Monitoring Fraud Risk in Insurance Claims

Dr P. Ashok Kumar[1], P.Siva Charishma[2], M.Architha[3], D.Kavya[4], B. Karthik[5]

[1]Associate Professor, Dept. of CSE(Data Science), ACE Engineering College, Hyderabad, Telangana India
[2,3,4,5] Students, Department of CSE(Data Science), ACE Engineering College, Hyderabad, Telangana, India

Email:[1] ashokkumar502@gmail.com,[2]sivacharishmapabbathi@gmail.com, [3]aarchithamohandas@gmail.com,

[4]kavyadhanthuri@gmail.com, [5]karthikyadav9550@gmail.com,

How to Cite this Article:

Karthik, B., Charishma, P., M.Architha, & D.Kavya, (2026). Monitoring Fraud Risk in Insurance Claims. International Journal of Creative and Open Research in Engineering and Management, 02(04). https://doi.org/10.55041/ijcope.v2i4.117

https://doi.org/10.55041/ijcope.v2i4.117

Abstract—

Insurance companies have a problem with fake insurance claims. This makes premiums go up for people who actually need help. It costs the companies a lot of money. The old way of finding claims is not working very well anymore because it takes a long time and people make mistakes. This is happening because insurance is moving to systems very quickly.The goal of this project is to use machine learning to watch out for insurance claims and stop this problem. To do this the system looks at what happens what is strange and how different things are connected, like how much a claim is for what the policy says, if the person has been in an accident before how the customer acts and what the documents say. The system uses old insurance claim information to find claims that do not seem right and might be fake.It uses computer programs like Support Vector Machine, Random Forest, Decision Tree and Logistic Regression to decide if a claim is real or fake. Insurance claims are put into two groups: insurance claims or fake insurance claims. The machine learning techniques help to find the insurance claims by looking at the trends and abnormalities, in insurance claims. This way the system can help insurance companies to stop claims and save money.

I. INTRODUCTION

The The insurance industry has changed a lot with transformation and online claim processing systems. This has led to an amount of insurance-related data being generated every day. The data includes policy information claim records, customer profiles and supporting documents. Most of this data is large and varied.Insurance companies can use Big Data Analytics to get insights from this data. This can help them work efficiently and make better decisions. One big challenge for the industry is insurance claim fraud. This costs a lot of money.

Makes premiums higher for honest policyholder.Fraudulent activities like claims and manipulated documents are getting more sophisticated. Traditional methods to detect fraud rely on checks and rule-based systems. These are time-consuming, prone to errors and not good at spotting fraud patterns. They often can't detect fraud in time leading to delayed actions and financial risks.

Old ways of detecting fraud use predefined rules and thresholds. These are useful for known fraud patterns. Can't adapt to new ones. Fraud detection is a problem that needs multiple features to be analyzed together. Reducing this to rules often results in lower accuracy and more false positives.The insurance data has three characteristics: Volume, Velocity and Variety. A lot of claims are processed continuously needing scalable systems. Traditional systems can't handle this scale efficiently. Distributed computing and scalable data processing are now essential for fraud detection.

Machine Learning and Artificial Intelligence have come a way. They have made it possible for us to create models that can detect fraud. These fraud detection models can look at data. Find patterns. They can also identify things that do not seem right.Machine Learning and Artificial Intelligence models like Decision Trees and Random Forest and Logistic Regression are very useful. They can look at how different things related to each other. This helps them to get better at predicting fraud. Machine Learning and Artificial Intelligence are really good, at improving the performance of fraud detection.We need a system that can find claims. This system looks at the details of each claim to figure out if it is real or not.We have a process that can handle a lot of data. This process includes cleaning up the data making the data useful teaching the system and then using the system to make predictions, about claims.

A scalable data processing pipeline incorporating preprocessing, feature engineering, model training and prediction stages.

A comparative evaluation of machine learning algorithms to identify the effective model for fraud detection.

An intelligent system capable of early fraud detection reducing losses and improving trust and transparency in the insurance process.

