting among all generated trees. The RF prediction function can be represented as Equation (2).

$$\hat{Y} = \text{mode}\{T_1(X), T_2(X), T_3(X), \dots, T_N(X)\} \quad (2)$$

where $\hat{Y}$ represents the final predicted class, $T_i(X)$ denotes the prediction of the i^{th} decision tree, X represents the input feature vector, and n indicates the total number of trees in the forest. The Ensemble Learning mechanism of RF enhances model stability, improves generalization capability, and provides effective classification performance for Vehicle Insurance Fraud Detection.

The probability estimation in the RFC is calculated by aggregating the outputs of all individual trees, which can be expressed as Equation (3):

$$P(Y = c|X) = \frac{1}{N} \sum_{i=1}^N I(T_i(X) = c) \quad (3)$$

where $P(Y = c|X)$ represents the probability of class c , N is the number of decision trees, and $I(\cdot)$ is an indicator function that returns 1 when the condition is satisfied. This probability-based decision mechanism enables RF to accurately distinguish between Vehicle Insurance Fraud Detection. The proposed RF model was configured with estimators = 200, criterion = Gini, max_depth = 20, min_samples_split = 2, min_samples_leaf = 1, max_features = sqrt, bootstrap = True, and random state = 42.

2) Proposed Extreme Gradient Boosting (XGBoost) Model

XGBoost is an Ensemble Learning algorithm that uses decision trees to make predictions. It can be used for regression problems by minimizing a loss function that measures a difference among the forecasted and actual target values. The mathematical model for XGBoost regression can be expressed as Equation (4):

$$y = f(x) \quad (4)$$

where y is the predicted property price, x is the vector of input features (such as square footage, number of bedrooms, etc.), and $f(x)$ is the XGBoost model that predicts y based on x . To compute $f(x)$, XGBoost builds an ensemble of decision trees that are trained to minimize the MeanSquaredError (MSE) loss function. The model combines the predictions from multiple DT to arrive at a final prediction. The general form of the XGBoost regression model can be expressed as Equation (5):

$$y = \sum_{k=1}^K f_k(x) \quad (5)$$

where $f_k(x)$ is the prediction of the k^{th} decision tree, and K is the total number of decision trees in the ensemble. The training process teaches each tree its leaf values, and the prediction is just a weighted average of those values. The prediction of the XGBoost model for a given input x is obtained by summing the predictions of all the decision trees in the ensemble. The proposed XGBoost model was configured with n_estimators = 200, max_depth = 6, learning rate = 0.1, subsample = 0.8, colsample_bytree = 0.8, min_child_weight = 1, gamma = 0, reg_alpha = 0, reg_lambda = 1, and random state = 42.

H. Evaluation Metrics

The effectiveness of the proposed design was assessed using multiple performance metrics. A Confusion Matrix was first constructed to summarize the classification results, showing the counts of correct and incorrect predictions across classes. From the confusion matrix, the following values were derived: True Positives (TP), False Positives (FP), True Negatives (TN), and False Negatives (FN). Based on these values, essential performance metrics including accuracy, precision, recall, and F1-score were calculated in Equation (6) to (9) as detailed below:

$$\text{Accuracy} = \frac{TP+TN}{TP+FP+TN+FN} \quad (6)$$

$$\text{Precision} = \frac{TP}{TP+FP} \quad (7)$$

$$\text{Recall} = \frac{TP}{TP+FN} \quad (8)$$

$$F1 - \text{score} = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \quad (9)$$

The precision reflects the percentage of properly predicted positive cases out of all predicted positive cases, whereas the accuracy indicates the proportion of correctly categorized instances out of all occurrences. Recall measures the proportion of actual positive cases correctly identified by the model, and the F1score provides the harmonic mean of Perc and Rec, with values ranging from 0 to 1. The Recurve illustrates the relationship among the TPR and FPR across different classification thresholds, where TPR corresponds to sensitivity or Recall, and FPR equals 1-specificity.

