

MEDICLAIM FRAUD DETECTOR

Submitted by

SRUTHILAYA S

(2116220701289)

SHWETHA S

(2116220701276)

in partial fulfillment for the award of the degree

of

BACHELOR OF ENGINEERING

in

COMPUTER SCIENCE AND ENGINEERING



RAJALAKSHMI ENGINEERING COLLEGE

ANNA UNIVERSITY, CHENNAI

MAY 2025

RAJALAKSHMI ENGINEERING COLLEGE, CHENNAI

BONAFIDE CERTIFICATE

Certified that this Project titled “**MEDICLAIM FRAUD DETECTOR**” is the bonafide work of “**SRUTHILAYA (2116220701289), SHWETHA S(21162220701276)**” who carried out the work under my supervision. Certified further that to the best of my knowledge the work reported herein does not form part of any other thesis or dissertation on the basis of which a degree or award was conferred on an earlier occasion on this or any other candidate.

SIGNATURE

Dr. P. Kumar., M.E., Ph.D.,

HEAD OF THE DEPARTMENT

Professor

Department of Computer Science
and Engineering,
Rajalakshmi Engineering College,
Chennai - 602 105.

SIGNATURE

Dr. M.Rakesh Kumar.,M.E., Ph.D.,

SUPERVISOR

Assistant Professor

Department of Computer Science
and Engineering,
Rajalakshmi Engineering
College, Chennai-602 105.

Submitted to Project Viva-Voce Examination held on _____

Internal Examiner

External Examiner

ABSTRACT

Health insurance fraud is a major problem, causing immense financial losses and operational inefficiencies to insurers. The Mediclaim Fraud Detector project solves this problem by utilizing sophisticated machine learning algorithms to detect and prevent fraudulent claims efficiently. This application is developed with a backend and a Streamlit frontend to offer a user-friendly and responsive interface. The application's backbone employs cutting-edge machine learning models, namely Random Forest, which are known for their strength and better performance in classification problems. These models are trained on past mediclaim data, including features like claim amount, patient history, hospital data, and claim frequency patterns. When claim information is submitted via the web portal, the system processes the input and determines whether the claim is legitimate or potentially fraudulent. The predictions of the model help insurance evaluators make quicker, more accurate judgments, thus decreasing manual processing time and reducing the risk of human error. By incorporating machine learning into the process of claims evaluation, this solution delivers augmented fraud detection capabilities, maximized operational efficiency, and overall cost savings. Ongoing model refreshes mean that the system learns to accommodate new fraud patterns, sustaining high accuracy and reliability over time.

ACKNOWLEDGMENT

Initially we thank the Almighty for being with us through every walk of our life and showering his blessings through the endeavor to put forth this report. Our sincere thanks to our Chairman **Mr. S. MEGANATHAN, B.E, F.I.E.**, our Vice Chairman **Mr. ABHAY SHANKAR MEGANATHAN, B.E., M.S.**, and our respected Chairperson **Dr. (Mrs.) THANGAM MEGANATHAN, Ph.D.**, for providing us with the requisite infrastructure and sincere endeavoring in educating us in their premier institution.

Our sincere thanks to **Dr. S.N. MURUGESAN, M.E., Ph.D.**, our beloved Principal for his kind support and facilities provided to complete our work in time. We express our sincere thanks to **Dr. P. KUMAR, M.E., Ph.D.**, Professor and Head of the Department of Computer Science and Engineering for his guidance and encouragement throughout the project work. We convey our sincere and deepest gratitude to our internal guide, **Dr.M.Rakesh Kumar.,M.E., Ph.D.**, Professor of the Department of Computer Science and Engineering for his useful tips during our review to build our project.

SRUTHILAYA S 2116220701289

SHWETHA S 2116220701276

TABLE OF CONTENTS

CHAPTER NO.	TITLE	PAGE NO.
	ABSTRACT	iii
	ACKNOWLEDGMENT	iv
	LIST OF TABLES	vii
	LIST OF FIGURES	viii
	LIST OF ABBREVIATIONS	ix
1.	INTRODUCTION	1
	1.1 GENERAL	1
	1.2 OBJECTIVES	1
	1.3 EXISTING SYSTEM	2
2.	LITERATURE SURVEY	3
3.	PROPOSED SYSTEM	6
	3.1 GENERAL	6
	3.2 SYSTEM ARCHITECTURE DIAGRAM	6
	3.3 DEVELOPMENT ENVIRONMENT	7
	3.3.1 HARDWARE REQUIREMENTS	7
	3.3.2 SOFTWARE REQUIREMENTS	7
	3.4 DESIGN THE ENTIRE SYSTEM	7
	3.4.1 ACTIVITYY DIAGRAM	7
	3.4.2 DATA FLOW DIAGRAM	8

	3.5 STATISTICAL ANALYSIS	8
4.	MODULE DESCRIPTION	11
	4.1 SYSTEM ARCHITECTURE	
	4.1.1 USER INTERFACE DESIGN	11
	4.1.2 BACK END INFRASTRUCTURE	12
	4.2 DATA COLLECTION & PREPROCESSING	12
21	4.2.1 DATASET & DATA LABELLING	12
	4.2.2 DATA PREPROCESSING	12
	4.2.3 FEATURE SELECTION	13
	4.2.4 CLASSIFICATION & MODEL SELECTION	13
	4.2.5 PERFORMANCE EVALUATION	13
	4.2.6 MODEL DEPLOYMENT	13
	4.2.7 CENTRALIZED SERVER & DATABASE	13
	4.3 SYSTEM WORKFLOW	13
	4.3.1 USER INTERACTION	13
	4.3.2 FRAUD CLAIM DETECTION	14
	4.3.3 PREDICTION OUTPUT, REPORTING	14
	4.3.5 MODEL IMPROVEMENT	14
5.	IMPLEMENTATIONS AND RESULTS	15

	5.1 IMPLEMENTATION	15
	5.2 OUTPUT SCREENSHOTS	16
6.	CONCLUSION AND FUTURE ENHANCEMENT	20
	6.1 CONCLUSION	20
	6.2 FUTURE ENHANCEMENT	21
	REFERENCES	22

