iJRASET
International Journal For Research in
Applied Science and Engineering Technology

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Volume: 13 **Issue:** V **Month of publication:** May 2025

DOI: https://doi.org/10.22214/ijraset.2025.70890

International Journal for Research in Applied Science & Engineering Technology (IJRASET)
ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.538
Volume 13 Issue V May 2025- Available at www.ijraset.com

Using Machine Learning Technique- Logistic Regression and Random Forest to Detect Fraud in Healthcare Insurance Claims Industry

Shailee Shah[1], Dr. Jyotindra Dharwa[2]
[1]Research Scholar, Ganpat University, Ganpat Vidyanagar Mehsana-Gozaria, Highway, Kherva, Gujarat 384012
[2]Associate Professor, Ganpat University, Ganpat Vidyanagar Mehsana-Gozaria, Highway, Kherva, Gujarat 384012

Abstract: Insurance fraud puts at risk the integrity of insurance systems around the world and can result in large financial losses. The stability and sustainability of the insurance markets depend on the detection and prevention of fraudulent activity. This study suggests a multipronged strategy to improve insurance fraud detection by utilizing cutting-edge technologies. The study starts by examining the state of insurance fraud today, identifying typical fraudulent schemes, and investigating the difficulties insurance firms have in spotting fraudulent activity. It then looks at conventional fraud detection techniques and their shortcomings in dealing with new fraudulent strategies.

We looked into the use of supervised machine learning techniques like decision trees. After data preparation and Principal Component Analysis, Random Forest and Logistic Regressions are used to analyse different aspects and classify claims as either fraudulent or non-fraudulent. Following preprocessing and PCA on the dataset, the outcomes of applying Random Forest and Logistic Regressions are presented in this work.

Keywords: Supervised learning, Machine learning, Fraud detection, Decision Trees, Principal Component Analysis, Data pre-processing, Random Forest, Logistic Regression

I. INTRODUCTION

A contract involving insurance is made between an insurance business (the insurer) and an individual or corporation (the insured). Under certain conditions, the insurer consents to provide financial protection or payment for specific losses or damages listed in the insurance policy. An individual or organization, known as the insured, and an insurance firm, known as the insurer, enter into a legal agreement when they purchase insurance.[2]

Reducing the financial effect of unanticipated events, such as accidents, illnesses, natural catastrophes, or death, is the main goal of insurance. There are various types of insurance, including house, auto, life, and health insurance[16]. Every kind of insurance is designed to offer defence against particular hazards

A key component of healthcare is health coverage, which increases access to medical treatments [1] and provides financial security Because it provides financial stability and increases access to medical care, health insurance is essential to the healthcare industry. These insurance policies assist in reducing the financial burden of medical expenses by covering all or some of the expenditures. This lowers the cost of essential medical operations, treatments, and prescription drugs for both individuals and families.

One effective statistical method for binary classification is logistic regression. Because logistic regression produces a binary output variable, it can be used in situations such as insurance claims. A helpful statistical method for determining a person's likelihood of being qualified to make an insurance claim is logistic regression. The dependent variable in the context of insurance eligibility may be whether a person is successful in obtaining insurance (1) or not (0). Investigating the usefulness of logistic regression in forecasting insurance claims and determining eligibility for health insurance coverage is the focus of this research work. Creating a predictive model enables insurers to detect high-risk policies or clients early on, streamline processes, enhance customer support, and reduce financial losses. It will also be assessed how well logistic regression models forecast insurance claims using the relevant predictors. The study will examine how various factors affect the probability of filing an insurance claim and evaluate the importance of each predictor.