IV. RESULTS AND DISCUSSION

This section describes the experimental setup and assesses the effectiveness of the proposed RF and XGBoost models for assessing automobile insurance fraud risk. An experiments were conducted on a high-performance system equipped with an Intel Core i9-13900K CPU (3.0 GHz), 64 GB DDR5 RAM, and NVIDIA RTX 4090 GPU (24 GB VRAM) running Windows 11 Pro, using Python libraries including Pandas, NumPy, Matplotlib, Seaborn, and Scikit-learn. The VehicleInsuranceClaim Fraud Detection Dataset was utilized

to develop the model, and performance was assessed using Accuracy, Precision, Recall, and F1score. As presented in Table II. the proposed RFmodel achieved 97.95% Accuracy, 97.98% Precision, 97.91% Recall, and 97.93% F1-score, whereas the proposed XGBoost model obtained 97.96% Accuracy, 97.94% Precision, 97.89% Recall, and 97.95% F1-score. These results demonstrate the consistent classification capability of both proposed models across the evaluated performance measures.

TABLE II. CLASSIFICATION RESULTS OF THE PROPOSED MODELS FOR VEHICLE INSURANCE FRAUD RISK ANALYSIS

Matrix	Proposed Random Forest (RF)	Proposed Extreme Gradient Boosting (XGBoost)
Accuracy	97.95	97.96
Precision	97.98	97.94
Recall	97.91	97.89
F1-score	97.93	97.95

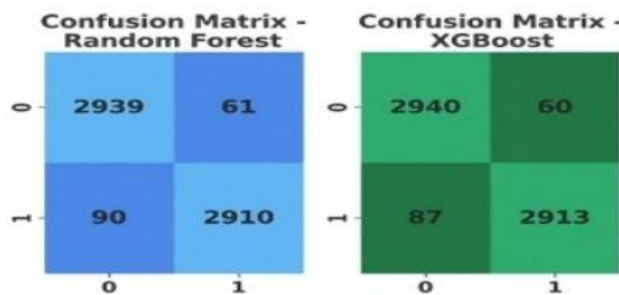


Fig. 6. Confusion Matrix for the Proposed Models

Figure 6. illustrates the confusion matrices comparing the performance of the Random Forest and XGBoost classification models. For class 0, Random Forest correctly predicts 2,939 true negatives with 61 false positives, whereas XGBoost achieves slightly better accuracy with 2,940 true negatives and 60 false positives. For class 1, Random Forest identifies 2,910 true positives alongside 90 false negatives, while XGBoost demonstrates superior sensitivity by recording 2,913 true positives and reducing false negatives to 87. Overall, XGBoost outperforms Random Forest across both categories, displaying a higher rate of correct classifications and lower overall error counts.

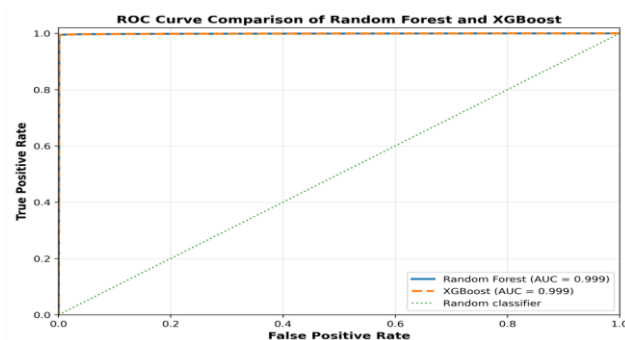


Fig. 7. ROC Curve for Proposed Models

Figure 7. ROC curve compares a performance of the RF and XGBoost classifiers against a baseline random classifier. The RF and XGBoost models both have an AUC of 0.999, indicating almost perfect predictive power. The solid blue and dashed orange lines closely track the top-left corner, indicating high true positive rates and extremely low false positive rates. The high values for the margin above the diagonal green line indicate that the datasets of both proposed

models have high discriminative power towards the classification task.