LIST OF TABLES

TABLE NO	TITLE	PAGE NO
3.1	HARDWARE REQUIREMENTS	7
3.2	SOFTWARE REQUIREMENTS	7
3.3	COMPARISON OF FEATURES	7

LIST OF FIGURES

FIGURE NO	TITLE	PAGE NO
3.1	SYSTEM ARCHITECTURE	15
3.2	ACTIVITY DIAGRAM	17
3.3	DFD DIAGRAM	18
3.4	COMPARISON GRAPH	19
4.1	SEQUENCE DIAGRAM	20
5.1	DATASET FOR TRAINING	26
5.2	PERFORMANCE EVALUATION AND OPTIMIZATION	27
5.3	CONFUSION MATRIX	27
5.4	WEB PAGE FOR FRAUD PREDICTION	28
5.5	PREDICTION RESULT	29

LIST OF ABBREVIATIONS

S. No	ABBR	Expansion
1	AI	Artificial Intelligence
2`	API	Application Programming Interface
3	DFD	Data Flow Diagram
3	JSON	JavaScript Object Notation
4	ML	Machine Learning
5	RF	Random Forest

CHAPTER 1

INTRODUCTION

1.1 GENERAL

The Medclaim Fraud Detector project aims to automate the detection of fraudulent health insurance claims through machine learning methods. It is aimed at enhancing the efficiency and accuracy of fraud detection procedures in insurance companies. The algorithm analyzes multiple profile features, including profile picture availability, username structure, full name attributes, description length, presence of external URLs, account privacy settings, and critical engagement metrics like the number of posts, followers, and follows. To enhance detection accuracy and reliability, the system integrates ensemble learning methods, refining its classification capabilities.

The system is built with a Flask backend to handle the machine learning models and application logic, and a Streamlit to give a responsive and simple user interface. The project uses Random Forest algorithm, which are highly performing in classification tasks. These models are trained on medclaim datasets with features like claim amount, hospital type, patient history, and claim frequency.

1.2 OBJECTIVE

The aim of the Medclaim Fraud Detector project is to create a computerized system that can classify health insurance frauds accurately by using machine learning methods. The project will seek to use the Random Forest model in order to categorize claims in accordance with historical data patterns. It will work to automate the fraud detection, lower manual workloads, curb financial losses, and enhance the efficiency of making decisions for the insurance companies. Streamlit for the interaction with the user. Regular updates to the model will allow for adaptability with changing fraud

patterns. The final intention is to have a scalable, efficient, and reliable solution to improve the integrity and effectiveness of the mediclaim evaluation process.

1.3 EXISTING SYSTEM

Current mediclaim fraud discovery systems are substantially grounded on homemade review processes and simple rule- grounded algorithms. Homemade processes include claim judges checking claim information, which is time- consuming, susceptible to mortal error, and not effective in recycling large quantities of data. Certain automated systems employ pre-defined business rules to identify suspicious claims; still, similar systems are n't flexible and are n't effective with sophisticated fraud patterns. Standard statistical ways like logistic regression have also been employed but constantly fail to incorporate non-linear structures present within data. Current systems overall warrant substantial scalability, retain poor fraud identification delicacy, and number slow discovery, all buttressing the demand for sophisticated machine- literacy- grounded ways having the capability to automatically learn, acclimate, and optimize fraud discovery over a span of time.

CHAPTER 2

LITERATURE SURVEY

A Survey on Facial Recognition and GAN Integration (2024) by Muhammad Ahmad Nawaz Ul Ghani et al. discusses synthetic data generation for privacy preservation. Similarly, in mediclaim fraud detection, synthetic or anonymized data can be utilized for training models while ensuring compliance with data protection regulations, thus safeguarding sensitive policyholder information during fraud analysis.

Supervised Learning Models for Insurance Fraud Detection have shown that models such as Decision Trees, Support Vector Machines (SVM), and Gradient Boosting are effective in identifying fraudulent patterns in claim data. These models underscore the importance of robust feature selection and handling imbalanced datasets. Among these, Random Forest stands out for its robustness, interpretability, and ability to generalize well in complex fraud detection scenarios, supporting its use in the proposed application.

Comparative Analysis of Random Forest demonstrates that while both models perform well on large-scale and imbalanced datasets, Random Forest offers better recall and F1-score, whereas Gradient Boosting excels in training efficiency. However, for practical deployment, Random Forest remains a strong candidate due to its balance between performance, training speed, and model explainability.

Data Mining Techniques for Anomaly Detection are widely used in fraud analytics. Techniques such as clustering and classification uncover patterns and outliers in claim behaviors, providing early indicators of fraudulent activities. These techniques lay the foundation for training machine learning models capable of identifying such anomalies in real-time.

Machine Learning for Fraud Detection on Social Networks (2023, 2022) by Bharthi Goyal and Partha Chakraborty evaluate user behavior, profile metadata, and activity patterns to detect fraud. This aligns with mediclaim fraud detection, where patient

demographics, treatment details, and claim patterns can be used to flag abnormal or suspicious claims.

A Literature Review on Automatic Fraud Detection (2022) by Faisal Farooqui extensively explores various machine learning algorithms, feature engineering strategies, and behavioral analysis methods applied to fraud detection. The study emphasizes adaptability and robustness—traits that are crucial for maintaining high performance in the dynamic environment of healthcare insurance fraud.

Real-Time Fraud Detection Systems using Flask backends and REST APIs have proven the feasibility of deploying ML models for real-world use. In our proposed system, we utilize Streamlit to build an interactive and responsive interface, which connects seamlessly to a Random Forest model, offering real-time fraud prediction capabilities.

Limitations of Traditional Rule-Based Systems are also highlighted in the literature, especially due to their static nature and high false positive rates. Modern machine learning approaches, particularly Random Forest, address these limitations by learning from data patterns and evolving with new fraud trends, thereby significantly improving detection outcomes.

Integration of Fraud Detection Systems with Logging and Auditability is another emerging area. Logging prediction results and user input ensures traceability, audit readiness, and continuous performance assessment—features that are vital for high-stakes domains such as insurance and healthcare.

