



FRADULANT INSURANCE CLAIMS DETECTION USING MACHINE LEARNING

A Phase 1 Report on Developing a Predictive Model for Fraudulent Insurance Claims Detection

¹Ms. VIDHISHA, ²Dr. Shashidhar Kini K

¹Student, ²Professor & Head

¹Department of Master of Computer Applications,

¹Srinivas Institute of Technology, Valchil Mangaluru, Karnataka, India

Abstract : This study presents a Fraudulent Insurance Claims Detection System developed using supervised machine learning techniques to identify deceptive claims with high accuracy. It utilizes historical insurance claim data, including features such as claim amount, incident type, claim history, policy details, and customer demographics, to classify claims as fraudulent or legitimate. The dataset was sourced from publicly available platforms and underwent comprehensive preprocessing, including handling missing data, feature selection, and encoding categorical variables. The modeling framework evaluates the performance of Logistic Regression, Support Vector Machine (SVM), Random Forest, and XGBoost classifiers using key evaluation metrics such as accuracy, precision, recall, F1-score, and AUC-ROC to determine the most effective approach for real-world deployment.

IndexTerms –*Fraudulent Insurance claim detection, Machine Learning, Classification, Fraud Detection, Random Forest, XGBoost, Predictive Analytics, Supervised Learning, Insurance Analytics*

I. INTRODUCTION

Insurance fraud is a major issue leading to significant financial losses for both insurers and policyholders. Early detection is crucial to reduce these losses and maintain trust in the system. Fraudulent claims often involve exaggerated or false information, which are hard to detect using traditional rule-based methods. Since manual audits and expert judgment may not scale well, there is a growing need for intelligent, data-driven models to analyze historical claim data and effectively identify fraudulent activities.

This project aims to develop a machine learning-based system for the detection of fraudulent insurance claims using structured claim data. The benefits of such a predictive system include:

- **Insurance Companies:** Improved fraud detection rates and reduced financial losses.
- **Policyholders:** Lower premiums through reduced fraudulent payouts.
- **Investigators:** Efficient prioritization of suspicious claims for deeper investigation.
- **Regulatory Bodies:** Support for designing fraud prevention policies based on real data trends.

II. EASE OF USE

The final fraudulent insurance claims detection model is designed for seamless integration into existing claims-processing platforms and mobile applications. Claims adjusters or policyholders simply enter relevant details—such as policy number, claimant age, claim type, claim amount, policy tenure, prior claim history, and any supporting documentation flags—into an intuitive web form or app interface. Once submitted, the model delivers a clear binary outcome (fraudulent/not fraudulent) along with a confidence score, giving users an immediate, data-backed recommendation without any need for specialized technical knowledge. Through a RESTful API deployment, insurers can embed the model directly within their portals or back-office workflows, ensuring a hassle-free, plug-and-play experience. To accommodate diverse operational settings, the system is both scalable and highly adaptable. Insurers can retrain the model periodically using their own historical claims data to capture evolving fraud patterns or demographic shifts. A lightweight variant of the model has also been optimized for mobile and offline use—ideal for field investigators operating in low-connectivity regions. Finally, the user interface supports multiple local languages and customizable validation prompts, making it accessible and user-friendly across global markets.

Prepare Your Paper Before Styling

Before formatting the paper, we focused on gathering reliable data for detecting fraudulent insurance claims. The dataset, sourced from platforms like Kaggle, included features such as claim amount, policyholder age, policy tenure, previous claims, and incident details. We cleaned the data by handling missing values, encoding categorical variables, scaling numerics, and removing outliers to improve model performance.

We tested multiple machine learning models—starting with Logistic Regression and progressing to Decision Tree, KNN, and Random Forest. Their performance was evaluated using accuracy, precision, recall, and F1-score. Hyperparameter tuning was applied to optimize the models and reduce overfitting. Cross-validation ensured the models performed consistently across different data subsets. Feature importance was also analyzed to understand which inputs had the strongest influence on fraud prediction.

After selecting the best-performing model, the paper was organized in IEEE format with clear sections, visuals, and proper references, followed by careful proofreading before submission.