II. RESEARCH SURVEY

Author	Key Findings	Methodology	References
Saraswat et al. (2023)	The goal of the project was to create a machine learning tool that would assist businesses in determining which workers should be covered by insurance. Conserve time and money. The subject of anticipating health insurance claims, which has not been thoroughly investigated previously, was tackled by the study using machine learning approaches. The study's conclusions include the potential to identify insurance fraud and the opportunity to save businesses time and dollars.	Health insurance claims were predicted using machine learning classification methods. devised a tool to assist companies in determining whether to offer insurance to their workers. identified internal insurance fraud in the company	[2]
DeVoe et al. (2009)	According to this survey, children between the ages of 2 and 27 who lived with at least one parent had 73.6% insurance, while those who lived with both parents had 8.0% uninsured. Discordant patterns of coverage, in which parents and children had distinct insurance statuses, were experienced by the remaining 18.4%. In contrast, 17.1% of Americans were uninsured over the same time period, while 82.9% of the country's population as a whole had insurance.	The findings made clear how crucial it is to comprehend how family coverage trends impact kids' access to medical care. It is challenging to find a relationship between family coverage and children's insurance treatment because of the substantial changes in coverage patterns over the last ten years.	[5]
Sun et al. (2019)	They pointed out that some scammers pose as patients in order to obtain insurance funds. Such occurrences are referred to in contemporary literature as disguise. The authors also mentioned the longitudinal and heterogeneous character of the healthcare insurance data. This gives the scammers the opportunity to conceal their actions inside the massive amount of data. Furthermore, the fraudsters' actions are constantly changing, making it challenging to anticipate and identify trends.	demonstrated techniques to detect the fraudulent or hidden behaviours by calculating the similarity between hospital admissions at the patient level, creating a similarity graph, and then clustering the data to extract the semantic meaning of each cluster using a density peak clustering algorithm based on graphs.	[7]
Ramani et al. (2024)	The outcomes of the experiment showed how well these machine learning models performed in producing precise forecasts. The Random Forest model outperformed the other models examined, with an astounding 96.7% accuracy rate.	In order to forecast the amounts of health insurance claims, the study investigated a number of machine learning models, such as Random Forest Regression, Multiple Linear Regression, XGBoost Regression, Gradient Boosting Regression, and Decision Tree .	[10]
Roy and George (2017)	Insurance fraud is a serious issue that costs the sector more than $40 billion a year. The study concentrated on detecting auto/vehicle insurance fraud using machine learning techniques. Metrics including accuracy, precision, and recall were used to assess the machine learning models' performance.	Auto/vehicle insurance fraud was detected using machine learning techniques. used a confusion matrix to compute measures like recall, accuracy, and precision in order to assess the machine learning model's performance.	[12]

Nabrawi and Alanazi (2023)	With an accuracy of 98.21%, the random forest model outperformed the others and identified age, education, and policy type as the key factors influencing healthcare fraud. With accuracies of 80.36% and 94.64%, respectively, the logistic regression and artificial neural network models also demonstrated strong performance. Strong evaluation metrics showed that all three models were successful in identifying healthcare fraud.	In order to create machine learning models for fraud detection, the study employed a retrospective cohort approach and examined a dataset of medical claims. With the majority of cases classified as fraud, the dataset was wildly unbalanced. To balance the dataset, the authors employed the SMOTE approach. To evaluate the dependability of the machine learning models, the authors employed a number of evaluation metrics, such as accuracy, precision, recall, ROC, and AUC. Artificial neural networks, logistic regression, and random forest were the three machine learning models that were employed.	[13]
Smith et al. (2000)	Handling and processing large volumes of insurance claim data necessitate the use of advanced computational tools. Machine learning techniques have emerged as critical in processing such data and extracting essential insights for decision-making. The authors illustrated how data mining and machine learning models are capable of analysing complex patterns in customer retention and insurance claims, where traditional methods fall short.	The study demonstrated how, in areas where conventional approaches are inadequate, data mining and machine learning models may analyse intricate patterns in insurance claims and client retention.	[4]
Konrad et al. (2019)	The authors analysed the insurance claims data using data mining techniques. They employed a data-driven clustering strategy to determine homogeneous service groupings for a specific condition. Within the discovered clusters, they extracted information regarding comorbidities, treatment quality, and illness progression using data analytics methods.	The study identifies the main obstacles that researchers have when utilizing health insurance claims data for sophisticated data analysis, including missing data, coding errors, and temporal shifts in coding methods. In order to overcome these difficulties, the study offers workable answers and suggestions, like classifying related codes, examining provider and temporal effects, and collaborating closely with data partners to comprehend the complexities of the data. To increase the caliber and reproducibility of study findings, the authors stress the significance of standardizing the fusion of aggregated claims data.	4
Seo et al. -2012	The authors used data from national health insurance claims from 2005 to 2008 to create an algorithm that estimates cancer incidence rates. Incident cancer cases were defined as those who were hospitalized in 2008 but had not previously been admitted for the same cancer diagnosis in 2005 or 2007. The incidence rates from the National Cancer Registry of Korea and the incidence rates computed from the claims data were compared.	The incidence rate of all cancers found using insurance claims data was very similar to the rate found in the national cancer registry. The age-, gender-, and disease-specific incidence rates were also similar between the two data sources. Insurance claims data can be a useful and cost-effective resource for health services research if appropriate algorithms are applied.	5