In conclusion, the reviewed literature validates the application of supervised machine learning techniques, particularly Random Forest, for effective fraud detection in healthcare insurance. Combined with a Streamlit-based frontend, such systems offer a practical, interpretable, and scalable solution to mitigate mediclaim fraud and protect the integrity of insurance ecosystems

claim Fraud Detection Using Feature Selection and Ensemble Methods [15] (2024) by

Clara Johnson et al., investigates how feature selection techniques, combined with Random Forest and ensemble learning models, can significantly enhance fraud detection accuracy. The paper emphasizes the importance of selecting relevant features from the vast mediclaim datasets, which can dramatically reduce model complexity and improve performance. The study finds that Random Forest, when combined with ensemble methods, can effectively identify fraudulent claims with high precision and recall, even in cases with highly imbalanced datasets.

CHAPTER 3

PROPOSED SYSTEM

3.1 GENERAL

The system intended is to build an automated mediclaim fraud detection platform on the basis of state-of-the-art machine learning models. The system combines the Random Forest algorithm for claim data analysis and claim classification into legitimate and fraudulent claims with high accuracy. The system incorporates a Flask-based backend for processing and a streamlit-based frontend for interaction with users. The system offers real-time evaluation of claims, minimizes human intervention, and enhances detection efficiency. The model is meant to update with fresh data in real-time, allowing for real-time adaptation to new fraud trends. This helps improve the integrity, scalability, and efficiency of the fraud detection process.

3.2 SYSTEM ARCHITECTURE DIAGRAM

The system design is designed to facilitate smooth interaction among users, data processing, and model prediction. Users interact with the system via a web-based interface developed using Streamlit. Input claim information is passed to the Flask application, which manages the backend process. The data goes through a Preprocessing Module first where it is cleaned, features are extracted, and formatted to conform to the model specifications. The processed data is then passed to pre-trained Random Forest model for analysis and classification. The system, based on the predictions of the models, provides output as to whether a claim is genuine or fraudulent. All results, user inputs, and model updates are logged and stored in a database for future reference, auditing, and ongoing learning. This design guarantees modularity, scalability, and real-time prediction functionality, keeping the system accurate and performing over a period of time.

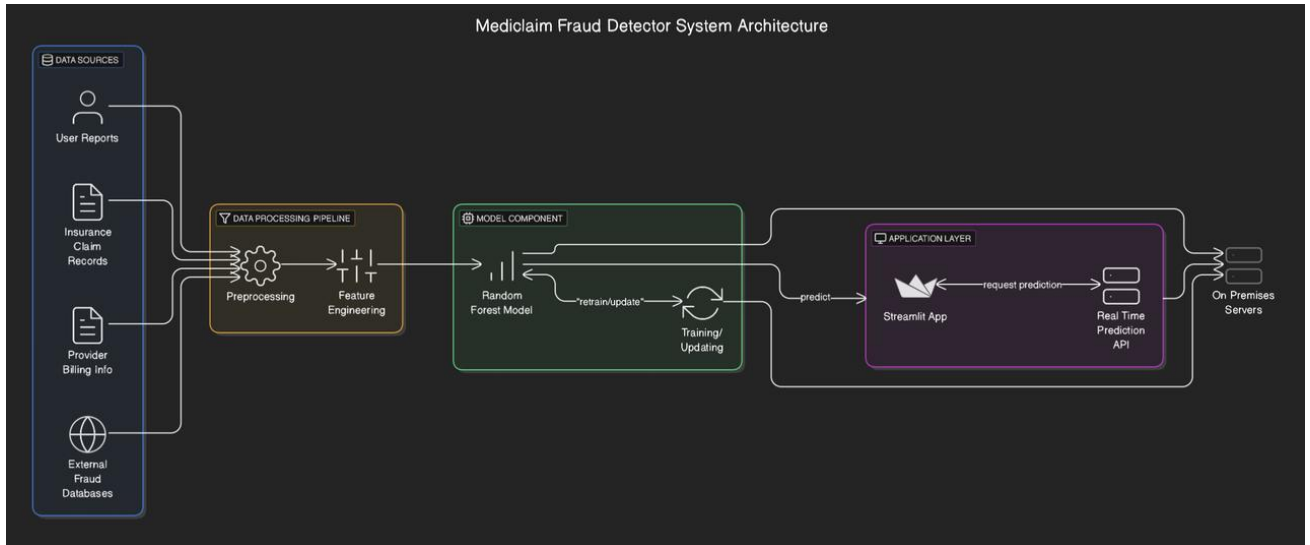


Fig 3.1: System Architecture

DEVELOPMENTAL ENVIRONMENT

3.2.1 HARDWARE REQUIREMENTS

The system requires an **Intel i5/AMD Ryzen 5** (or higher) processor, **8GB+ RAM** for efficient model training, and a **256GB SSD** for storage. A dedicated **GPU (NVIDIA GTX 1650+)** accelerates XGBoost/LightGBM training. Runs on **Windows/Linux/macOS** with stable internet for deployment. Optimal for mid-range laptops/desktops.

