ISSN 2277-2685
IJESR/April. 2026/ Vol-16/Issue-2s1/290-294
Mohammed Zohaib Hussain *et. al.*, / International Journal of Engineering & Science Research
*Full Length Article*

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

**Mohammed Zohaib Hussain[1], Syed Saqlain Ahmed[2], Mohammed Abdul Raqeeb Haqqani[3], Dr . K Upendra Babu[4]**
[1,2,3]B.E.Students; Department of Information Technology, ISL Engineering College, Hyderabad, India
[4]Assistant Professor; Department of Information Technology, ISL Engineering College, Hyderabad, India
Email: zohaibmohammed229@gmail.com,syedsaqlain6251@gmail.com,maraqeebhaqqani@gmail.com
***Accepted 27-04-2026***


***Abstract***
*Insurance fraud has become one of the fastest-growing financial crimes in the global insurance industry. Fraudulent activities negatively impact insurance companies by increasing operational losses, damaging customer trust, and raising insurance premiums for honest policyholders. Traditional fraud detection systems rely mainly on manual investigations and predefined business rules. These systems are no longer sufficient because fraudsters continuously develop more sophisticated methods to bypass verification mechanisms. Machine learning has emerged as an effective solution for identifying fraudulent insurance claims automatically by analyzing hidden patterns within historical data. However, insurance datasets usually suffer from two major problems: class imbalance and missing values. Fraudulent claims represent only a very small portion of the total claims, making prediction difficult for conventional machine learning algorithms. In addition, incomplete customer records and missing claim information reduce the reliability of predictive systems. This paper presents a detailed machine learning-based framework for insurance fraud detection. The proposed system integrates preprocessing techniques, missing value handling, Synthetic Minority Oversampling Technique (SMOTE), feature engineering, and advanced classification algorithms such as Logistic Regression, Decision Tree, Random Forest, and XGBoost. Experimental results demonstrate that ensemble learning algorithms significantly improve fraud detection accuracy, precision, recall, and F1-score compared to traditional approaches.*

***Keywords:*** *Insurance Fraud Detection, Machine Learning, Class Imbalance, SMOTE, Missing Values, Random Forest, XGBoost, Gradient Boosting, Logistic Regression, Anomaly Detection.*

## Introduction

The insurance industry plays a critical role in economic stability by protecting individuals and businesses Against uncertain risks. Despite its importance, insurance fraud has become a major challenge worldwide. Fraudulent claims lead to billions of dollars in losses every year and create financial pressure on insurance companies. As a result, companies increase premiums for genuine customers to compensate for these losses. Insurance fraud can occur in multiple forms. Examples include fake accident claims, staged vehicle collisions, false medical expenses, duplicate claims, and intentional damage to insured property. Some fraud cases involve organized criminal networks that exploit weaknesses in claim verification systems.

Traditional fraud detection methods are mainly based on manual review processes and rule-based systems. Manual verification is slow, labor-intensive, and inefficient when dealing with millions of records. Rule-based systems also fail when fraud patterns evolve over time because fraudsters adapt quickly to predefined conditions. Machine learning techniques provide an automated and scalable solution for fraud detection. These algorithms can learn complex behavioral patterns from historical datasets and identify suspicious claims more accurately than traditional methods. However, real-world insurance datasets contain highly imbalanced class distributions. Fraudulent claims are usually much fewer compared to legitimate claims, causing machine learning models to become biased toward majority classes. Another major issue is missing data. Insurance records often contain incomplete information because of human errors, inconsistent reporting, or system limitations. Missing values reduce prediction quality and negatively impact model stability. Therefore, advanced preprocessing techniques are required

before model training. The primary goal of this research is to design a robust fraud detection framework capable of handling class imbalance and missing values while improving predictive performance using ensemble learning techniques.