Arunkumar et al. (2021)	Using machine learning techniques, the researchers were able to identify false medical insurance claims. To solve the issue of class imbalance in the data, the researchers employed SMOTE. To find fraud, the researchers employed a hybrid strategy based on classification and clustering.	used a Medicare dataset that was made publicly available to categorize providers as either fraudulent or not. used the Synthetic Minority Over-sampling Technique (SMOTE) to rectify the dataset's class imbalance. used a hybrid strategy that included classification and clustering methods. Several machine learning algorithms were tested to find the best effective one for the job.	[14]
Saripalli et al. (2017)	The study created machine learning techniques to precisely pinpoint medical claims that are likely to be rejected or denied. In order to increase the accuracy of the machine learning models, the study looked into the causes of claims denial and suggested techniques for engineering pertinent characteristics using CARCs. In terms of automating and reducing the risks associated with healthcare claims denials, the study's machine learning technique represents a novel and noteworthy breakthrough in current practice.	identified claims that are likely to be denied or rejected using machine learning classification techniques. investigated the causes of claim denials and used high information gain Claim Adjustment Reason Codes (CARCs) to engineer features. created a new, cutting-edge machine learning engine to automate and reduce the risk of claims denial	[11]

III. DATA PREPROCESSING

A. Data Set Samples

- Patient characteristics -age, gender, race, country, and insurance history
- provider specifics - location, specialization, and status
- claim details - diagnosis codes, bill amounts, and service dates
- outcome factors - processing time, payment amounts, and claim status.

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q
1	BeneID	ClaimID	ClaimStartDt	ClaimEndDt	Provider	InscClaim	AttendingF	OperatingF	OtherPhysi	AdmissionDt	ClmAdmitD	Deductible	DischargeD	Diagnosis	ClmDiagnc	ClmDiagnc	ClmDiagnc
2	BENE1100	CLM46614	12-04-2009	18-04-2009	PRV55912	26000	PHY39092	NA	NA	12-04-2009	7866	1068	########	201	1970	4019	5853
3	BENE1100	CLM66048	31-08-2009	02-09-2009	PRV55907	5000	PHY31849	PHY31849	NA	31-08-2009	6186	1068	########	750	6186	2948	56400
4	BENE1100	CLM68358	17-09-2009	20-09-2009	PRV56046	5000	PHY37239	NA	PHY32468	17-09-2009	29590	1068	########	883	29623	30390	71690
5	BENE1101	CLM38412	14-02-2009	22-02-2009	PRV52405	5000	PHY36965	PHY39296	PHY34976	14-02-2009	431	1068	########	67	43491	2762	7843
6	BENE1101	CLM63689	13-08-2009	30-08-2009	PRV56614	10000	PHY37937	PHY39825	NA	13-08-2009	78321	1068	########	975	42	3051	34400
7	BENE1101	CLM70950	06-10-2009	12-10-2009	PRV54986	8000	PHY40271	PHY40271	PHY40271	06-10-2009	1749	1068	########	597	1745	V4571	78702
8	BENE1101	CLM32075	02-01-2009	07-01-2009	PRV54090	8000	PHY41231	PHY34749	NA	02-01-2009	5699	1068	########	390	1536	73300	7230
9	BENE1102	CLM62376	03-08-2009	07-08-2009	PRV51148	6000	PHY34628	PHY40551	NA	03-08-2009	78605	1068	########	379	56212	25000	30000
10	BENE1103	CLM62784	06-08-2009	09-08-2009	PRV55839	7000	PHY38503	NA	NA	06-08-2009	2859	1068	########	294	42823	4280	6822
11	BENE1103	CLM31519	29-12-2008	05-01-2009	PRV55215	29000	PHY35560	PHY41586	NA	29-12-2008	41401	1068	########	262	41041	3669	V851
12	BENE1103	CLM57949	01-07-2009	09-07-2009	PRV55193	102000	PHY39797	PHY41825	NA	01-07-2009	78605	1068	########	857	3842	25541	78552
13	BENE1103	CLM70083	30-09-2009	07-10-2009	PRV51145	30000	PHY32977	NA	NA	30-09-2009	29590	1068	########	876	29623	2875	27651
14	BENE1103	CLM65412	26-08-2009	29-08-2009	PRV55846	3000	PHY36358	PHY36433	NA	26-08-2009	72989	1068	########	30	43411	27651	2724
15	BENE1104	CLM54944	10-06-2009	15-06-2009	PRV52283	6000	PHY38290	NA	NA	10-06-2009	49121	1068	########	202	49121	2752	7812
16	BENE1104	CLM78682	07-12-2009	13-12-2009	PRV52283	17000	PHY43279	PHY43279	NA	07-12-2009	51881	1068	########	165	51881	2859	4659
17	BENE1104	CLM57153	26-06-2009	30-06-2009	PRV56588	9000	PHY40186	PHY39234	NA	26-06-2009	82100	1068	########	482	82021	E8859	73300
18	BENE1104	CLM34469	19-01-2009	20-01-2009	PRV56118	7000	PHY37795	NA	NA	19-01-2009	29570	1068	########	882	29590	25000	5990

Train_Inpatientdata-15428656275

B. Data Cleaning

Data cleaning, which involves finding and removing any unnecessary or missing duplicate data, is an essential step in the machine learning (ML) pipeline. Raw data is often noisy, inconsistent, and incomplete, which can negatively impact the accuracy of the model and the dependability of the insights it generates. This is why data cleaning attempts to ensure that the data is accurate, consistent, and error-free.