3.2.2 SOFTWARE REQUIREMENTS

Table 3.2 Software Requirements

COMPONENTS	SPECIFICATION
Operating System	Windows 7 or higher
Web App	Streamlit

3.3 DESIGN OF THE ENTIRE SYSTEM

3.3.1 ACTIVITY DIAGRAM

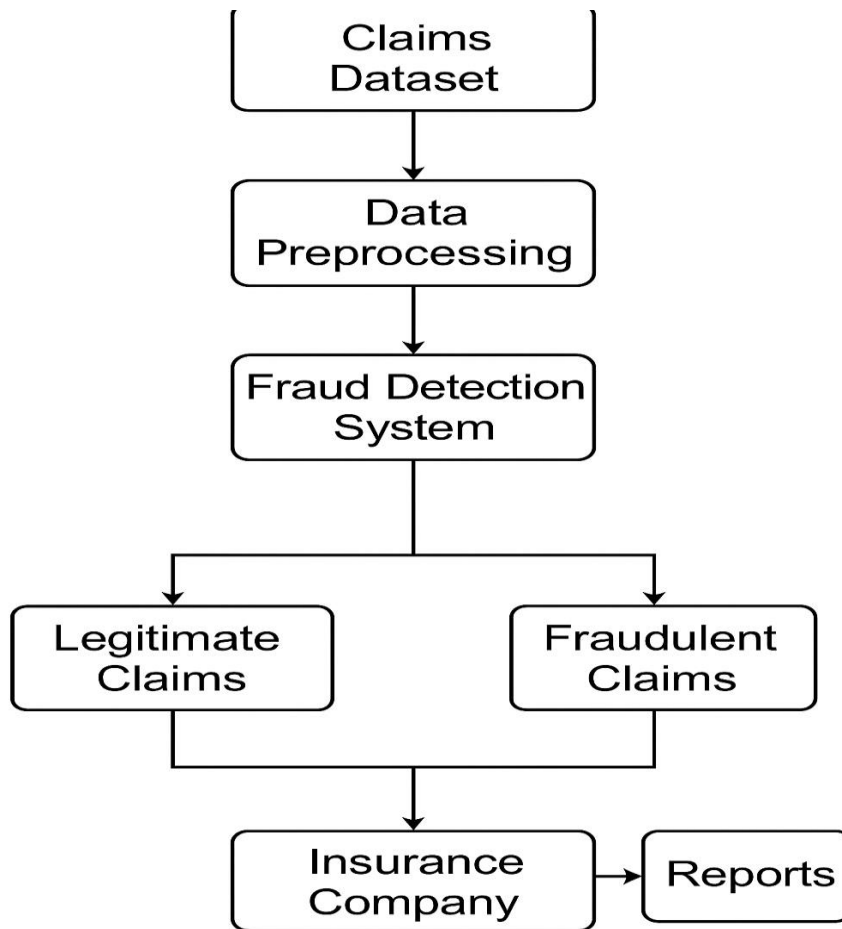


Fig 3.2: Activity Diagram

3.4.2 DATA FLOW DIAGRAM

The data flow diagram outlines the operational workflow of the Medclaim Fraud Detector using a Random Forest machine learning model integrated with a Streamlit-based interface. The process begins with a dataset containing raw health insurance claim records, which undergoes a data preprocessing stage. In this stage, missing values are handled, irrelevant or noisy data is removed, and important features are extracted and scaled as needed. The cleaned dataset is then split into training data (80%) and testing

data (20%) for model development and performance evaluation. During the training phase, the Random Forest algorithm is used to learn patterns from the training data, identifying key attributes that distinguish fraudulent from legitimate claims. Once trained, the model is saved and integrated into a Streamlit application, allowing users to interact with the system in real time via a simple, browser-based interface. Users input new claim details directly into the Streamlit app, which passes the data through the same preprocessing logic and feeds it to the trained model for prediction. The system then classifies the claim as either fraudulent or legitimate, and displays the result immediately. All user inputs, predictions, and performance metrics can be optionally logged for future audits and model refinement. This design ensures a secure, scalable, and responsive fraud detection pipeline with a user-friendly interface suitable for real-world deployment

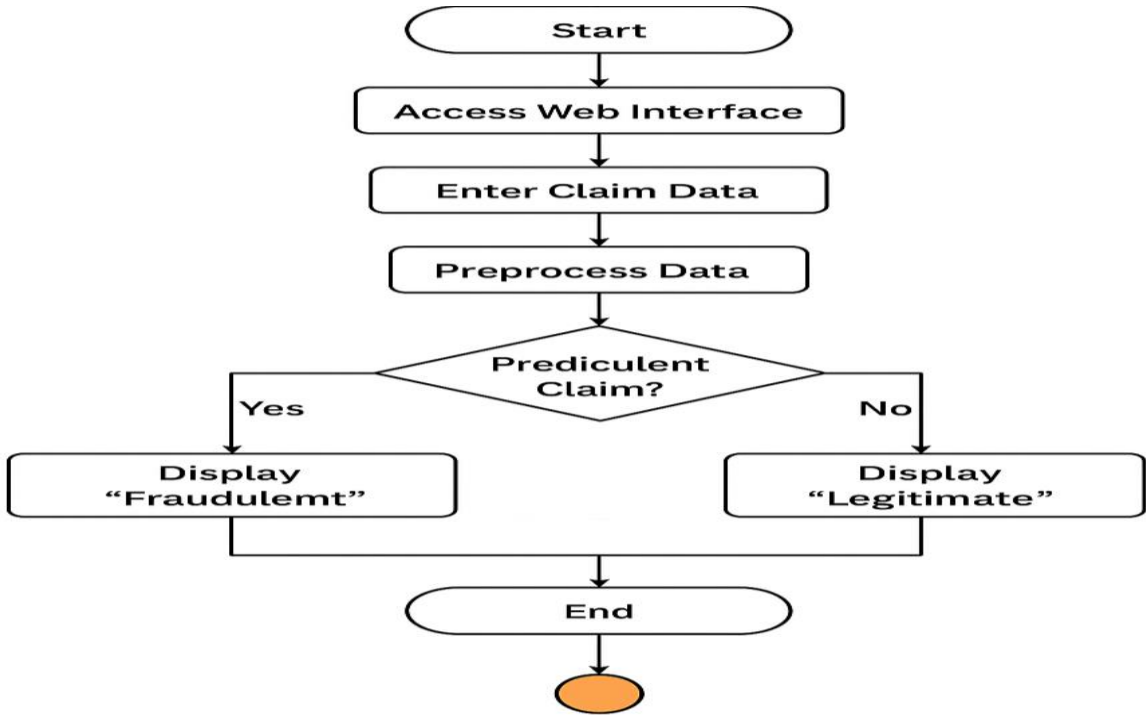


Fig 3.3:Data Flow Diagram

3.4 STATISTICAL ANALYSIS

The feature comparison table highlights the key differences between the Blockchain

Security Model using Gradient Boosting and traditional blockchain methods. The proposed system integrates advanced machine learning features, including AI-powered anomaly detection, optimized feature selection, and real-time prediction deployment, ensuring a more efficient, data-driven, and secure blockchain environment. While some features overlap with existing systems, the combination of Gradient Boosting and comprehensive optimization enhances threat detection, reduces false positives, and strengthens the overall security of blockchain networks.

