IJIREEICE ISSN (O) 2321-2004, ISSN (P) 2321-5526

International Journal of Innovative Research in Electrical, Electronics, Instrumentation and Control Engineering

Impact Factor 8.021 ⯊ Peer-reviewed & Refereed journal ⯊ Vol. 13, Issue 3, March 2025

DOI: 10.17148/IJIREEICE.2025.13319

FRAUD DETECTION AND ANALYSIS FOR INSURANCE CLAIMUSING MACHINE LEARNING

D.MANOJ KUMAR[1], Dr. K. BANUROOPA[2]

Department of Information Technology, Dr.N.G.P. Arts and Science College, Coimbatore, Tamil Nadu, India[1]

Department of Information Technology, Dr.N.G.P. Arts and Science College, Coimbatore, Tamil Nadu, India[2]

Abstract: Insurance claim fraud detection is a serious problem, resulting in heavy financial loss to insurers and impacting policyholder premium levels. The current methods for fraud detection are rule-based and manual inspection-based and are mostly inefficient and error-prone. The use of Machine Learning (ML) methods to enhance the detection and investigation of fraudulent insurance claims is explored in this paper. Employing several ML algorithms, i.e., supervised learning techniques including Random Forests, Support Vector Machines, and neural networks, we illustrate how fraud activity from historical claims data can be discovered and leveraged to predict fraud risk. The article describes the preprocessing of the claims data, feature engineering, model estimation, and validation procedures in the design of successful fraud detection models.

I. INTRODUCTION

Introduction to Insurance Claims Frauds Detection and Analysis Using Machine Learning It is a cause of serious concern to the insurance industry, the yardstick of financialloss and inefficacy on an enormous scale. Rule-based or manual fraud-detection methods are inefficient in detectinghigh-tech and ever-changing fraud patterns.

Machine learning fraud detection enables insurers to:

1. Automate detection: ML can rapidly and effectively search through huge amounts of claims data to detect suspected fraudulent claims, if any, with minimal or no human intervention.
2. Improve accuracy: Machine learning is able to detect subtle patterns and relationships which may be beyond the reach of conventional methods, thus enabling it to detect fraud accurately with fewer false positives.
3. Evolve with changing fraud strategies: Fraudsters continuously evolve strategies. ML algorithms can beevolved and adapted on the basis of new data so that frauddetection works in the future.
4. Establish customer trust: Through effective detection and clearing of fraud, insurers can offer genuine customers a fair premium and establish customer trust inthe claims process.

II. LITERATURE REVIEW

1. **Feature Engineering and Preprocessing:** Featureengineering and selection are significant in improving the performance of ML models by selecting the most critical data features distinguishing fraudulent from genuine claims.
2. **Detection of Anomalies:** Outlier detection andclustering methods are widely used to identify unusual behavior or patterns within claims, which could indicate fraud.
3. **Performance Metrics:** Accuracy, precision, recall, F1-score, and AUC-ROC are utilized for gauging the performance of ML models for detecting fraud whilekeeping false positives and true positives in balance to obtain optimal results.
4. **Problems with Fraud Detection:** Frauds, which are high-dimensional, biased classes, and dynamic and evolving attacks on the system pose serious problemsin developing good models.
5. **Future Directions:** Deep learning, reinforcement learning, and innovative unsupervised techniques being deployed may potentially yield more effective fraud detection capabilities in the future to render systems more adaptive.