C. Taking Care Of Missing Values

Using the missing no library, an exploratory investigation revealed missingness patterns.

- Features that have missing values more than 30% were assessed for possible removal.
- Mode replacement was used to impute categorical missing data.
- Mean values for normal distributions and median values for skewed distributions were used to impute numerical missing values.
- More advanced imputation techniques, such as KNN imputation, were used for crucial features.

```
# Average features grouped by BeneID
# Average features grouped by Operating Physician

columns_to_transform = [
    "InscClaimAmtReimbursed",
    "DeductibleAmtPaid",
    "IPAnnualReimbursementAmt",
    "IPAnnualDeductibleAmt",
    "OPAnnualReimbursementAmt",
    "OPAnnualDeductibleAmt",
    "DurationofClaim",
    "NumberofDaysAdmitted"

]

for column in columns_to_transform:

    df_train1[f"PerBeneIDAvg_{column}"] = df_train1.groupby('BeneID')[column].transform('mean')
    df_test1[f"PerBeneIDAvg_{column}"] = df_test1.groupby('BeneID')[column].transform('mean')

    df_train1[f"PerAttendingPhysician Avg_{column}"] = df_train1.groupby('AttendingPhysician')[column].transform('mean')
    df_test1[f"PerAttendingPhysician Avg_{column}"] = df_test1.groupby('AttendingPhysician')[column].transform('mean')
9.6s
```

D. Feature transformation

Several transformation techniques were applied to optimize the dataset for modelling:

- Numerical features were standardized using Standard Scalar to achieve zero mean and unit variance
- Categorical features with high cardinality were grouped to reduce levels
- Categorical variables were encoded using either Label Encoder (for ordinal data) or One Hot Encoder (for nominal data)
- Temporal features were extracted from date fields, including day of week, month, and time intervals between service and claim submission

```
from sklearn.preprocessing import StandardScaler
import pandas as pd

# Assume X_train and X_test are your input DataFrames
scaler = StandardScaler()
X_train_scaled = scaler.fit_transform(X_train)
X_test_scaled = scaler.transform(X_test)

# Convert scaled data back to a DataFrame for better readability
X_train_scaled_df = pd.DataFrame(X_train_scaled, columns=X_train.columns)
2.0s

# Save the scaled data to a CSV/Excel file
X_train_scaled_df.to_csv('data/output/scaler/X_train_scaled.csv', index=False)
print("StandardScaler applied, and results saved to 'X_train_scaled.csv'.")

1m 19.2s
StandardScaler applied, and results saved to 'X_train_scaled.csv'.
```

E. One-hot encoding

Categorical data can be transformed into a numerical representation that models can comprehend using this machine learning (ML) technique. When working with categorical variables—such as colours, cities, or animal species—that lack inherent order, it is very helpful.

Before Preprocessing

ChronicCond_Alzheimer	ChronicCond_Heartfailure
2	1
2	1
1	2
1	1
2	1
1	1
1	1
1	2
2	2

After Preprocessing

ChronicCond_Alzheimer_1	ChronicCond_Alzheimer_2	ChronicCond_Heartfailure_1	ChronicCond_Heartfailure_2
1	0	1	0
1	0	1	0
0	1	1	0
0	1	0	1
0	1	0	1
1	0	1	0
1	0	1	0
0	1	1	0
1	0	1	0