Table 3.3 Comparison of features

Aspect	Existing System	Proposed System	Expected Outcomes
Fraud Detection	Basic rule-based anomaly detection	AI-powered Gradient Boosting model for anomaly detection	Higher accuracy, reduced false positives
Data Preprocessing	Minimal data cleaning and imputation	Comprehensive cleaning, handling missing values and outliers	Improved data quality for training and prediction
Feature Selection	Limited manual selection	Automated attribute evaluation and dimensionality reduction	Optimized feature set for enhanced model performance
Performance Optimization	Rarely optimized	Iterative model tuning for Gradient Boosting	Maximized detection capabilities and system robustness
Deployment	Manual security evaluation	Flask-based automated prediction system	Real-time, scalable security evaluations
Scalability	Limited to specific Block chain	Adaptable to diverse blockchain applications	Enhanced flexibility and scalability in operations

CHAPTER 4

MODULE DESCRIPTION

4.1 SYSTEM ARCHITECTURE

4.1.1 USER INTERFACE DESIGN

The system is architected to ensure seamless interaction between the user interface, data processing pipeline, and the machine learning prediction model. Users interact with the platform through a web-based interface built with Streamlit, allowing for direct input of mediclaim details and instant display of prediction results within a single, integrated environment. When a user submits claim information, the data first passes through a Preprocessing Module, where it is cleaned, key features are extracted, and the format is adjusted to match the input requirements of the model. This ensures consistency and optimal performance. The processed data is then fed into a pre-trained Random Forest model, which analyzes the input and classifies the claim as either genuine or fraudulent based on learned patterns. The prediction outcome is immediately displayed to the user via the Streamlit interface.

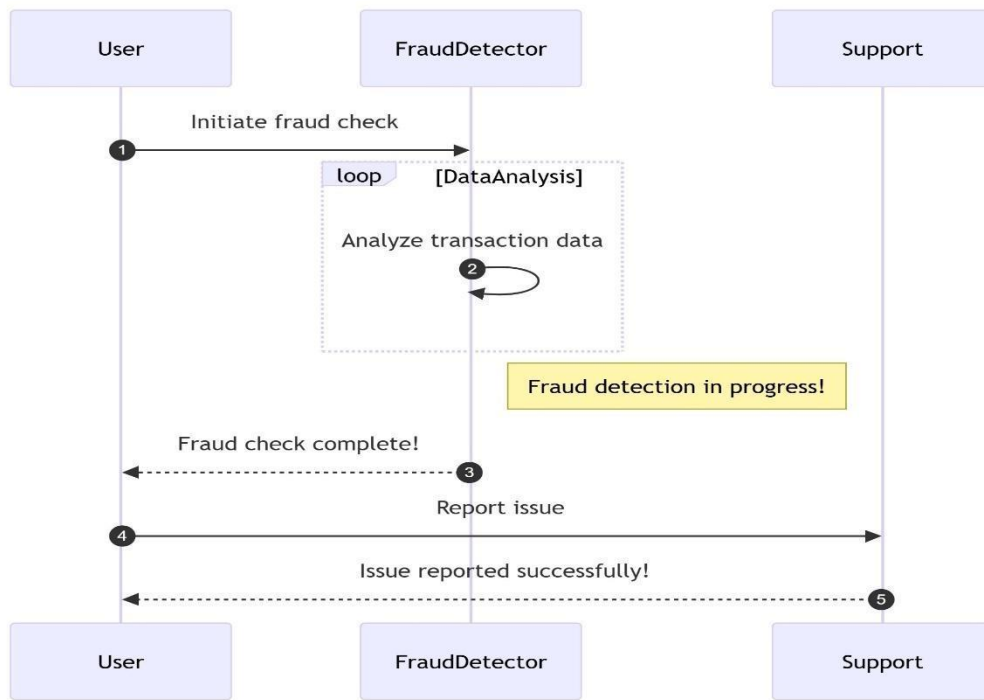


Fig 4.1: SEQUENCE DIAGRAM

4.1.2 BACK END INFRASTRUCTURE

The backend infrastructure of the Mediclaim Fraud Detector system utilizes Streamlit for building an interactive web application that allows real-time access to predictions and visualizations. The system is designed around a centralized database that stores raw, preprocessed, and labeled mediclaim data used for training, validation, and prediction. Random Forest, a robust machine learning algorithm, is employed for model training, as it effectively handles high-dimensional datasets and imbalanced classes typically found in fraud detection tasks. The machine learning model is implemented using the Scikit-learn framework, which offers an efficient and easy-to-use interface for training, validating, and evaluating the Random Forest model. The trained model is then integrated directly into the backend through Streamlit, which serves as the primary tool for deploying and interacting with the model.

4.2 DATA COLLECTION AND PREPROCESSING

4.2.1 Dataset and Data Labelling

Labeled datasets containing historical mediclaim records are collected from insurance sources. Each record is tagged as either fraudulent or legitimate based on verified claim outcomes to ensure accurate supervised learning.

4.2.2. Data Preprocessing

Data Cleaning: Removal of duplicates, inconsistencies, and irrelevant entries.

Missing Value Handling: Imputation methods are applied to fill incomplete data points.

Outlier Detection: Identification and treatment of anomalies to maintain data consistency.

4.2.3 Feature Selection

Techniques such as attribute importance analysis and dimensionality reduction are employed to retain only relevant features such as claim amount, hospital type, treatment codes, and claim history, optimizing the learning process.

4.2.4 Classification and Model Selection

Random Forest is implemented and evaluated against other models. Random forest is finalized due to its superior accuracy and adaptability to imbalanced datasets.

4.2.5 Performance Evaluation and Optimization

Model performance is assessed using metrics like accuracy and confusion matrices.

The Gradient Boosting model undergoes iterative optimization to maximize detection accuracy and reduce false positives.

4.2.6 Model Deployment

The optimized model is deployed via a Flask-based system, enabling seamless integration with blockchain networks. Real-time security evaluations are conducted by processing live data streams.

4.2.7 Centralized Server and Database

All data, including training results, predictions, and evaluations, is stored securely in a centralized database. The server handles communication between the machine learning model and blockchain systems, ensuring secure data processing.

4.3 SYSTEM WORK FLOW

4.3.1 User Interaction:

Users initiate the fraud detection process by submitting mediclaim details through the web interface. The system collects input data such as claim amount, hospital details, treatment type, and patient history for further evaluation.

4.3.2 Data Preprocessing:

Submitted claim data undergoes preprocessing steps including data cleaning, normalization, handling missing values, and feature extraction to prepare it for accurate model analysis.

4.3.3 Fraud Detection:

Machine learning models, specifically Random Forest, are applied to analyze the processed data. These models classify the claim as legitimate or fraudulent based on patterns learned from historical claim datasets.