A. Comparative Analysis

To determine an effectiveness of a proposed models, their classification performance was compared with existing Machine Learning (ML) approaches, as presented in Table III. The Naïve Bayes (NB) model achieved 68.29% accuracy, 83.94% precision, 91.44% recall, and 58.59% F1-score, while the Decision Tree (DT) obtained 74.90% accuracy, 58.20% recall, and 23.06% F1-score. The SVM recorded 85.60% accuracy, 76.80% precision, 74.10% recall, and 75.40% F1-score. In comparison, the proposed Random Forest (RF) achieved 97.95% accuracy, 97.98% precision, 97.91% recall, and 97.93% F1-score, whereas Extreme Gradient Boosting (XGBoost) attained 97.96% accuracy, 97.94% precision, 97.89% recall, and 97.95% F1-score. These results demonstrate the strong classification performance of the proposed approaches across all evaluated metrics.

TABLE III. COMPARATIVE ANALYSIS OF MACHINE LEARNING MODELS FOR VEHICLE INSURANCE FRAUD DETECTION

Model	Accuracy	Precision	Recall	F1-score
NB[22]	68.29	83.94	91.44	58.59
DT[23]	74.9		58.2	23.06
SVM[24]	85.6	76.8	74.1	75.4
Proposed RF	97.95	97.98	97.91	97.93
Proposed XGBoost	97.96	97.94	97.89	97.95

The proposed Random Forest (RF) and XGBoost models achieved high classification accuracy of 97.95% and 97.96%, respectively, demonstrating effective vehicle insurance fraud detection. Both models can capture complex relationships among claim-related features and identify fraudulent patterns efficiently. Their strong predictive performance supports reliable classification of legitimate and fraudulent insurance claims. The outcomes show that the proposed models can provide an effective machine learning-based solution for automated fraud detection and risk assessment.

V. CONCLUSION AND FUTURE STUDY

Fraudulent insurance claims place a tremendous financial strain on the automobile insurance sector, resulting in higher rates for honest consumers and large losses for insurers. Conventional manual techniques for identifying fraudulent claims are laborious, ineffective, and prone to errors. This study addresses these challenges by applying and evaluating multiple machine learning techniques to accurately distinguish between genuine and fraudulent vehicle insurance claims. The research specifically aims to develop an effective fraud detection framework using the VehicleInsuranceClaim Fraud Detection Dataset, with appropriate preprocessing, Z-score normalization, and ADASYN-based data balancing. The experimental outcomes show that a proposed RF and XGBoost models achieved accuracies of 97.95% and 97.96%, respectively, outperforming NB (68.29%), DT (74.90%), and SVM (85.60%). The results show that ensemble-based machine learning models can accurately detect insurance claim fraud and provide accurate risk assessment. The suggested framework offers a means for insurers to enhance fraud detection efficacy, minimize monetary damages, and bolster data-driven claims evaluation.

The limitation of this study is its restriction to a single vehicle insurance claim dataset, and another is the structured

nature of the attributes of the claims used, which could affect generalizability to other insurance fields and real-world datasets. Future work could focus on larger multi-source datasets, explainability, deep learning methods, real-time fraud detection and enhanced optimization techniques that would further improve robustness and practical implementation.

- **Funding:** This research received no external funding.
- **Institutional Review Board Statement:** Not applicable.
- **Informed Consent Statement:** Not applicable.
- **Data Availability Statement:** The data supporting the findings of this study are available from the corresponding author upon reasonable request.
- **Conflicts of Interest:** The author declares no conflict of interest.