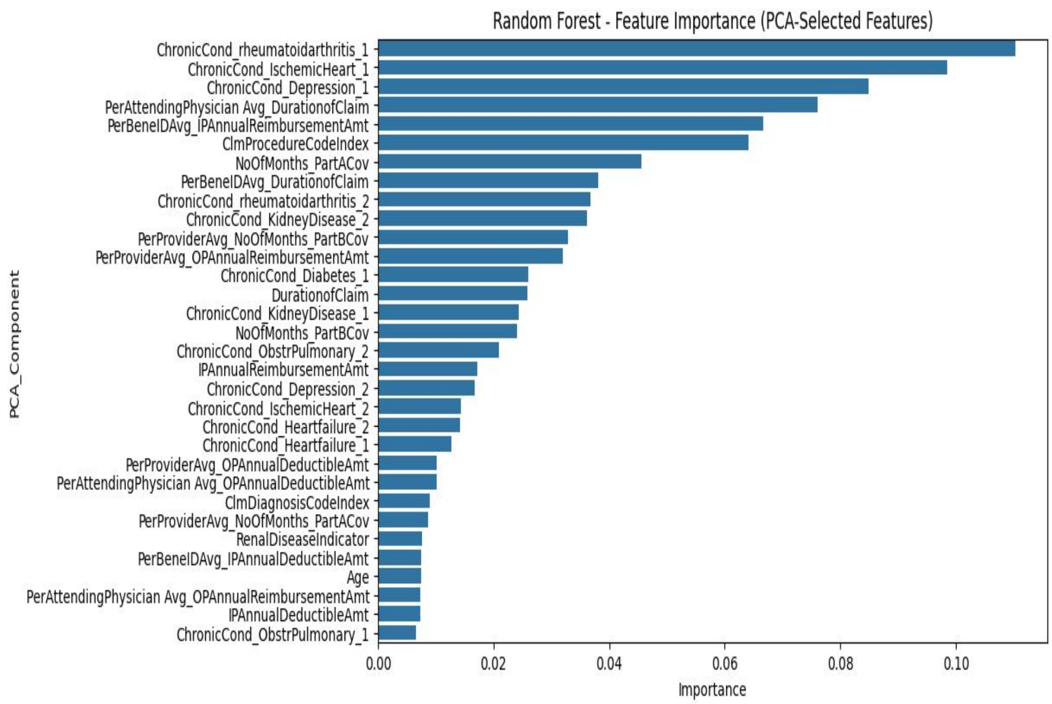
```
from sklearn.preprocessing import OneHotEncoder

# Select only 'ChronicCond_' columns for one-hot encoding
categorical_cols = [col for col in X_train.columns if col.startswith('ChronicCond_')]

# Use OneHotEncoder with sparse_output=False for dense output
encoder = OneHotEncoder(sparse_output=False, handle_unknown='ignore')

# Fit and transform on training data; transform on test data
encoded_train = encoder.fit_transform(X_train[categorical_cols])
encoded_test = encoder.transform(X_test[categorical_cols])  # Transform only

# Convert encoded arrays to DataFrame with appropriate column names
encoded_cols = encoder.get_feature_names_out(categorical_cols)
encoded_df_train = pd.DataFrame(encoded_train, columns=encoded_cols, index=X_train.index)
encoded_df_test = pd.DataFrame(encoded_test, columns=encoded_cols, index=X_test.index)

# Drop original categorical columns and concatenate encoded data
X_train = pd.concat([X_train.drop(columns=categorical_cols), encoded_df_train], axis=1)
X_test = pd.concat([X_test.drop(columns=categorical_cols), encoded_df_test], axis=1)
```

✓ 8.0s

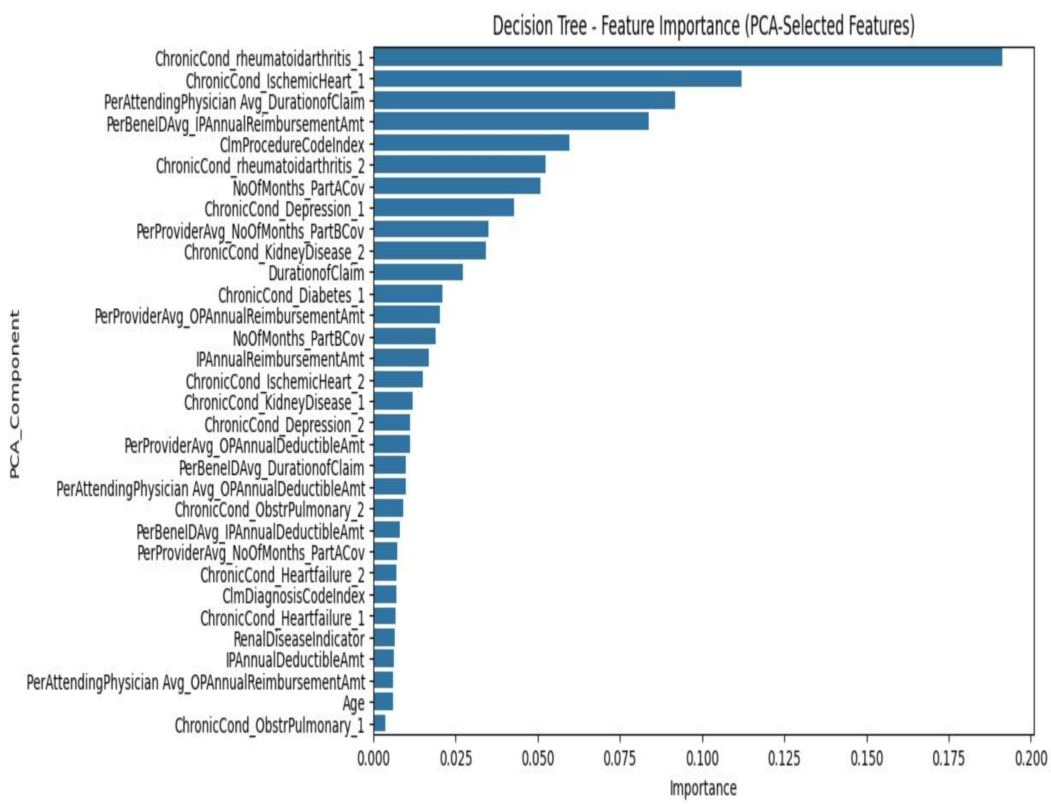
F. Principal Component Analysis

Using Principal Component Analysis (PCA), the data's dimensionality is decreased while as much variance (useful information) is preserved as possible. dealing with multi-featured, high-dimensional datasets, which might lead to issues.