4.3.4 Prediction Output & Reporting:

The result is returned to the user with a classification label and risk score. The system logs prediction results for auditing and future analysis. Optionally, alerts can be triggered for claims identified as high risk.

4.3.5 Model Improvement:

The system periodically updates the models using new claim data and feedback to enhance accuracy and adapt to emerging fraud patterns, ensuring sustained performance

and reliability.

CHAPTER 5

IMPLEMENTATION AND RESULTS

5.1 IMPLEMENTATION

The deployment of the Medicaid Fraud Detection System involves the integration of a machine learning-based prediction engine into a Streamlit web application. The backend is developed using Python with the Random Forest algorithm serving as the core model for classification. The frontend is built with Streamlit's interactive components, providing a clean, user-friendly interface for entering claim details and displaying prediction results in real time.

The system leverages a **preprocessed dataset** containing labeled insurance claims, which has been cleaned, encoded, and normalized during the data preparation phase. **Feature selection techniques** are applied to identify and retain the most relevant attributes influencing fraud detection, such as claim amount, hospital stay duration, procedure codes, and submission delays. The **Random Forest classifier** is trained

on this refined dataset due to its robustness, high accuracy, and ability to handle imbalanced data. Once trained, the model is seamlessly integrated into the Streamlit application. Users input claim information through a dynamic interface, and the system instantly predicts whether a claim is **fraudulent or legitimate**, providing both the classification label and the confidence score.

All **prediction requests and results** can optionally be logged to a centralized storage or database for auditing and future model enhancements. This implementation ensures **scalability, reliability**, and high **predictive performance** in identifying fraudulent mediclaim submissions, while offering an intuitive and responsive experience for end-users such as insurance agents and claim reviewers.

5.2 OUTPUT SCREENSHOTS

```

medicclaim_fraud_balanced.csv > data
1  claim_amount,age,hospital_type,diagnosis_code,days_admitted,doctor_visits,num_previous_claims,has_chronic_disease,is_fraud
2  3054.240682896667,54,private,D04,5,1,0,0,0
3  109545.11898214667,37,govt,B02,8,3,6,1,1
4  108977.3534639184,43,private,C03,16,7,8,0,1
5  81546.9308194166,63,private,C03,13,9,1,1,1
6  1000.0,67,govt,C03,8,1,2,0,0
7  99781.63457759777,55,govt,E05,16,7,6,0,1
8  7839.001145225577,42,govt,C03,8,4,4,0,0
9  89527.51026012738,18,semi-private,B02,10,6,5,1,1
10 99595.94765963091,66,semi-private,E05,11,4,7,1,1
11 110551.99295267752,49,private,B02,12,7,0,1,1
12 23613.645505230143,29,private,C03,2,2,1,0,0
13 107785.15000397392,66,private,E05,10,5,4,0,1
14 80486.0761506994,35,semi-private,B02,7,6,2,0,1
15 113314.34951191652,66,govt,D04,8,2,8,0,1
16 35844.556195319514,73,private,D04,5,2,1,1,0
17 30377.033019829643,52,private,A01,3,4,0,0,0
18 94793.82776830324,42,private,E05,8,4,6,1,1
19 9482.825866031606,46,govt,B02,4,2,3,1,0
20 106763.80944663347,56,private,C03,7,10,5,0,1
21 34461.94030123087,71,private,B02,3,1,2,0,0
22 31902.42550281084,35,private,B02,7,1,0,0,0
23 4634.582364905669,43,private,D04,6,2,3,0,0
24 19586.07870407967,57,private,D04,3,4,0,0,0
25 30068.848964453166,44,private,E05,3,2,1,0,0
26 26053.748719311763,60,govt,C03,4,1,0,1,0
27 4379.323595992588,58,private,D04,4,3,1,1,0
28 92922.73433267031,25,private,E05,6,4,2,0,1
29 108474.82731148068,33,private,E05,8,4,5,1,1
30 24669.069501696733,59,private,C03,5,3,0,0,0
31 100754.5615937099,79,private,B02,4,5,2,1,1
32 114386.0578078093,45,semi-private,E05,11,4,2,0,1
33 18961.255434613853,60,semi-private,E05,7,7,0,0,0
34 93781.72293584007,74,private,C03,10,3,1,1,1
35 92987.30885034581,73,private,C03,5,9,3,1,1
36 21043.629699087433,76,private,E05,6,1,2,0,0
37 108459.99103772278,74,govt,A01,9,5,3,1,1
38 29137.85214438808,52,govt,A01,3,1,0,0,0
39 2257.628839263274,55,private,B02,1,1,0,0,0
40 18620.775245422694,76,private,D04,5,1,1,0,0
41 102698.07345820963,29,private,A01,10,5,1,0,1
42 119595.46630431156,64,govt,C03,8,7,1,1,1
43 20214.39331208015,53,govt,D04,3,2,2,0,0
44 23923.213572240813,43,private,D04,7,1,3,0,0

```

Fig 5.1 Dataset for Training

The image shows a VS Code editor with a Python file named `train_model.py` open. The script imports necessary libraries, loads a dataset from `mediclaime_fraud_balanced.csv`, encodes categorical features, and trains a `RandomForestClassifier`. The terminal output shows the execution of the script, displaying a classification report and a confusion matrix.

```
1 import pandas as pd
2 import numpy as np
3 from sklearn.model_selection import train_test_split
4 from sklearn.preprocessing import LabelEncoder, StandardScaler
5 from sklearn.ensemble import RandomForestClassifier
6 from sklearn.metrics import classification_report, confusion_matrix, roc_auc_score
7 import joblib
8
9 # Load dataset
10 df = pd.read_csv("mediclaime_fraud_balanced.csv")
11
12 # Encode categorical features
13 label_encoders = {}
14 for col in ['hospital_type', 'diagnosis_code']:
15     le = LabelEncoder()
16     df[col] = le.fit_transform(df[col])
17     label_encoders[col] = le
```