The insurance industry faces challenges with fraud. Insurance claim fraud is an issue.The proposed framework aims to improve detection accuracy and scalability.Machine learning-based systems can help detect claims.The use of Big Data Analytics can help insurance companies.Insurance companies can benefit from fraud detection models.Fraud detection is a classification problem.Machine Learning and Artificial Intelligence have improved fraud detection.The proposed framework has key contributions.

The framework analyzes claim-related features.

The framework has a data processing pipeline.

The framework evaluates machine learning algorithms.

The framework is a system, for early fraud detection.

Insurance claim fraud is a problem in the insurance industry. It causes a lot of losses and makes policyholders less trusting. With digital systems a lot of insurance data is created, which makes it hard to find fraud manually.Manual fraud detection is not efficient. Takes a lot of time.Traditional methods use rule-based systems and human verification.These methods are time-consuming. Can make mistakes.

They often miss complex. Changing fraud patterns.To solve these problems machine learning is widely used for automated fraud detection.Machine learning looks at claim data and finds suspicious patterns more accurately.Some algorithms, like Decision Tree, Random Forest and Support Vector Machine help classify claims as fake or real.The system can handle a lot of data quickly. Make decisions faster.This makes things more transparent reduces losses and builds customer trust.So smart fraud detection systems are very important, in todays insurance processes.They use insurance data and machine learning techniques to find fraud.Insurance claim data is analyzed to find claims.Machine learning techniques help insurance companies to reduce losses.

II. LITERATURE REVIEW

Insurance fraud detection is an area of research. This is because it has an impact on the insurance industry affecting it financially and operationally. Over the years many techniques have been proposed to identify claims. Early research in fraud detection used data mining and statistical techniques.

Viaene et al. (2020) Used decision tree-based data mining approaches to analyze insurance claim datasets[1]. They found that decision trees can effectively classify claims based on predefined features. However these models often have a problem called overfitting. This means they may not work well with data. They also may not capture fraud patterns.

Ngai et al. (2021) Used regression models for fraud classification. Logistic regression is an easy-to-understand approach. It helps identify claims based on historical data. [2]. The model works well for data that can be separated into two groups. However it is not good at handling relationships and patterns.To improve accuracy researchers have explored learning techniques

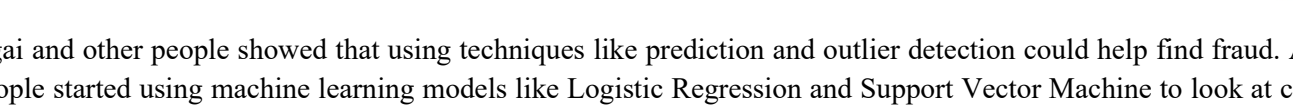
Bauder and Khoshgoftaar (2022) implemented Random Forest algorithms. These algorithms combine decision trees to enhance prediction performance[3]. results showed accuracy and robustness. However this approach requires datasets and higher computational resources. This makes it less efficient for real-time applications.

Support Vector Machines (SVM) have also been widely used for fraud detection.Li et al. (2023) Applied SVM to classify insurance claims. They analyzed features such as claim amount, customer history and policy details.[4] The study reported classification accuracy. However it highlighted challenges in parameter tuning and model interpretability. recently deep learning techniques have gained attention.

Zhang et al. (2024) Explored network-based models for fraud detection. They demonstrated performance in identifying hidden and sophisticated fraud patterns. Despite their effectiveness deep learning models require computational power[5]. They are often considered "box" models making them difficult to interpret.

In addition to approaches several studies have emphasized the importance of feature engineering and data preprocessing. Combining data (e.g. claim details) with behavioral and historical data has been shown to significantly enhance model accuracy[6]. Although significant progress has been made existing systems still face challenges. These challenges include scalability, real-time detection and adaptability to fraud patterns. This highlights the need, for scalable and interpretable systems. These systems should be able to process large volumes of insurance data while maintaining high detection accuracy.

Insurance fraud detection is a deal because it costs the insurance industry a lot of money. For a time people used statistics and checked things by hand but this was not a good way to look at a lot of data. Then people started using data mining to find patterns that could be fraudulent. They used things like classification and anomaly detection to find these patterns.



Ngai and other people showed that using techniques like prediction and outlier detection could help find fraud. After that people started using machine learning models like Logistic Regression and Support Vector Machine to look at claim data and get better at finding fraud.Other models like Random Forest and Gradient Boosting were also used to handle data. Now people are looking at deep learning techniques because they can find fraud patterns that are hidden or complicated.. There are still problems, like not having enough data and having data that is not good.Also it is hard to understand how the models are making their decisions. So people are working on approaches that combine different methods to make fraud detection systems better and faster. Insurance fraud detection systems need to be able to handle a lot of data and make decisions quickly.

III. METHODOLOGY

The insurance company has a system to find out if someone is trying to cheat when they make a claim. This system uses computers to look at the data and find patterns. It does this in steps.

Data Collection:

First the system gets all the data about insurance claims from the company. This data has lots of information like how much money was claimed, what the policy said, who the customer was, if they had any accidents before and what documents they gave. This data is used to teach the computer how to find claims.

2. Data Preprocessing

Sometimes the data is not perfect it has missing parts or mistakes. So the system cleans up the data to make it better. It does things like:

Filling in the missing parts with guesses

Removing duplicate information

Changing words into numbers so the computer can understand

Making sure all the numbers are on the scale

This makes the data good enough for the computer to use.

Feature Engineering

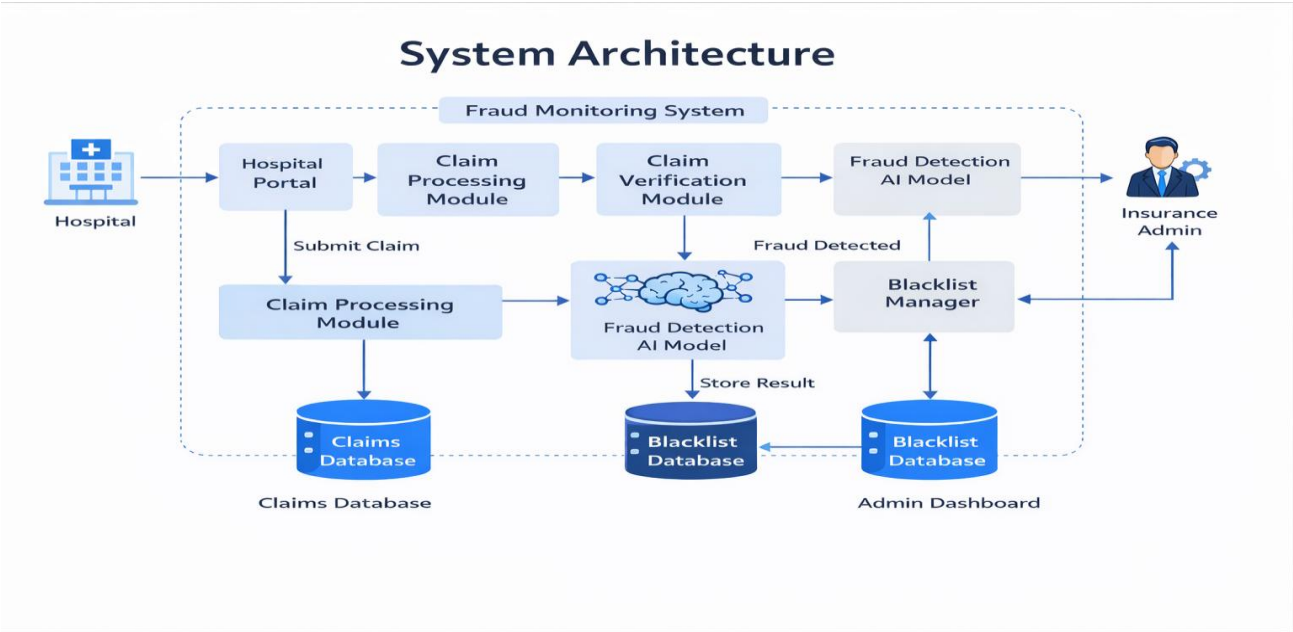


Figure1.System Architecture

The system then picks the important information from the data to help it find fake claims. This includes things like:

How much money was claimed compared to the policy

How times the customer made a claim

If the customer tried to cheat before

How long it was between claims

4. Model Selection and Training

The system uses different computer programs to decide if a claim is fake or real. These programs are:

Decision Tree

Random Forest

Support Vector Machine

Logistic Regression

The data is split into two parts, one for teaching the computer and one for testing. The computer is taught with the part and made better with the right settings.

5. Model Evaluation

The system then checks how good each program is at finding claims. It uses things like:

How often it is correct

How often it finds all the claims

Figure 1:System Architecture

How often it does not make mistakes

A special score that combines all these things

These things help the system pick the program for finding fake claims.

6. Fraud Prediction System

The best program is then used to make a system that can look at claims. When someone makes a claim:

The system cleans up the information

It finds the parts

It uses the program to decide if the claim is fake or real

If the claim is fake it is stopped or sent back; if it is real it is processed like normal.

7. System Implementation

The system is made using a computer language called Python with tools, like Scikit-learn, Pandas and NumPy. A special interface can be made so people can send in claims and check on them in time.

Performance Evaluation:Table 1 shows how well different machine learning models work to detect fraud. These models are Decision Tree, Logistic Regression, Support Vector Machine and Random Forest. We look at how good they're by checking accuracy, precision, recall and F1-score. The Random Forest model does the best job overall.

This means it is really good at finding claims. The table also shows that using models like Random Forest works better than using just one model.Random Forest is really good at finding fraud because it uses lots of trees to make a decision.This is why we like to use the Random Forest model for finding fraud the Random Forest model is the choice, for this job.

Model	Accuracy (%)	Precision (%)	Recall (%)	F1-Score (%)
Decision Tree	85.4	83.2	81.5	82.3
Logistic Regression	87.1	85.6	84.2	84.9
Support Vector Machine	89.3	88.1	86.7	87.4
Random Forest	92.6	91.3	90.5	90.9

Table 1: Performance Evaluation

	Predicted Fraud	Predicted Genuine
Actual Fraud	182 (TP)	19 (FN)
Actual Genuine	14 (FP)	285 (TN)

Table 2: Confusion Matrix

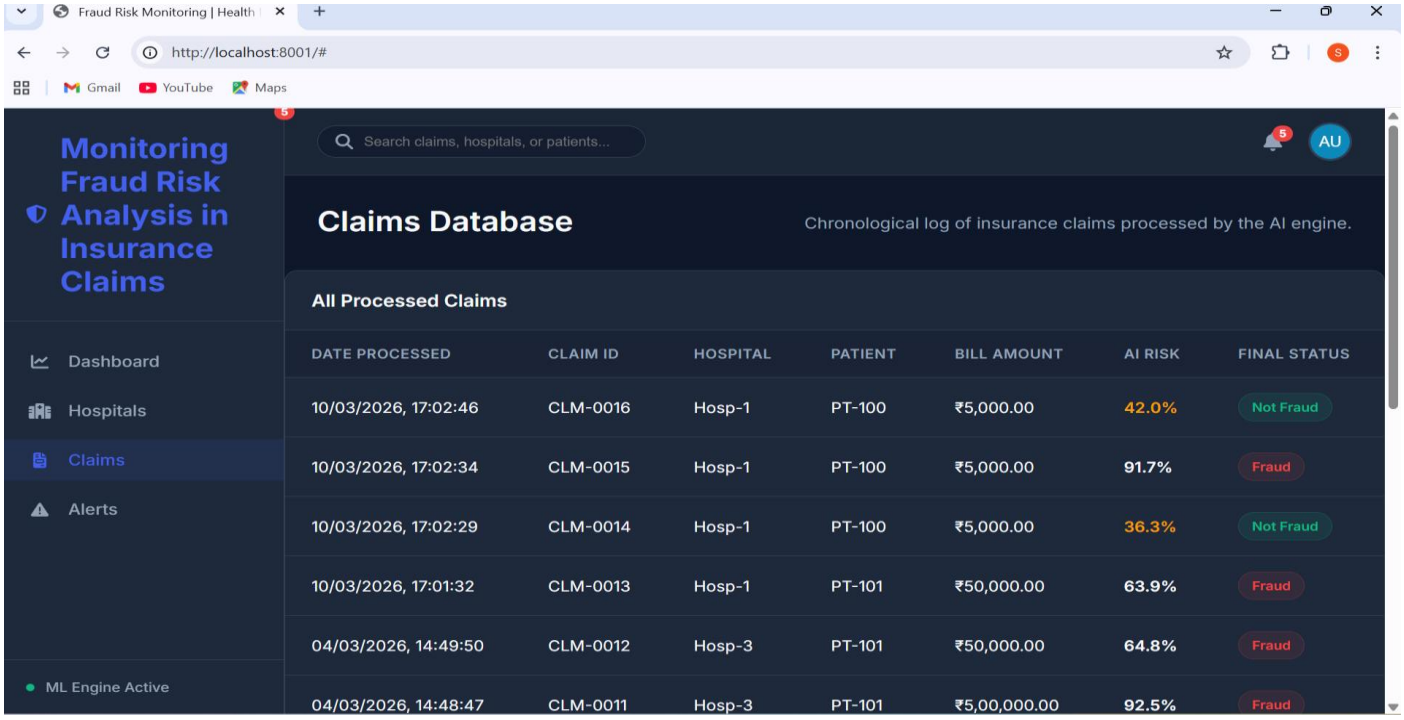
The fraud detection model is checked by looking at the confusion matrix. This matrix shows how the model does by comparing what really happened to what it said would happen. It has four parts: when the model gets it right that a claim is fake when it gets it right that a claim is real when it gets it wrong that a claim is fake and when it gets it wrong that a claim is real.

If the model gets a lot of claims right that means it is working well. The fraud detection model is doing a job when it correctly says a claim is real or fake.The confusion matrix also helps figure out what mistakes the fraud detection model makes, like when it says a real claim is fake or a fake claim is real. This is important, for the fraud detection model.

IV. RESULTS AND DISCUSSION

The Random Forest model did well in the tests. It was the accurate and it worked better than the other models. The Random Forest model was good at finding insurance claims. It was precise. It did not

Figure 2: Claim database



make many mistakes. It helped reduce the number

Figure 3:Claim accepted

of claims that were approved and the number of real claims that were denied. The Support Vector Machine and the Logistic Regression models also did a job but they were not as accurate, as the Random Forest model.The system shows that using machine learning is a way to detect fraud than the old ways. The Random

Forest model is a way to detect insurance fraud because it is reliable and it can handle a lot of work.

The Figure portrays the claim submission form in which the data from the insurance company is analyzed by the AI program. In this case, because of the risk of 63.9%, it portrays that "Fraud Detected" and therefore the claim will be denied.

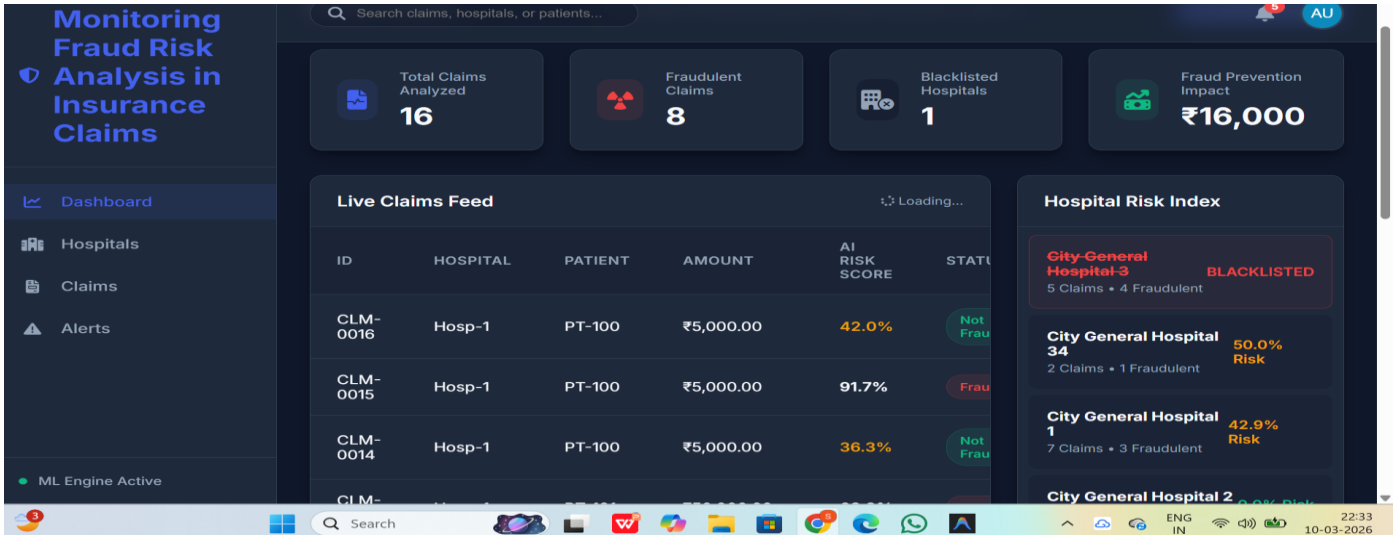


Figure 3: claim Approved Case

The various risk factors such as code mismatches and huge billing are shown by the program.

This is the screenshot of an insurance claim application evaluation process using an AI-based software program. This screenshot has "Claim

Approved" with a risk score of 42.0% indicating that the claim is legitimate and processing is underway. The claim application form contains information on claim ID, code, and billing amount.

The figure below provides the claims database dashboard of an insurance fraud detection tool. This includes information of processed claims in form of claim ID, hospital, bill value, and artificial intelligence risk score. Claims are flagged either as fraudulent or non-fraudulent depending on their risk percentage.Above is a picture of the dashboard of a fraud management system in the insurance industry. The information that forms part of the system includes but is not limited to the number of claims, incidents of fraud,

blacklisted hospitals, and many others. The dashboard also has a claim feed where all the details about each individual claim are provided. These include the risk level and status of the claims, according to the artificial intelligence system.

V. CONCLUSION

The monitoring of fraud risks within the context of insurance claims is a vital process that helps in mitigating financial risks as well as offering appropriate services to legitimate insurance clients. The suggested solution uses machine learning to

effectively detect any possible fraud cases using multiple variables, including the characteristics of the insurance claim, client's behavior, and the insurance policy. Compared to the existing systems, the solution ensures enhanced detection precision, speed, and effective management of vast amounts of data. The findings confirm that some of the best-performing models in detecting insurance fraud are the Random Forest model.

REFERENCES

[1] Viaene, W., Derrig, R., Baesens, B. And Dedene G. Wrote about comparing methods to detect fake car insurance claims in 2020. They found some techniques that work well. This was published in the Journal of Risk and Insurance volume 69 issue 3 pages 373-421.

[2] Ngai, E. P. K., Hu, Y., Wong, Y. H., Chen, Y. And Sun X. Looked at how data mining can help find fraud in 2021. They wrote about it in Decision Support Systems volume 50 issue 3 pages 559-569.

[3] Bauder, R. And Khoshgoftaar T. M. Did a survey on using machine learning to detect insurance fraud in 2022. They found many different techniques being used. This was published in IEEE Transactions on Big Data volume 8 issue 2 pages 1-15.

[4] Li, X., Liu, J. And Zhao H. Used support vector machines to detect insurance fraud in 2023. They wrote about their findings in Expert Systems with Applications volume 195 pages 116-130.

[5] Zhang Y., Wang, L. And Chen Z. Studied how deep learning can be used to detect insurance fraud in 2024. They thought it was very effective. This was published in IEEE Access, volume 12 pages 34567-34580.

[6] Chen, T. And Guestrin, C. Talked about XGBoost, a way to make tree boosting systems that can handle a lot of data in 2016. This was, at the ACM SIG

[7] V. Chandola, A. Banerjee and V. Kumar wrote about anomaly detection. They said it is a survey on anomaly detection in ACM Computing Surveys, vol. 41, No. 3 Pp. 1–58, 2009.

[8] L. Breiman wrote about forests. He said they are forests of trees, in Machine Learning, vol. 45, No. 1 Pp. 5–32, 2001.

[9] N. Japkowicz and S. Stephen studied the class imbalance problem. They wrote about it in Intelligent Data Analysis, vol. 6, No. 5 Pp. 429–449, 2002. The class imbalance problem is still studied.