1) Principal Component Analysis selected features for Random Forest



2) Principal Component Analysis selected features for Logistic Regression



International Journal for Research in Applied Science & Engineering Technology (IJRASET)
ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.538
Volume 13 Issue V May 2025- Available at www.ijraset.com

IV. METHODOLOGY

A. *Random Forest*

In order to increase accuracy and decrease false fraud alarms, Random Forest, an ensemble learning technique, is used in healthcare Mediclaim fraud detection.

1. Data Preparation & Feature Selection
   The system gathers and examines previous Mediclaim fraud cases, choosing pertinent features such as
   - Claim Amount (unexpectedly high charges)
   - Claim Frequency (many claims filed in a short period of time)
   - Hospital/Provider Reputation (previous fraud activity)
   - Patient History (medical treatments inconsistent with diagnosis)
   - Billing Patterns (duplicate procedures paid separately).
2. Constructing the Random Forest Framework
   Several decision trees are produced by the method, each of which was trained using a distinct subset of data.
   - Every tree determines if a claim is authentic or fake on its own.
   - The following is used to make the final choice:
   1. **Majority Voting** (for classification)
   2. **Average Prediction** (for regression)
3. Fraud Identification and Categorization
   - Every tree in the forest analyses new claims.
   - A claim is marked for additional examination if the majority of trees determine that it is fraudulent.
   - It is handled normally if the majority of trees consider it to be legitimate.

Formula for Classification:
where $T_i(x)$T_i(x) is the prediction from the i-th tree for input x, and mode represents the majority vote [2]

$$\hat{y} = \text{mode}\{T_1(x), T_2(x), \ldots, T_n(x)\}$$

B. *Logistic Regression*

In the healthcare industry, logistic regression is frequently used to determine if a medical insurance claim is authentic or fraudulent. This is how it operates:

1. Gathering Data and Choosing Features
The model extracts fraud signs by analysing past claims:
   - Claim amount (quite large bills)
   - Claims frequency (many claims in a little period of time)
   - Details about the hospital or provider (previous fraudulent activity, unusual billing)
   - Medical records of patients (unnecessary treatments billed)
   - Geographic discrepancies (claims submitted from odd places)
2. Model of Logistic Regression
The logistic (sigmoid) function is used to forecast the likelihood of fraud:

$$P(Y=1|X) = \frac{1}{1 + e^{-(\beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_n X_n)}}$$

   Where X are input features (provider behaviour, claim details),
   β are where β are training weights,
   PY=1) is the likelihood that a claim is false.
3. (Classification Based on Threshold
- If P(Y=1) **exceeds a predefined threshold** (e.g., 0.75), the claim is flagged as suspicious.
- Lower probabilities indicate **legitimate claims**.

4. Model Training & Optimization
- The model is trained using labeled datasets (fraud vs. non-fraud claims).

V. RESULT DISCUSSION

A. *Confusion Matrix*

A confusion matrix is a simple table that contrasts the actual results with the expectations of a classification model. That distinction is created into four categories: correct forecasts for both classes (true positives and true negatives) and erroneous predictions (false positives and false negatives). [2\

- True Positive (TP): As the model had correctly predicted, the actual outcome was positive.
- True Negative (TN): As the model had correctly predicted, the actual outcome was negative.
- A false positive, or (FP), occurs when a model predicts a positive outcome but yields a negative one instead.
- False Negative (FN): The model generated a positive outcome while it was expecting a negative one.
- Accuracy: Accuracy gauges how well the model predicts outcomes overall.
- Precision: quantifies the percentage of accurately identified fraudulent claims, which is crucial for reducing inflated accusations.
- Recall: measures the capacity to identify all real fraud cases, which is crucial for thorough fraud prevention.
- F1-Score: Offers a balanced metric between precision and recall.