**Literature Review**

Researchers have explored several approaches for fraud detection in insurance and financial systems. Early fraud detection systems were primarily based on statistical analysis and expert-defined business rules. Although these methods were simple to implement, they lacked flexibility and adaptability. Logistic Regression became one of the earliest machine learning techniques used for binary classification problems such as fraud detection. It provides interpretable results and works effectively when relationships between variables are linear. However, Logistic Regression struggles with highly nonlinear and complex fraud patterns. Decision Tree algorithms gained popularity because they produce human-readable decision rules. Insurance companies prefer interpretable models because investigators can understand why a claim is classified as fraudulent. However, Decision Trees are prone to overfitting, especially when datasets contain noise and imbalance. Random Forest algorithms improved prediction performance by combining multiple Decision Trees through ensemble learning. This method reduces overfitting and improves generalization capability. Similarly, Gradient Boosting methods such as XGBoost demonstrated superior performance because they iteratively improve weak learners and capture complex data relationships.

Several studies also focused on handling class imbalance problems. One of the most widely use techniques is SMOTE, which generate synthetic minority class samples instead of simply duplicating fraud records. This improves the ability of machine learning models to learn minority fraud patterns effectively.

Recent studies explored deep learning approaches such as Artificial Neural Networks and Autoencoders for fraud detection. Although these methods show promising results, they require large computational resources and massive datasets. Many previous studies also ignored missing value handling, which remains a major issue in real-world datasets. The proposed framework combines preprocessing, missing value handling, balancing techniques, feature engineering, and ensemble learning into a unified architecture for efficient fraud detection.

**Methodology Block Diagram**

The proposed methodology follows a structured pipeline consisting of multiple stages. The process begins with data collection from insurance claim records. The collected dataset undergoes preprocessing to remove duplicates, inconsistencies, and invalid entries. Missing value handling is performed using statistical imputation methods. Numerical features are filled using mean or median values, while categorical attributes are handled using mode replacement techniques. This step improves dataset completeness and consistency. Feature engineering is then applied to identify important variables influencing fraud detection. Examples include claim amount, customer age, accident location, claim frequency, policy duration, and historical customer behavior. After preprocessing, the dataset is balanced using SMOTE. This technique generates synthetic fraud samples by interpolating minority observations with their nearest neighbors.
SMOTE significantly improves minority class learning and reduces model bias toward majority classes. The balanced dataset is divided into training and testing subsets using an 80:20 split ratio. Multiple machine learning models are trained and compared:

Logistic Regression
Decision Tree
Random Forest
XGBoost

Cross-validation techniques are used to evaluate model reliability. Performance metrics such as Accuracy, Precision, Recall, and F1-Score are used to measure fraud detection capability.


**Mathematical Concepts**

The fraud detection process is mathematically represented as a binary classification problem. Let X represent the input feature vector and Y represent the output label. If Y = 1, the claim is fraudulent. If Y = 0, the claim is genuine.

The Logistic Regression probability function is:

$$P(Y=1|X)=1/(1+e^{-z})$$

Where:

$$z = b0 + b1x1 + b2x2 + ... + bnxn$$

$$\text{F1-Score} = 2 \times (\text{Precision} \times \text{Recall})/(\text{Precision} + \text{Recall})$$

Recall is particularly important in fraud detection because missing fraudulent claims can result in major financial losses.

**Implementation Flowchart**

The proposed system is implemented using Python because of its extensive support for data science and machine learning libraries. Several libraries are used during implementation:

Pandas for data manipulation and preprocessing
NumPy for numerical operations
Scikit-learn for machine learning algorithms
Imbalanced-learn for SMOTE implementation
XGBoost for gradient boosting models

Random Forest classification combines multiple Decision Trees and aggregates their outputs through majority voting. XGBoost uses gradient boosting optimization to iteratively minimize classification error. Performance evaluation metrics are defined as:

$$\text{Accuracy} = (TP + TN)/(TP + TN + FP + FN)$$

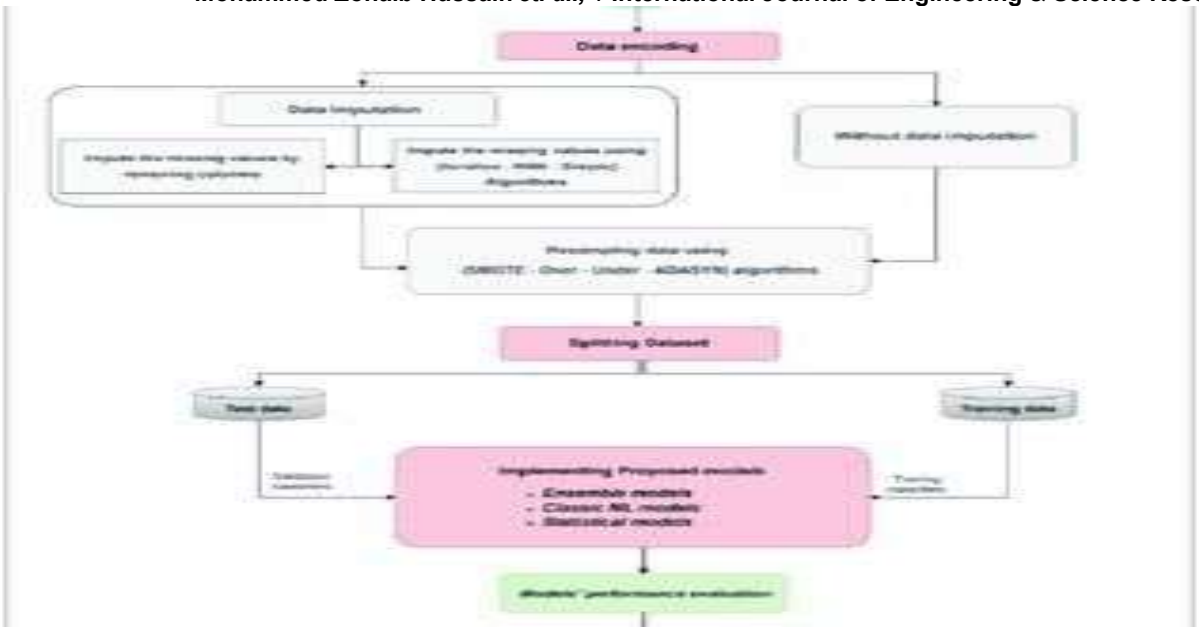
$$\text{Precision} = TP/(TP + FP)$$

$$\text{Recall} = TP/(TP + FN)$$

effectively.
The implementation workflow includes:

1. Loading insurance datasets
2. Cleaning and preprocessing data
3. Handling missing values
4. Encoding categorical variables
5. Applying SMOTE balancing
6. Splitting data into training and testing subsets
7. Training machine learning models
8. Evaluating performance metrics
9. Predicting fraudulent claims

Random Forest and XGBoost models demonstrated superior performance because of their ability to handle nonlinear relationships and noisy datasets

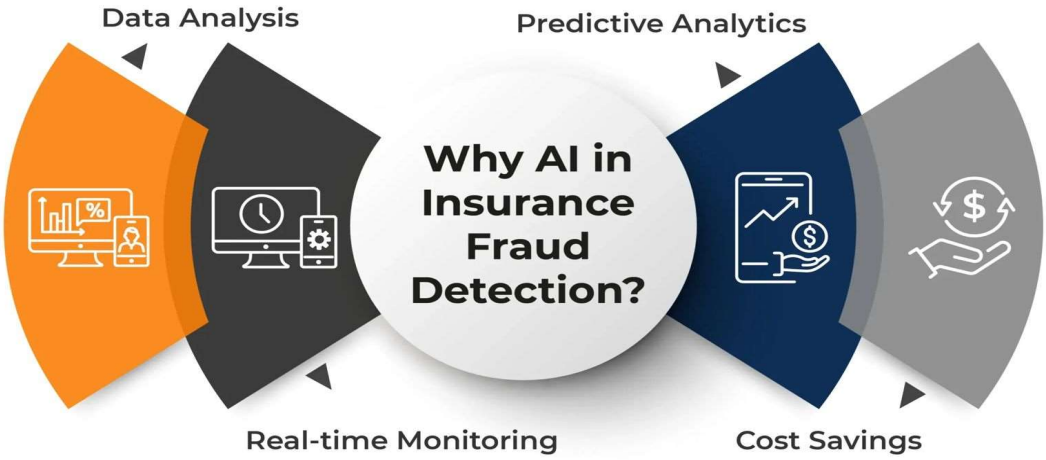


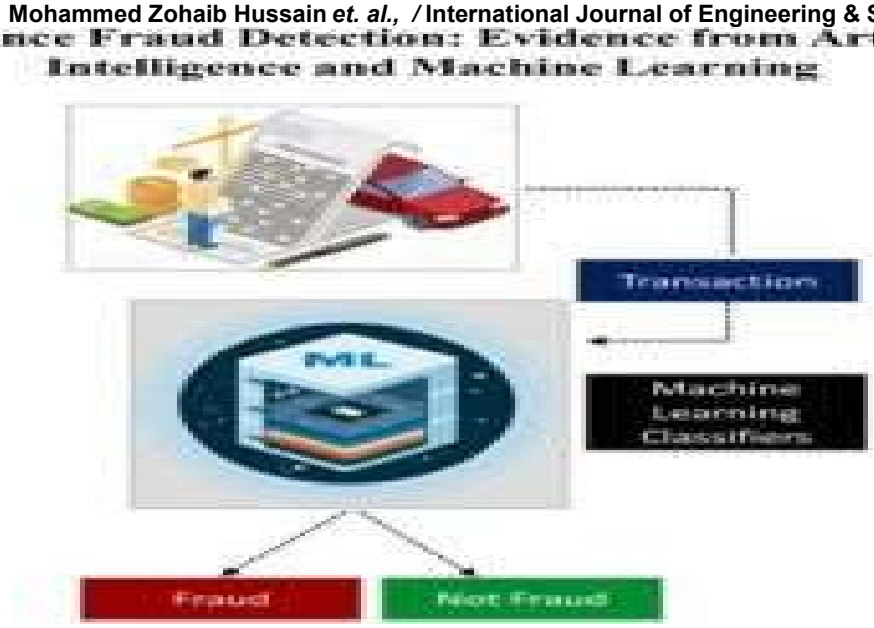
**Testing and Validation**

Testing was conducted using insurance claim datasets containing both genuine and fraudulent records. The dataset was divided into training and testing subsets using an 80:20 ratio. Cross-validation techniques were applied to ensure model stability under different testing conditions. Performance metrics such as Accuracy, Precision, Recall, and F1-Score were analyzed carefully. The experimental analysis showed that Random Forest and XGBoost outperformed traditional classifiers. SMOTE significantly improved the detection of minority fraud cases, while preprocessing improved model consistency and reliability. The proposed framework demonstrated stable performance under multiple validation scenarios and reduced false classification rates significantly.

**Results Screenshots**

The experimental results demonstrated that ensemble learning techniques provide better fraud detection capability compared to traditional machine learning methods. Logistic Regression achieved moderate performance due to its linear assumptions. Decision Trees improved interpretability but suffered from overfitting problems. Random Forest reduced variance and improved stability through ensemble learning. XGBoost achieved the highest overall performance because of its boosting mechanism and optimization capability. The use of SMOTE significantly improved recall by enabling models to learn minority fraud patterns more effectively.






**Conclusion and Future Scope**

This research presented a detailed machine learning-based framework for insurance fraud detection on class imbalance datasets with missing values. The framework integrates preprocessing, missing value handling, SMOTE balancing, feature engineering, and ensemble learning algorithms to improve fraud prediction performance. Experimental results demonstrated that Random Forest and XGBoost outperform traditional classifiers such as Logistic Regression and Decision Trees in terms of Accuracy, Precision, Recall, and F1-Score. Future research can focus on integrating deep learning architectures, blockchain-based claim verification systems, explainable AI techniques, and real-time fraud detection using cloud-based streaming analytics.

**Comparative Performance Table**

| Model | Accuracy | Precision | Recall | F1-Score |
|---|---|---|---|---|
| Logistic Regression | 84% | 81% | 78% | 79% |
| Decision Tree | 87% | 84% | 82% | 83% |
| Random Forest | 92% | 90% | 89% | 89% |
| XGBoost | 94% | 92% | 91% | 91% |

**References**

1) J. Han and M. Kamber, Data Mining: Concepts and Techniques, Morgan Kaufmann, 2022.
2) T. Chen and C. Guestrin, XGBoost: A Scalable Tree Boosting System, ACM SIGKDD, 2016.
3) N. Chawla et al., SMOTE: Synthetic Minority Over-sampling Technique, JAIR, 2002.
4) L. Breiman, Random Forests, Machine Learning Journal, 2001.
5) Scikit-learn Documentation, Machine Learning in Python, 2024.
6) IEEE Research Papers on Insurance Fraud Detection, 2023.