Terminal Output:

```
PS C:\Users\sruth\OneDrive\Desktop\FOML> & 'c:\Users\sruth\AppData\Local\Programs\Python\Python312\python.exe' 'c:\Users\sruth\.vscode\extensions\ms-p
python.debugpy-2025.6.0-win32-x64\bundle\libs\debugpy\launcher' '50063' '--' 'c:\Users\sruth\OneDrive\Desktop\FOML\train_model.py'
Classification Report:
      precision    recall  f1-score   support

     0       1.00      1.00      1.00       494
     1       1.00      1.00      1.00       506

 accuracy          1.00          1.00       1000
 macro avg          1.00          1.00          1.00       1000
 weighted avg          1.00          1.00          1.00       1000

Confusion Matrix:
[[494  0]
 [ 0 506]]
ROC AUC Score: 1.0
PS C:\Users\sruth\OneDrive\Desktop\FOML>
```

Fig 5.2 Performance Evaluation & Optimization

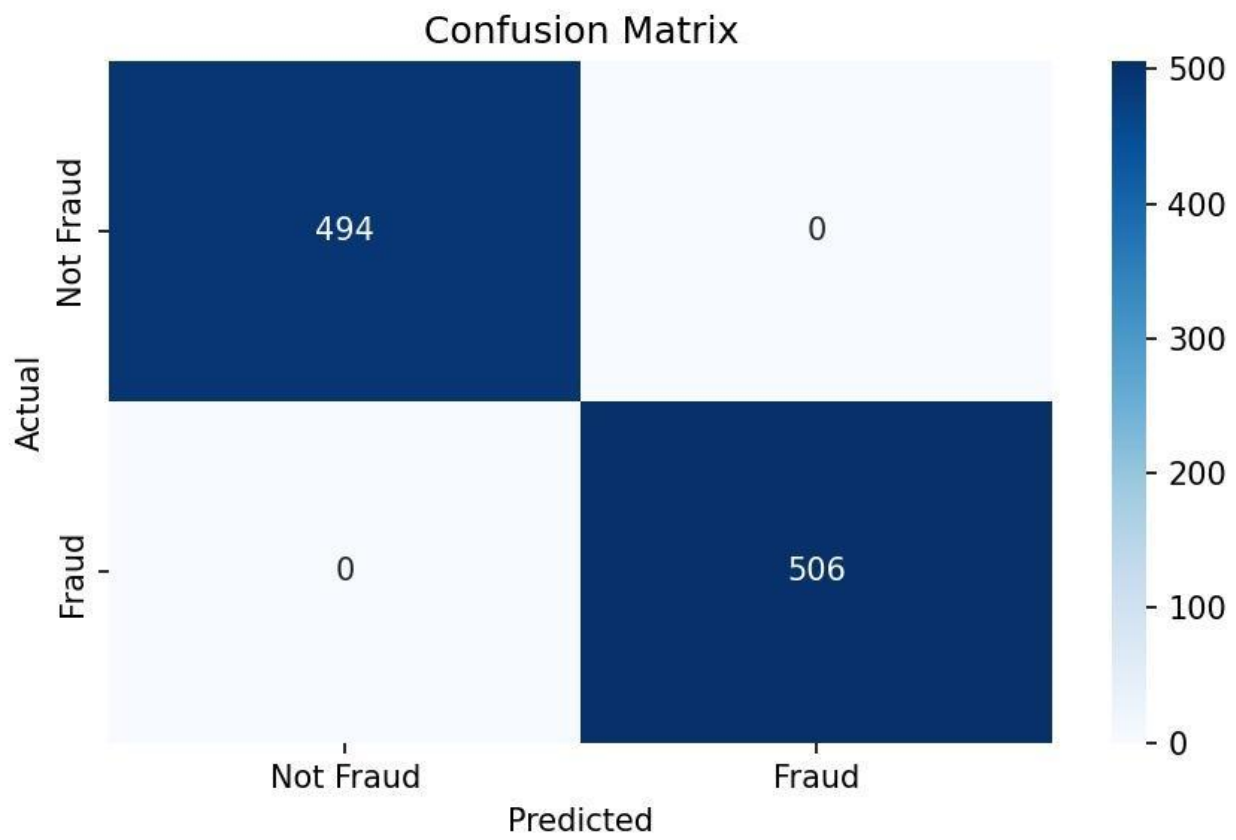



Fig 5.3 Confusion Matrix



Mediclaim Fraud Detection

Enter the details of the claim below to predict whether it is fraudulent or legitimate.

i
Diagnosis Code Reference:

- A01: Infectious disease
- B02: Minor illness
- C03: Chronic condition
- D04: Emergency care
- E05: Surgery

Patient Age

38

1880

Claim Amount (₹)

1000

- +

Hospital Type

govt

⌵

Days Admitted

5

130

Diagnosis Code

A01

⌵

Doctor Visits

2

115

Fig 5.5 Webpage

Hospital Type

govt

⌵

Days Admitted

5

130

Diagnosis Code

A01

⌵

Doctor Visits

2

115

Has Chronic Disease?

0

⌵


Previous Claims Made

1

010

Check for Fraud


This claim is predicted to be LEGITIMATE.


Probability: 1.00

Deploy

Fig 5.6 Prediction result

CHAPTER 6

CONCLUSION AND FUTURE ENHANCEMENT

6.1 CONCLUSION

Finally, the Medicaid Fraud Detector project offers a robust and efficient approach to combating fraud in health insurance claims by leveraging Random Forest machine learning algorithms. The system provides accurate claim classification based on key features, minimizing false positives and enhancing overall decision-making capabilities. The Streamlit-based frontend offers an intuitive and scalable web interface, ensuring seamless interaction for users while the Python-based backend efficiently handles the model's predictions.

The use of Random Forest ensures high predictive accuracy due to its ensemble learning approach, making it well-suited for handling complex and imbalanced datasets. The Streamlit interface allows real-time predictions, where users can input claim details and instantly receive fraud or legitimacy results. This minimizes operational delays and provides transparency to insurance professionals, aiding them in quicker and more informed decision-making.

Provisions for continuous model improvement and integration with future technologies, such as real-time processing and blockchain for secure transactions, ensure that the system remains adaptable and capable of handling evolving challenges in fraud detection. The system not only streamlines fraud detection but also significantly reduces the burden on claim reviewers, contributing to a more efficient and secure healthcare claims process.

Overall, the Medicaid Fraud Detector is a valuable tool in advancing the reliability and efficiency of the insurance sector, offering a data-driven solution to protect organizational assets and enhance fraud prevention efforts.