$$Accuracy = \frac{Correct\ prediction}{Total\ cases} * 100\%$$

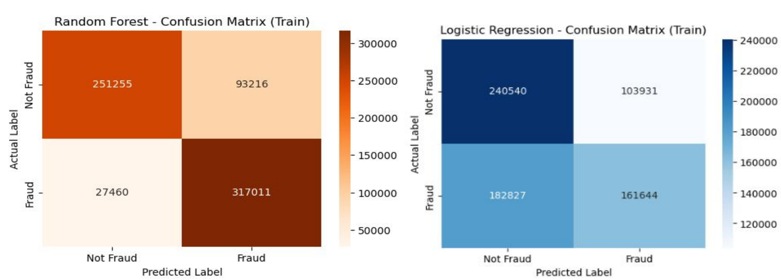
$$Precision = \frac{True\ Positive}{All\ Predicted\ Positives} * 100\%$$

$$Accuracy = \frac{(TP + TN)}{(TP + TN + FP + FN)} * 100\%$$

$$Precision = \frac{TP}{TP + FP} * 100\%$$

B. *Train Set*

Random Forest	Logistic Regression
Accuracy: 0.7910	Accuracy: 0.5848
Precision: 0.7398	Precision: 0.6096
Recall: 0.8979	Recall: 0.4714
F1 Score: 0.8112	F1 Score: 0.5317



VI. CONCLUSION AND FUTURE SCOPE

The empirical investigation in this paper has provided important new information about how well different fraud detection models and strategies work. Every technique has different benefits and drawbacks when it comes to spotting fraudulent activity, ranging from rule-based systems to machine learning algorithms and ensemble methods. Utilizing relevant and informative features from insurance data is crucial, as feature engineering and selection have also been identified as crucial elements in increasing the accuracy of fraud detection. This paper gives result of Random Forest model shows the accuracy of 79% and Logistic Regression model shows the accuracy of 58% for detecting fraudulent claims.

In future we can apply other supervised learning techniques and unsupervised learning techniques for particular disease's impact on the claim for considering it as a legitimate or fraud.

REFERENCES

[1] Thakre V P, Poul R D, Sawarkar A D (March 05, 2025) Predictive Precision: Unraveling Health Insurance Claim Patterns with Logistic Regression and Decision Trees. Cureus J Computer Sci 2 : es44389-025-03010-y. DOI https://doi.org/10.7759/s44389-025-03010-

[2] Saraswat BK, Singhal A, Agarwal S, Singh A: Insurance claim analysis using traditional machine learning algorithms. 2023 International Conference on Disruptive Technologies (ICDT), Greater Noida. 2023, 623- 628. 10.1109/ICDT57929.2023.10150491

[3] Seo HJ, Oh IH, Yoon SJ: A comparison of the cancer incidence rates between the National Cancer Registry and insurance claims data in Korea. Asian Pacific Journal of Cancer Prevention. 2012, 13:6163-6168. 10.7314/apjcp.2012.13.12.6163

[4] Smith KA, Willis RJ, Brooks M: An analysis of customer retention and insurance claim patterns using data mining: a case study. Journal of the Operational Research Society. 2000, 51:532-541. 10.1057/palgrave.jors.2600941

[5] DeVoe JE, Tillotson CJ, Wallace LS: Children's receipt of health care services and family health insurance patterns. The Annals of Family Medicine. 2009, 7:406-413. 10.1370/afm.1040

[6] Antwi S, Zhao X: A logistic regression model for Ghana National Health Insurance claims. International Journal of Business and Social Research. 2012, 139-47.

[7] Seo HJ, Oh IH, Yoon SJ: A comparison of the cancer incidence rates between the National Cancer Registry and insurance claims data in Korea. Asian Pacific Journal of Cancer Prevention. 2012, 13:6163-6168. 10.7314/apjcp.2012.13.12.6163

[8] Sun C, Li Q, Li H, Shi Y, Zhang S, Guo W: Patient cluster divergence-based healthcare insurance fraudster detection. IEEE Access. 2019, 7:14162-14170. 10.1109/access.2018.2886680

[9] Rayan N: Framework for analysis and detection of fraud in health insurance. 2019 IEEE 6th International Conference on Cloud Computing and Intelligence Systems (CCIS), Singapore. 2019, 47-56. 10.1109/CCIS48116.2019.9073700

[10] Ramani K, Kumar ST, Datta PP, Jamuna P, Nithin KS: Predicting health insurance claim amount through machine learning algorithms. 2024 IEEE International Conference on Information Technology, Electronics and Intelligent Communication Systems (ICITEICS), Bangalore, India. 2024, 1-6. 10.1109/ICITEICS61368.2024.10625132

[11] Saripalli P, Tirumala V, Chimmad A: Assessment of healthcare claims rejection risk using machine learning. 2017 IEEE 19th International Conference on e-Health Networking, Applications and Services,Dalian, China. 2017, 1-6. 10.1109/HealthCom.2017.8210758