2. Abbreviations and Acronyms

Define abbreviations and acronyms the first time they are used in the text, even if they have already been defined in the abstract. Common abbreviations such as IEEE, SI units (e.g., kg, km), and standard terms do not need to be redefined. Avoid using abbreviations in the title or section headings unless necessary.

In this paper, the following abbreviations and acronyms are used:

- **ML** – Machine Learning
- **KNN** – K-Nearest Neighbors
- **AUC** – Area Under the Curve
- **ROC** – Receiver Operating Characteristic
- **BMI** – Body Mass Index
- **CSV** – Comma-Separated Values
- **RF** – Random Forest
- **TP** – True Positive
- **FP** – False Positive
- **FN** – False Negative
- **TN** – True Negative

RESEARCH METHODOLOGY

This study developed a machine learning model to detect fraudulent insurance claims. It involved collecting data from reliable sources, selecting key predictive features, and applying preprocessing techniques like encoding and scaling. Various ML algorithms were tested, and their performance was evaluated using standard metrics to identify the most accurate and reliable model.

3.1 Population and Sample

The population for this study includes individuals who have filed insurance claims, both genuine and potentially fraudulent. A sample dataset of around 5,000 records was collected from public sources like Kaggle and insurance-related databases. The data covers diverse policyholders across different claim types, regions, and demographic profiles.

Each record contains features such as claim amount, policyholder age, claim type, policy duration, and prior claim history. To ensure data quality, only complete and consistent entries were used; records with missing or illogical values were removed during preprocessing. The dataset is cross-sectional and well-suited for training supervised machine learning models to classify claims as fraudulent or not.

3.2 Data and Sources of Data

The data used in this study was sourced from publicly available insurance claim datasets on platforms such as Kaggle. These datasets are intended for academic and research use and contain anonymized records of insurance claims. The selected dataset includes both genuine and fraudulent claims, making it suitable for supervised learning. It features a variety of claim types, including health, auto, and property, ensuring a diverse representation of real-world insurance data. The dataset also includes demographic details and claim history, providing a comprehensive foundation for fraud detection. The records were carefully selected to maintain data quality and ensure consistency for training machine learning models.

Key features used for modeling include:

- Claim Amount
- Policyholder Age
- **Claim Type** (e.g., Auto, Health, Property)
- Policy Tenure
- Number of Previous Claims
- **Incident Severity**
- Days Since Policy Inception
- **Fraud Reported** (Target variable: 0 = No, 1 = Yes)

The dataset underwent several preprocessing steps to ensure quality and consistency. Missing values were handled appropriately, and categorical variables such as claim type and incident severity were encoded into numerical format. Continuous features like claim amount and policy duration were normalized to bring them to a common scale. Outliers were identified and removed using statistical thresholds to maintain a representative and unbiased training dataset.

3.3 Theoretical framework

The objective of this study is to develop predictive models to identify fraudulent insurance claims based on various claim and policyholder features. The dependent variable is whether a claim is fraudulent (binary classification: 0 = Not Fraudulent, 1 = Fraudulent), while the independent variables include:

- **Claim Amount** (numerical)
- **Policyholder Age** (numerical)
- **Claim Type** (categorical: Auto, Health, Property)
- **Policy Tenure** (numerical)
- **Previous Claims History** (binary: 0 = No, 1 = Yes)
- **Incident Severity** (categorical)
- **Fraud Reported** (target variable)

The relationship between these variables and the likelihood of fraud is expected to be complex and non-linear, making advanced machine learning algorithms such as logistic regression, random forest, and XGBoost ideal for this task.

3.4 Statistical tools and econometric models

This section outlines the methods used to predict fraudulent insurance claims.

3.4.1 Descriptive Statistics

We first analyzed the dataset to understand its distribution. For numerical features like claim amount and policy tenure, we calculated basic statistics such as mean, median, and standard deviation. For categorical variables (e.g., claim type, fraud status), we examined the frequency distribution to identify patterns and any class imbalances.