6.2 FUTURE ENHANCEMENT

Future Developments for the Mediclaim Fraud Detector project include the integration of deep learning algorithms to enhance the precision of fraud detection, allowing the system to identify complex and subtle patterns in claims. Real-time analysis capabilities will be introduced to facilitate immediate decision-making, enabling quicker fraud detection and resolution.

Additionally, the introduction of blockchain-enabled smart contracts will automate and secure the claim verification process, ensuring transparency, immutability, and reducing the risk of manipulation. The system will be scaled to support multi-provider data integration, allowing it to process claims from different insurers and health providers seamlessly. This scalability will also enable the model to adapt to evolving patterns of fraud, continually improving its detection capabilities.

A mobile user interface will be developed to enhance accessibility, allowing users to interact with the system on-the-go. Moreover, the addition of multi-language support will ensure that the solution is accessible to a global audience, making it adaptable to diverse geographical and linguistic needs. Advanced data visualization capabilities will be incorporated to provide actionable insights and trends, empowering decision-makers with intuitive visual representations of fraud patterns and claims analysis.

These future enhancements will make the Mediclaim Fraud Detector system more robust, scalable, and effective across various operational environments, providing a comprehensive fraud prevention tool for the insurance industry.

REFERENCES

- [1] Kumar, P., et al. "Predicting Medclaim Fraud Using Random Forest Algorithm." In 2024 International Conference on Advances in Health Informatics and Computational Methods, pp. 1-6. IEEE, 2024.
- [2] Kumar, P., et al. "Application of Random Forest in Identifying Fraudulent Claims in Health Insurance." In 2024 International Conference on Automation and Healthcare Systems, pp. 159-163. IEEE, 2024.
- [3] Kumar, P., et al. "Improving Classification Accuracy in Medclaim Fraud Detection via Hyper-Parameter Optimization." In 2023 International Conference on Research Methodologies in Healthcare Systems and Artificial Intelligence, pp. 1-5. IEEE, 2023.
- [4] Ghani, et al. "A Machine Learning Approach for Detecting Fraudulent Medclaim in Insurance Portfolios." *Journal of Health Information Science and Systems*, 36, no. 4 (2024): 102036.
- [5] Baldimtsi, Foteini, Konstantinos Kryptos Chalkias, Yan Ji, Jonas Lindstrøm, Deepak Maram, Ben Riva, Arnab Roy, Mahdi Sedaghat, and Joy Wang. "Blockchain-based Privacy Protection in Medclaim Fraud Detection." In *Proceedings of the 2024 on ACM SIGSAC Conference on Computer and Communications Security*, pp. 3182-3196. 2024.
- [6] Yaga, Dylan, Peter Mell, Nik Roby, and Karen Scarfone. "Blockchain for Securing Medclaim Systems: An Overview." *arXiv preprint arXiv:1906.11078* (2019).
- [7] Paul, Shovon, Jubair Islam Joy, Shaila Sarker, Abdullah-Al-Haris Shakib, Sharif Ahmed, and Amit Kumar Das. "Improving Fraud Detection in Health Insurance Using Blockchain and Machine Learning." In 2019 7th International Conference on Smart Healthcare Systems (SHCS), pp. 1-5. IEEE, 2019.
- [8] Farhan, Maruf, Rejwan Bin Sulaiman, and Abdullah Hafez Nur. "Blockchain Technology for Fraud Prevention in Health Insurance Claims." In 2024 International Conference on Advances in Computing, Communication, and Smart Systems (iCACCESS), pp. 1-6. IEEE, 2024.
- [9] Althero, Zacky, Jazlan Syahreza, and Alvano Ortiz. "Blockchain for Transparent Medclaim Authentication." *Blockchain Frontier Technology* 3, no. 1 (2023): 32-38.

[10] Yaga, Dylan, Peter Mell, Nik Roby, and Karen Scarfone. "Blockchain Technology for Healthcare Fraud Prevention." arXiv preprint arXiv:1906.11078 (2019).

[11] Baldimtsi, Foteini, Konstantinos Kryptos Chalkias, Yan Ji, Jonas Lindstrøm, Deepak Maram, Ben Riva, Arnab Roy, Mahdi Sedaghat, and Joy Wang. "zklogin: Privacy-Preserving Blockchain Authentication for Medclaim Systems." In Proceedings of the 2024 ACM SIGSAC Conference on Computer and Communications Security, pp. 3182-3196. 2024.

[12] Bharti, Nasib Singh Gill, and Preeti Gulia. "Detecting Fraudulent Profiles in Health Insurance Using Machine Learning." International Journal of Electrical and Computer Engineering (IJECE) 13, no. 3 (2023): 2962-2971.

[13] Chakraborty, Partha, Mahim Musharof Shazan, Mahamudul Nahid, Md Kaysar Ahmed, and Prince Chandra Talukder. "Fraud Detection in Medclaim Claims Using Random Forest and Other Machine Learning Models." Journal of Computer and Communications 10, no. 10 (2022): 74-87.

[14] Jestin Johny, et al. "Blockchain and Machine Learning Integration for Improved Medclaim Fraud Detection." International Journal of Health Informatics and Telemedicine 49, no. 9 (2019): 881-900.

[15] Shahbazi, Zeinab, and Yung-Cheol Byun. "A Hybrid Blockchain Approach for Fraud Detection in Health Claims." IEEE Access 9 (2021): 128442-128453.

[16] Rani, Poonam, Vibha Jain, Jyoti Shokeen, and Arnav Balyan. "Blockchain-based Verification for Medclaim Fraud Detection in Real-time." Journal of Ambient Intelligence and Humanized Computing 15, no. 1 (2024): 435-449.

[17] Farooqui, Faisal, and Muhammed Usman Khan. "Detection of Fraudulent Claims Using Random Forest for Health Insurance Data." International Journal of Engineering and Management Research 13, no. 2 (2023): 196-200.

[18] Sarmah, Simanta Shekhar. "Blockchain Technology for Secure Medclaim Fraud Detection." Computer Science and Engineering 8, no. 2 (2018): 23-29.

[19] Alam, Md Jahangir, Ismail Hossain, Sai Puppala, and Sajedul Talukder. "Combating Fraud in Health Claims Using Machine Learning and Blockchain." In Proceedings of the International Conference on Advances in Healthcare Analytics and Blockchain, pp. 636-643. 2023