How to cite this article:

C. K. Patel, "Machine Learning-Based Vehicle Insurance Claim Fraud Detection and Risk Assessment," *Journal of Global Research in Multidisciplinary Studies*, vol. 2, no. 9, pp. 8–14, Sep. 2026. Doi.XXX

REFERENCES

- [1] A. Al Doulat, O. E. Ayo-Bali, and S. Shaik, "Fraud Detection in Insurance Claims Using Supervised Machine Learning Models," in *2025 International Conference on Smart Applications, Communications and Networking (SmartNets)*, IEEE, Jul. 2025, pp. 1–7. doi: 10.1109/SmartNets65254.2025.11106803.
- [2] K. M. Hussain and S. C. Pallaprolu, "Improving Passenger Safety with Machine Learning Techniques on 2022 Incident Data," in *2025 International Conference on Sensors and Related Networks (SENNET) Special Focus on Digital Healthcare (64220)*, Vellore, India: IEEE, 2025, pp. 1–6, July. doi: 10.1109/SENNET64220.2025.11136102.
- [3] R. K. Ramadhan, L. M. Putranto, R. Inawan, and M. Yasirroni, "Data Driven Frequency Stability Detection in Power Systems Using XGBoost Model," in *2024 Beyond Technology Summit on Informatics International Conference (BTS-I2C)*, IEEE, Dec. 2024, pp. 7–12. doi: 10.1109/BTS-I2C63534.2024.10942308.
- [4] N. Ding, X. Ruan, H. Wang, and Y. Liu, "Automobile Insurance Fraud Detection Based on PSO-XGBoost Model and Interpretable Machine Learning Method," *Insur. Math. Econ.*, vol. 120, pp. 51–60, Jan. 2025, doi: 10.1016/j.insmatheco.2024.11.006.
- [5] N. S. Patil, S. Kamanavalli, S. Hiregoudar, S. Jadhav, S. Kanakraddi, and N. D. Hiremath, "Vehicle Insurance Fraud Detection System Using Robotic Process Automation and Machine Learning," in *2021 International Conference on Intelligent Technologies (CONIT)*, IEEE, Jun. 2021, pp. 1–5. doi: 10.1109/CONIT51480.2021.9498507.
- [6] E. Nabrawi and A. Alanazi, "Fraud Detection in Healthcare Insurance Claims Using Machine Learning," *Risks*, vol. 11, no. 9, p. 160, Sep. 2023, doi: 10.3390/risks11090160.
- [7] H. Lv, X. Liu, S. Lin, X. Ruan, and N. Ding, "Auto insurance fraud detection based on Logistic-SVM algorithm," in *International Conference on Electronic Information Engineering and Computer Communication (EIECC 2021)*, Z. Zhu and F. Cen, Eds., SPIE, May 2022, p. 58. doi: 10.1117/12.2634651.
- [8] C. Muranda, A. Ali, and T. Shongwe, "Detecting Fraudulent Motor Insurance Claims Using Support Vector Machines with Adaptive Synthetic Sampling Method," in *2020 61st International Scientific Conference on Information Technology and Management Science of Riga Technical University (ITMS)*, IEEE, Oct. 2020, pp. 1–5. doi: 10.1109/ITMS51158.2020.9259322.
- [9] 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)*, IEEE, Sep. 2018, pp. 528–531. doi: 10.1109/ICDSBA.2018.00104.
- [10] V. P. Rambabu, J. Sreerama, and J. T. S. Singh, "AI-Driven Data Integration: Enhancing Risk Assessment in the Insurance Industry," *Aust. J. Mach. Learn. Res. Appl.*, vol. 2, no. 2, pp. 130–179, July, 2022.
- [11] 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)*, IEEE, Dec. 2015, pp. 1–7. doi: 10.1109/ICCIC.2015.7435726.
- [12] S. Agarwal, "An Intelligent Machine Learning Approach for Fraud Detection in Medical Claim Insurance: A Comprehensive Study," *Sch. J. Eng. Technol.*, vol. 11, no. 09, pp. 191–200, Sep. 2023, doi: 10.36347/sjet.2023.v11i09.003.