3.4.2 Machine Learning Models Used

Several machine learning models were applied and compared to predict fraudulent claims:

a) Logistic Regression (Baseline Model)

This model provides insights into the relationship between features like claim amount, policyholder age, and fraud risk. It predicts the likelihood of fraud based on these factors.

b) Random Forest Classifier

This model builds many decision trees and combines their results. It's good at handling complex relationships and avoids overfitting, meaning it's less likely to make mistakes on new data.

c) XGBoost Classifier

XGBoost is an advanced, high-performance model well-suited for complex data. Known for its speed and accuracy, it can handle missing values and is effective in capturing non-linear relationships, making it ideal for predicting fraudulent insurance claims.

d) K-Nearest Neighbour

KNN classifies claims by comparing the features of a new claim to existing cases in the dataset. It uses factors such as claim amount, policyholder details, and previous claims history to assess the likelihood of fraud. KNN is particularly useful for detecting fraud by identifying patterns in similar historical claims.

3.4.3 Evaluation Metrics

The following metrics were used to evaluate the performance of each model in predicting fraudulent insurance claims:

- **Accuracy:** The percentage of correct predictions (fraudulent and non-fraudulent).
- **Precision:** The proportion of predicted fraudulent claims that were actually fraudulent.
- **Recall:** The proportion of actual fraudulent claims correctly identified by the model.
- **F1-Score:** The harmonic mean of precision and recall, balancing both metrics.
- **ROC-AUC:** Measures the model's ability to distinguish between fraudulent and non-fraudulent claims.

3.4.4 Comparison of the Models

The models were compared based on their performance, with particular emphasis on the F1-Score and ROC-AUC. The model with the highest F1-Score and best ROC-AUC was selected for final deployment. Additionally, we analyzed feature importance to identify which factors, such as claim amount, policyholder age, or previous claim history, were most influential in predicting fraudulent claims.

IV. RESULTS AND DISCUSSION

4.1 Results of Descriptive Statics of Study Variables

Table 4.1: Descriptive Statics

Variable	Minimum	Maximum	Mean	Std. Deviation
Claim Amount	50	50000	1200	3500
Previous Claims	0	5	1.2	1.3
Fraud Reported	0	1	0.22	0.33

Table 4.1 Logistic Regression performed well on balanced data but struggled with recall. Random Forest achieved the highest AUC (0.89) and demonstrated robustness to outliers and feature interactions. The model was able to generalize well and showed consistent results across different folds.

ACKNOWLEDGMENT

The author wishes to express sincere gratitude to the Project Guide and Head of the Department of MCA, Dr. Shashidhar Kini K, for his invaluable guidance, constant encouragement, and kind support throughout this research work. Appreciation is also extended to the Principal, Dr. Shrinivasa Mayya D, for fostering an environment conducive to completing this project within the institution. The author thanks the management of Srinivas Institute of Technology for their direct and indirect support. Gratitude is also due to all the faculty members and non-teaching staff of the MCA department for their constant help and support. Finally, the author is indebted to parents and friends for their unwavering support and belief throughout this endeavor.

REFERENCES

- [1] Kaggle. (2023). Stroke Prediction Dataset. <https://www.kaggle.com/datasets/fedesoriano/stroke-prediction-dataset>
- [2] Aslam, A., e2021t al. (). Machine Learning Based Stroke Prediction: A Comparative Study. *International Journal of Biomedical Engineering and Technology*, 37(1), 25-34.
- [3] Sharma, S., & Gupta, R. (2022). Detecting Fraudulent Claims Using Data Mining Techniques. *International Journal of Computer Applications*, 183(35).
- [4] Kaur, G., & Singh, D. (2021). Comparative Analysis of Machine Learning Algorithms for Insurance Fraud Detection. *Journal of Insurance Technology*, 47(4), 11-17.
- [5] Thamarai, M., & Malarvizhi, S. P. (2020). Fraudulent Claim Detection Using Machine Learning. *International Journal of Information Engineering and Electronic Business (IJIEEB)*, 12(2), 23-30. <https://doi.org/10.5815/ijieeb.2020.02.04>
- [6] Dabreo, S., Rodrigues, S., Rodrigues, V., & Shah, P. (2021). Insurance Fraud Detection using Predictive Analytics. *Fr. Conceicao Rodrigues College of Engineering, Mumbai*.