DOI: 10.17148/IJIREEICE.2025.13319

III. METHODS AND MATERIAL

1. Obtaining Dataset: We collected dataset from different sources like haggle , Google. Also we created dummy dataset for analysis and detection of frauds. We used three types of dataset raw dataset which is dummy dataset, processed insurance claim dataset and Integrated dataset.
2. Loading Dataset: For reading dataset we used panda'spython library. We loaded dataset in csv File format. For visualization the features of dataset used python libraries which are matplotlib, seaborn plotly etc. We loaded dataset in jupyter notebook
3. Preprocessing Dataset: Preprocessing the dataset means data wrangling. In preprocessing dataset we reduced reduced redundancy of data. Under preprocessingstep we cleaned the dataset, treated the null values, we didnormalization of datta removed the outliers. Encoded variables categorical variables into continuous variables
4. Comparative Analysis: In comparative analysis we divided dataset into training and testing sets .we Visualized heat map of training and testing dataset. We compared model with Respect to accuracy parameter using different classifier .Finally we generated the ROC curve using the different classifier.
5. Training and Validation: For Training and Testing we splited dataset using sklearn library into 80-20. For training we feuded or trained model with 100% accuracy for testing we got 65% accuracy. For visualizing training and testing dataset we used heat map.

IV.FEATURE EXTRACTION

Sometimes too much information can reduce the effectiveness of data mining. Some of the columns of dataattributes assembled for building and testing a model maynot contribute meaningful information to the model. Somemay actually detract from the quality and accuracy of the model. Adaptive genetic algorithm is a process that identifies important features or attributes of the data. The technique of randomly generate an initial population of feature subsets then Fitness Evaluation perform on every attributes. Finally based on fitness scores important and unique features will consider for fraudulent claims prediction .

V.MODULES

A. DATA COLLECTION:

Data collection is very important process .In this modulesdata is collected from UCI data mining repository which contains the parameters including Policyholder information such as name, policy number, date of policy start, policy deductable, policy annual premium, fraud reported, .Each parameter will be used for fraudulent claims prediction.

B. PRE PROCESSING:

The data we get from different sources may contain inconsistent data, missing values and repeated data. To getproper prediction result, the dataset must be cleaned, missing values must be taken care of either by deleting orby filling with mean values or some other method. Fraudulent claims datasets often have imbalanced classesproposed system uses Resampling technique to used to balance dataset effectively.

C. FEATURE EXTRACTION:

Sometimes too much information can reduce the effectiveness of data mining. Some of the columns of dataattributes assembled for building and testing a model maynot contribute meaningful information to the model. Somemay actually detract from the quality and accuracy of the model. Adaptive genetic algorithm is a process that identifies important features or attributes of the data. The technique of randomly generate an initial population of feature subsets then Fitness Evaluation perform on every attributes .Finally based on fitness scores important and unique features will consider for fraudulent claims prediction .

D. SYSTEM ARCHITECTURE:

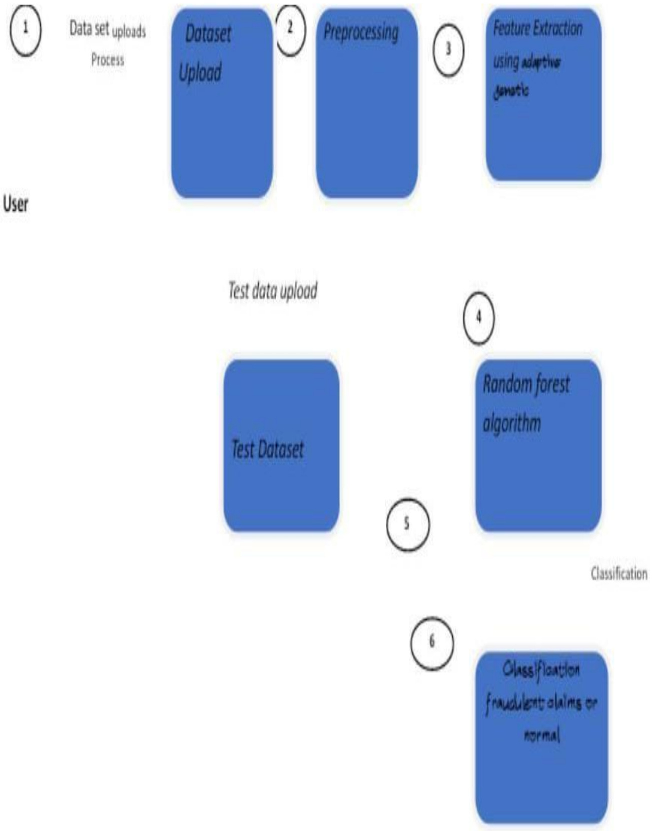
Machine learning model is built with different algorithmsthat is trained by information and data set provided which predict new classification as “fraud” or “not” These algorithms implemented for building model that is trainedusing historical data and that predict unseen data with mostmatching features and then model is tested and validated to evaluate its performance. At first, we have taken the insurance claim dataset (raw) then, preprocessing activities are carried out to improve the quality of dataset

In preprocessing dataset we reduced reduce redundancy ofdata. Under preprocessing step we cleaned the dataset, thenull values, we did normalization of data removed the outliers. Encoded variables dividing the data we usedsclera library.
In training phase development of model using machine learning is done. We calculated performance of model with

DOI: 10.17148/IJIREEICE.2025.13319

respect to accuracy parameters .In testing phase model tested for unknown datasets. Result is calculated onbasis of confusion matrix, precision recall and f1 score. The categorization report includes a number of metrics that are crucial for assessing any model. Accuracy, precision, recall, and F1 are the included measures.

VI.ARCHITECTURAL DIAGRAM



FUTURE WORK

The machine learning models applied on these datasets were able to determine most of the fallacious cases with low false positive rate which suggests with cheap exactness. Certain knowledge sets had severe challenges around data quality, resulting in comparatively poor levelsof prediction. Given inherent characteristics of varied datasets, it would not be sensible to outline optimum algorithmic techniques or use feature engineering process for a lot of higher performance. The models would then beused for specific business context and user priorities. Thishelps loss management units to specialize in a replacementof fraud situations and then guaranteeing that modelssquare measure adapting to spot them. However, it might be cheap to counsel that supported the model performanceon back- testing and talent to spot new frauds, the set of models work the cheap suite to use within the space of theinsurance claims fraud detection.

In order to compare the effectiveness of machine learningand deep learning methodologies, future research should focus on attempting to use an advanced or recently obtained dataset. Additionally, it is advised to utilize a different dataset in light of the fact that the one being usedis unbalanced. Additional evaluation should be done to determine feature relevance across various datasets that may or may not have similar characteristics in order to develop a much more universal method to feature selection and focus. Because this research has been done by using all features in the future, we will do the feature selection to measure the variance between the total and selected features.

VIII.RESULTS AND DISCUSSION

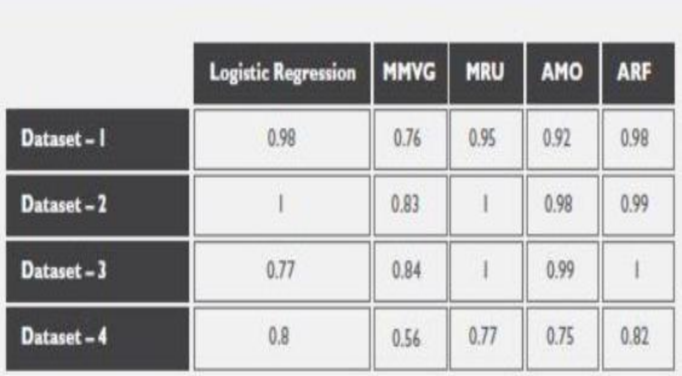
This is most important process in data mining. Data inbroken down into two parts. Training and Testing. There are 80% of data is used for training purpose andreaming 20% used for testing purpose. The training set include target variable. The model is trained by using Random Forest classifier data miningalgorithms.

DOI: 10.17148/IJIREEICE.2025.13319

The training data is used to fit the model. The algorithm uses the training data to learn the relationship between the features and the target. Thencreate an instance of the Random forest model with the default parameters. Then fit this to our training data. We pass both the features and the target variableso the model can learn and build.

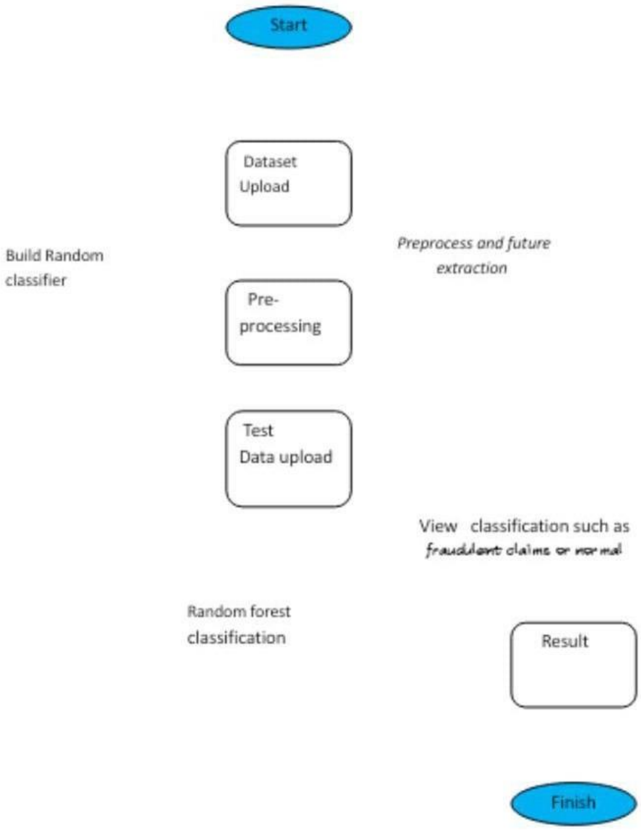
Test data classification Finally, the trained model is applied to test dataset for prediction. Random forest classification data mining algorithm which combinesthe output of multiple decision trees to reach a singleresult. Finally test data is applied to Random forest classification this will classify the test data result as fraudulent claims or normal with more accuracy.

PREDICTIVE ANALYTICS TABLE



	Logistic Regression	MMVG	MRU	AMO	ARF
Dataset – 1	0.98	0.76	0.95	0.92	0.98
Dataset – 2	1	0.83	1	0.98	0.99
Dataset – 3	0.77	0.84	1	0.99	1
Dataset – 4	0.8	0.56	0.77	0.75	0.82

WORK FLOW DIAGRAM



Start
Dataset Upload
Build Random classifier
Preprocess and future extraction
Pre-processing
Test Data upload
View classification such as fraudulent claims or normal
Random forest classification
Result
Finish

DOI: 10.17148/IJIREEICE.2025.13319

IX.SYSTEM TESTING AND MAINTENANCE

A. Objectives of Testing

Software testing is a critical element of software quality assurance that represents the ultimate review of specifications, design and coding. The user tests the developed system and changes are made according to theirneeds. The testing phase involves the testing of developedsystem using various kinds of data. It involves user training, system testing and successful running of thedeveloped system.

The changes are made according to their needs. The testing phase involves the testing of the developed systemusing various kinds of data. While testing, errors are noted and corrections are made system testing is the stage of implementation, which is aimed at ensuring that the system works accurately and efficiently before live operation commences. The candidate system is subject toa variety of test: stress recovery, and security and usabilitytests.

B. Test Plan

Testing is the process of executing a program with the intent of finding any errors. A good test of course has the high probability of finding a yet undiscovered error. A successful testing is the one that uncovers a yet undiscovered error.A test is vital to the success of the system; system test makes a logical assumption that if all parts of the system are correct, then goal will be successfully achieved. The candidate system is subjected to a verity of tests online like responsiveness, its value, stress and security. A series of tests are performed before the system is ready for user acceptance testing.

X.SYSTEM IMPLEMENTATION

Implementation is the stage where the theoretical design is turned into a working system. The most crucial stage in achieving a new successful system and in giving confidence on the new system for the users that it will work efficiently and effectively. The system can be implemented only after thorough testing is done and if it is found to work according to the specification. It involves careful planning, investigation of the current system and its constraints on implementation, design of methods to achieve the change over and an evaluation of change over methods a part from planning. Two major tasks of preparing the implementation are education and training of the users and testing of the system.

The more complex the system being implemented, the more involved will be the systems analysis and design effort required just for implementation. The implementation phase comprises of several activities. The required hardware and software acquisition is carried out. The system may require some software to be developed. This proposed software application is implemented in python as front end.

X.DISCUSSION

1. **Supervised and Unsupervised Learning**: Supervised learning algorithms (decision trees and random forests) areapplied for detecting fraud when one has labeled data available. Unsupervised learning (anomaly detection and clustering) is applied when one has few labeled data and identifies outliers or suspicious trends in claims.
2. **Feature Engineering:** The right features are important for efficient fraud detection, such as claim amounts, number of claims, policyholder behavior, and geographies.Such features can identify patterns that indicate fraudulent behavior.
3. **Imbalanced Data Challenge**: Fraudulent claims are not typical, hence the dataset is extremely imbalanced.
4. **Evaluation Metrics**: Because of the class imbalance nature of fraud detection, precision, recall, F1-score, and ROC-AUC are critical to apply in performance evaluationfor models. They help balance the trade-off between fraud detection and minimizing false alarms.
5. **Ongoing Evolution:** Scammers continuously evolvetheir methods, so machine learning algorithms need to be updated and retrained on fresh data on a regular basis to remain effective. This is important for long-term accuracyand identifying new fraud patterns.
6. **Imbalanced Data Challenge: Fraudulent claims are not typical, hence the dataset is extremely imbalanced.**
7. **Evaluation Metrics**: Because of the class imbalance nature of fraud detection, precision, recall, F1-score, and ROC-AUC are critical to apply in performance evaluationfor models. They help balance the trade-off between fraud detection and minimizing false alarms.
8. **Ongoing Evolution:** Scammers continuously evolvetheir methods, so machine learning algorithms need to be updated and retrained on fresh data on a regular basis to remain effective. This is important for long-term accuracyand identifying new fraud patterns.

DOI: 10.17148/IJIREEICE.2025.13319

XI.CONCLUSION

As the different countries around the world evolve into a more economical-based one, stimulating their economy isthe goal. To fight these fraudsters and money launderers was quite a complex task before the era of machine learning but thanks to machine learning and AI we are ableto fight these kinds of attacks. The proposed solution can be used in insurance companies to find out if a certain insurance claim made is a fraud or not. The model was designed after testing multiple algorithms to come up withthe best model that will detect if a claim is fraudulent or not. This is aimed at the insurance companies as a pitch tocome up with a more tailored model for their liking to theirown systems. The model should be simple enough tocalculate big datasets, yet complex enough to have adecent successful percentile. Under this project, we choosethe sample of more than 1000 data and the data divided into training and testing data. We can see that, compared with the algorithms XGboost and random forest algorithms; have better performance than KNN. We just brought out the feature of machine learning algorithms. We worked with more algorithms and finally calculate which provide more accuracy, precision, and recall.

REFERENCES

1. The Python Language Reference Manual For PythonVersion 3.2 By Guido Van Rossum, Fred L. Drake · 2021
2. "Hands-On Machine Learning with Scikit-Learn, Keras, and Tensor Flow" Author: Aurelian Géron 2022
3. "Hands-On Machine Learning with Scikit-Learn,Keras, and TensorFlow" by Aurélien Géron 2020.
4. "Python Machine Learning" by Sebastian Raschka andVahid Mirjalili.
   "Machine Learning Yearning" by Andrew Ng 2019.
5. "Deep Learning" by Ian Goodfellow, Yoshua Bengio,and Aaron Courville 2018.
6. "Python for Data Analysis" by Wes McKinney."Building Machine Learning PoweredApplications: Going from Idea to Product" byEmmanuel Ameisen 2017.