- [13] M. R. HaratiNik, M. Akrami, S. Khadivi, and M. Shajari, "FUZZGY: A hybrid model for credit card fraud detection," in *6th International Symposium on Telecommunications (IST)*, IEEE, Nov. 2012, pp. 1088–1093. doi: 10.1109/ISTEL.2012.6483148.
- [14] T. Badriyah, L. Rahmania, and I. Syarif, "Nearest Neighbour and Statistics Method based for Detecting Fraud in Auto Insurance," in *2018 International Conference on Applied Engineering (ICAE)*, IEEE, Oct. 2018, pp. 1–5. doi: 10.1109/INCAE.2018.8579155.
- [15] S. Padhi and S. Panigrahi, "Decision Templates based Ensemble Classifiers for Automobile Insurance Fraud Detection," in *2019 Global Conference for Advancement in Technology (GCAT)*, IEEE, Oct. 2019, pp. 1–5. doi: 10.1109/GCAT47503.2019.8978332.
- [16] S. S. B. Sree, S. S. S. M. R. V. R. Sekaran, and J. S., "Auto insurance fraud detection through uncertain reasoning and data-based inferential modelling," in *2026 International Conference on Data Science and Business Systems (ICDSBS)*, IEEE, May 2026, pp. 1–6. doi: 10.1109/ICDSBS69077.2026.11634831.
- [17] A. Kalaivani, V. Asha, J. Sathya, A. Prasad, R. Roy, and S. Kehri, "Fraud Detection in Insurance Claims Using XGBoost and Blockchain," in *2026 International Conference on Intelligent and Innovative Technologies in Computing, Electrical and Electronics (IITCEE)*, IEEE, Jan. 2026, pp. 1–5. doi: 10.1109/IITCEE67948.2026.11394356.
- [18] A. P. Satone and S. Daronde, "Assessing the Effectiveness of Machine Learning Algorithms in Vehicle Insurance Fraud Prediction: A Comparative Approach," in *2025 International Conference on Electronics and Renewable Systems (ICEARS)*, IEEE, Feb. 2025, pp. 1999–2004. doi: 10.1109/ICEARS64219.2025.10941222.
- [19] G. S. Sargam and R. Kalapala, "AI-Driven Claim Fraud Detection in Health Insurance Using Federated Anomaly Detection Networks with Cloud Computing on AWS for Privacy-Preserving Financial Security," in *2025 Third International Conference on Cyber Physical Systems, Power Electronics and Electric Vehicles (ICPEEV)*, IEEE, Sep. 2025, pp. 1–6. doi: 10.1109/ICPEEV67897.2025.11291290.
- [20] Z. Bolukbasi, H. Yurtoglu, E. Kugu, and H. Cukurtepe, "Insurance Fraud Detection Using Machine Learning and XAI Algorithms," in *2025 IEEE International Conference on Electro Information Technology (eIT)*, IEEE, May 2025, pp. 463–468. doi: 10.1109/eIT64391.2025.11103653.
- [21] S. Y. N. Victor, G. Srivastava, and T. R. Gadekallu, "A Hybrid Federated Learning Model for Insurance Fraud Detection," in *2023 IEEE International Conference on Communications Workshops (ICC Workshops)*, IEEE, May 2023, pp. 1516–1522. doi: 10.1109/ICCWorkshops57953.2023.10283682.
- [22] R. Ming, O. Abdelrahman, N. Innab, and M. H. K. Ibrahim, "Enhancing fraud detection in auto insurance and credit card transactions: a novel approach integrating CNNs and machine learning algorithms," *PeerJ Comput. Sci.*, vol. 10, pp. 1–35, 2024, doi: 10.7717/peerj-cs.2088.
- [23] Z. Zhang, "Fraud Detection in Vehicle Insurance Claims using Machine Learning," *Ayar*, vol. 15, no. 1, pp. 37–48, 2024.
- [24] H. Patil and S. K. Nanda, "Multimodal AI for Automated Vehicle Insurance Claim Assessment and Fraud Detection," vol. 3, pp. 1709–1721, 2026.