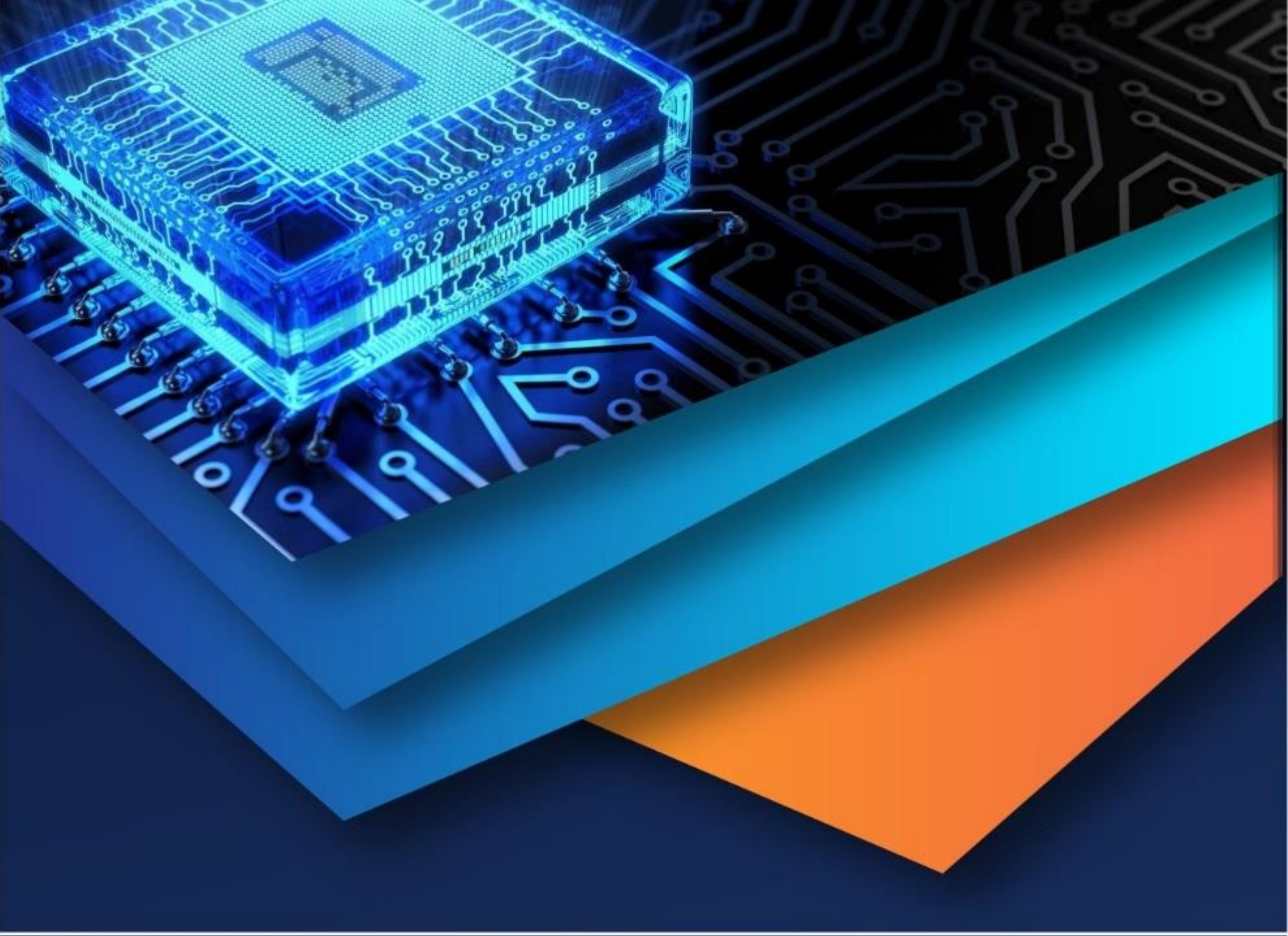
[12] Roy R, George KT: Detecting insurance claims fraud using machine learning techniques. 2017 International Conference on Circuit, Power and Computing Technologies (ICCPCT), Kollam, India. 2017, 1-6. 10.1109/ICCPCT.2017.8074258

[13] Arunkumar C, Kalyan S, Ravishankar H: Fraudulent detection in healthcare insurance. Advances in Electrical and Computer Technologies. Sengodan T, Murugappan M, Misra S (ed): Springer, Singapore; 2021. 711:1-9. 10.1007/978-981-15-9019-1_1

[14] Nabrawi E, Alanazi A: Fraud detection in healthcare insurance claims using machine learning. Risks. 2023, 11:160. 10.3390/risks11090160

[15] https://www.healthcare.digital/single-post/future-of-telemedicine-and-virtual-care-key-trends-and-predictions

[16] https://proassurance.com/knowledge-center/different-types-of-insurance



10.22214/IJRASET

INDEX COPERNICUS 45.98

IMPACT FACTOR: 7.129

IMPACT FACTOR: 7.429

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY