IJCRT.ORG ISSN : 2320-2882

INTERNATIONAL JOURNAL OF CREATIVE RESEARCH THOUGHTS (IJCRT)

An International Open Access, Peer-reviewed, Refereed Journal

Data- Driven Fraud Detection In Medical Insurance System

Udayagiri Lakshmi Sailaja [1], Kosireddi Alekya [2], Kari Sushanth Kumar [3], Vegi Kumar Sankar Naidu [4], Mrs.M.Kalyani [5], Mr.A.Venkateswara Rao[6], Mr.Y.Someswara Rao[7]

[1,2,3,4] **B.Tech Students**, Department of CSE(Artificial Intelligence and Machine Learning), Dadi Institute of Engineering and Technology, NH-16, Anakapalle, Visakhapatnam-531002,A.P

[5] **Assistant Professor**, Department of CSE (AI & ML), Dadi Institute of Engineering and Technology , NH-16, Anakapalle, Visakhapatnam-531002,A.P

Abstract :

Medical insurance fraud is a critical issue for the insurance providers, which finally means business losses and increased premiums for the policyholders. In this line, this project proposes a Fraud Detection System for medical insurance claims using Machine Learning algorithms in order to determine possibly fraudulent claims. The advanced machine-learning techniques within this system make use of the classification algorithms, Logistic Regression and CatBoost, to identify patterns and anomalies indicative of fraud from an analysis of historical claims data. The models are trained using historical insurance claims data to bring out patterns that predict the likelihood of fraud on new or continuing claims. It will use a key combination of such features as the claim amount, claim frequency, policyholder demographics, and diagnosis codes in finding outliers or patterns in behavior that could show fraud. The system optimizes fraudulent activity detection accuracy of the ML models through data preprocessing, feature selection, and then choosing the best suited ML algorithm for the task at hand. This fraud detection system is going to decrease false claims, minimize financial losses for the insurance companies, and strengthen policyholder trust by recognizing and stopping fraudulent activity.

Keywords: — Machine Learning, Fraud Detection, Logistic Regression, XGBoost, CatBoost, Medical Insurance.

INTRODUCTION

This very pervasive medical insurance fraud costs the health care industry billions of dollars yearly. It burdens not only insurance providers but also inflates premiums for honest policyholders, undermining the integrity of healthcare systems worldwide. Traditionalmethods of fraud detection become less adequate with fraudulent techniques increasingly sophisticated. Consequently, there is a big demand for sophisticated, accurate, and efficient systems that could detect and prevent fraudulent activities linked with the medical insurance claim. This project will focus on developing a sophisticated Fraud Detection System (FDS) for medical insurance claims using state of-the-art Machine Learning (ML) technologies. The system will allow insurance companies to enhance their capability in detecting fraudulent claims more accurately and at speed, thanks to the integration of ML algorithms. Core to the functioning of the system is the capability to analyze vast reams of historical claims data for the establishment of anomalies and patterns characteristic of fraud. The project resorts to the strong classification algorithms of Logistic Regression, XGBoost, and CatBoost for that, which have been proved effective in leading anomaly detection tasks across industries. These will be trained on a rich dataset including various features of insurance claims: claim amounts, frequencies, policyholder demographics, and diagnosis codes. Through this training, the models learn irregular patterns and can, therefore, predict potential fraud with high accuracy. It will focus on how to optimize the machine learning models by using heavy data preprocessing, feature selection, and algorithm tuning. The optimization will not only increase the accuracy of fraud detection but also make the system efficient in processing the claims fast while minimizing the computation resources. In the long run, this fraud detection system, if implemented successfully, will help insurance companies bring down the number of false claims, decrease financial losses, and restore faith among policyholders. This project is an important step forward in the fight against medical insurance fraud and a perfect example of how machine learning can transform traditional business processes in the insurance sector.

Dataset : This dataset includes information on the claims data used in Medicare fraud detection and was downloaded from Kaggle. It can give the name of potential health care providers and variables necessary for fraud detection. The dataset has three basic elements: Inpatient Claims, Outpatient Claims, and Beneficiaries Information. These elements provide a comprehensive outline of the claims process as it includes hospital admission, outpatient treatment, and even demographic and health information regarding the beneficiaries.

Data Description :

Potential Fraud or Not (Target variable): Indicates whether a claim is fraudulent or not according to the provider. **Inpatient Data:** This dataset contains information about the patients admitted with admission and discharge dates along with diagnosis codes to the hospital.

Outpatient Data: This contains claims for patients who visit the hospitals but are not admitted.

Beneficiary Details: Information on the patients, including their health conditions, area they live in, and so on.

Claim ID: A value uniquely identifying an insurance claim.

BeneID: A value that uniquely identifies beneficiaries or patients that have received service.

MOTIVATION/ LITERATURE SURVEY

Motivation:

Health insurance faces enormous financial losses and increased premiums due to the high prevalence of fraudulent claims. The current manual and rule-based techniques that are in vogue mostly are inefficient and lack the ability to adapt with new, fast-changing fraudulent techniques. Machine Learning can deliver the transformative solution by using models such as Logistic Regression, XGBoost, and CatBoost to analyze vast data, detect anomalies, and improve fraud detection accuracy over time. This may help insurers automate the process of reducing losses, lowering operational costs, and stabilizing premiums, hence restoring trust and affordability in the sector. Development of an ML-based fraud detection system will help solve one of the critical economic and societal challenges and assure a more sustainable and trustworthy healthcare system.

Literature Review:

Healthcare insurance fraud detection has shifted from traditional rule-based systems, which struggle with modern complexities, to Machine Learning (ML) approaches. Models like Logistic Regression, XGBoost, and CatBoost excel in analyzing large, imbalanced datasets, identifying patterns, and improving accuracy.

While deep learning shows promise for anomaly detection, challenges like interpretability and resource demands remain. ML systems address limitations of traditional methods, offering enhanced efficiency and accuracy in combating fraud, though issues like class imbalance and privacy require ongoing research. This advancement positions ML as a transformative tool in restoring trust and reducing losses in the insurance sector.

ALGORITHMS AND IMPLEMENTATION

Logistic Regression for Predictive Modeling and Fraud Detection :

Logistic Regression is used as a baseline model for binary classification to predict whether an insurance claim is fraudulent or not. Logistic Regression: This is used for binary classification in predicting whether an insurance claim is fraudulent or not. The target variable is Potential Fraud: fraudulent vs. not fraudulent. Logistic Regression computes the probability of fraud by modeling a linear relationship with the features on the target variable. The algorithm trains iteratively, at each epoch, minimizing the log-loss function. Simple, efficient, and interpretable. Hence suitable for handling large datasets while providing a baseline model for fraud detection.

XGBoost for improving prediction accuracy and handle complex patterns in the data:

XGBoost creates an ensemble of decision trees in a sequential manner, where each tree is dedicated to the task of correcting the errors of the previous ones. It optimizes performance by a combination of regularization, gradient descent, and hyper parameter tuning, therefore having the capability to achieve high precision and recall in detecting fraudulent claims. This has been shown to be very effective when there are nonlinear relations and interactions among features.

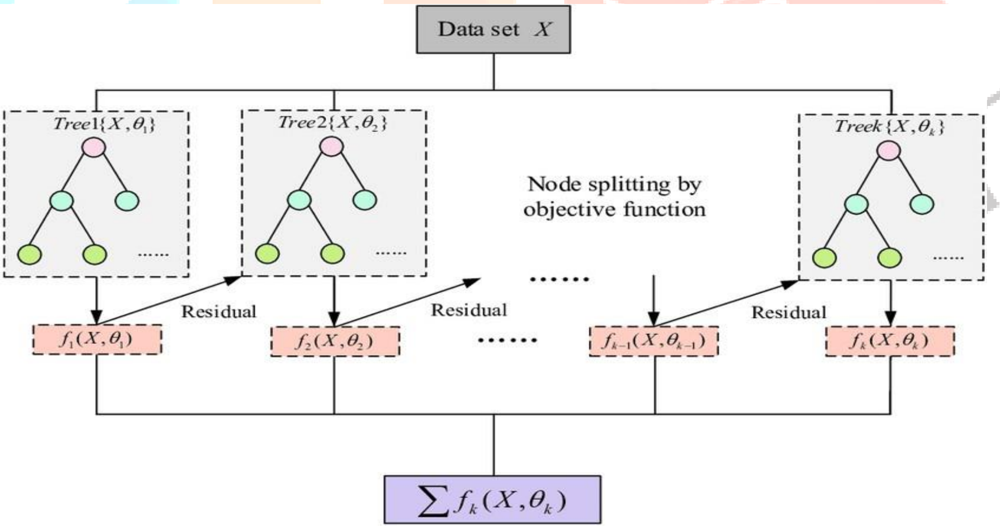


Fig-1 : Flow chart of XGBoost

- **Dataset Preparation:**

  In your code, data.apply (pd.to_numeric) converts data to numeric, and the dataset is prepared using train_test_split.

- **Hyperparameter Tuning:**

  You use Optuna to find optimal hyper parameters like n_estimators, max_depth, and learning_rate.

- **n_estimators:**

  It sets how many trees the model builds. More trees can make the model better, but too many may overfit.

- **max_depth:**

  It controls how deep each tree grows(depth limit)

- **learning_rate:**
  It decides how much each tree impacts the final result.
- **Model Training:**

  Train the model iteratively, where each tree predicts residuals of the previous tree.Residuals are carried forward, improving predictions iteratively.
- **Final Prediction:**

  The sum of all tree outputs ($\sum f_k(X,\theta_k)$ representing the final prediction.

IMPLEMENTATION

1) **Setting Up the Environment:**

- **Install Required Libraries:**

  Ensure you have Python installed.
- **Install dependencies using the requirements.txt file by running:**
  pip install -r requirements.txt
- **Organize the Data:**

  Place all provided.csv files in the appropriate directories
- **Launch Jupyter Notebook:**

  Open Jupyter Notebook for running.ipynb files:

2) **Data Exploration:**

Open EDA.ipynb to:

Load and examine Train-*.csv and Test-*.csv datasets.

Use visualizations to understand the distributions and outliers better.

Identify and note important features useful for model t raining.

Save the findings or graphs as per requirements to guide model development.

3) **Data Preprocessing Preprocessing Steps:**

Use insurance-fraud.ipynb or insurance.py for data preprocessing. Tasks include:
1.Mapping categorical values (e.g.,OperatingPhysician_mapping.csv).

2.Handling missing data.

3.Normalizing or scaling numerical data.

4.Splitting the dataset into training and testing subsets.

5.Ensure the preprocessed data is saved for model training.

4) **Model Training:**

Train Machine Learning Models:

CatBoost Model : Open catboost_model.ipynb. Load preprocessed data. Train the CatBoost model with optimization for categorical features. Other Models: Train other models like Logistic Regression and XGBoost using insurance-fraud.ipynb.

Hyper parameter Tuning:

Experiment with parameters like learning rate, depth, and number of estimators. Apply cross-validation to improve accuracy and reduce overfitting.

Save Models:

Save trained models in .joblib files (e.g., catboost_model.joblib, xgboost.joblib).

5) **Model Evaluation :**

Evaluate Performance:
Metrics to be used: accuracy, precision, recall, F1 score, and AUC- ROC.

Visualization of Results:

Use plots for better visualization: confusion matrix and ROC curve

Model Comparison:

Compare the results and select the model that is most appropriate for fraud detection.

6) **Fraud Prediction:**

1. Use Saved Models:

Load models like catboost_model.joblib in insurance.py. Provide new claim datasets for prediction (e.g., Test-*.csv).

2. Output Predictions:

Generate predictions indicating fraud likelihood or classification.Save results to a file or display them.

7) **Deployment:**

1. API Integration:

Package the trained model in an API to make fraud detection real time.Use Flask or FastAPI frameworks to develop the API.

2. System Integration**:**

Integrate API into the insurance claim processing system.Set up monitoring for prediction accuracy and model performance.

ARCHITECTURE

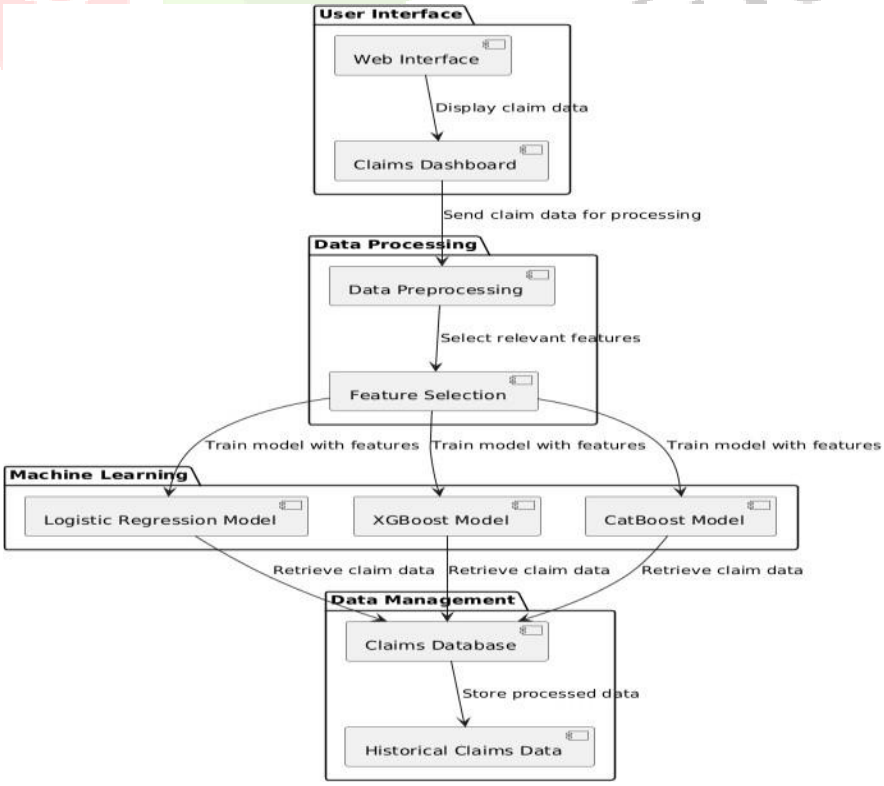


Fig – 2: Architecture of the project(Work Flow)

WORK FLOW:

1) **Data Preprocessing:** The data is cleaned and transformed into the proper form for training the model; that is, handling missing values, encoding categorical variables, and dropping unnecessary columns.

Cleaning Data: Fill missing values with 0 using df.fillna(0).

Encoding Categorical Variables: The categorical variables Provider, Physician, and Diagnosis Codes are encoded with numerical values, which are preserved in CSV files for mappings.

Dropping Unnecessary Columns: Drops column(s) which are not in use for the prediction model, such as df.drop().

2) **Feature Selection:**

Features relating to the amount and frequency of a claim are selected, among others, in preparation for analysis by machine learning.

3) **Model Training:**

The data gets trained by some machine learning models like Logistic Regression, XGBoost, and CatBoost.

Logistic Regression : The model was chosen as a baseline model because of its simplicity in understanding and interpreting the model.

XGBoost and CatBoost: These two models are powerful predictors in classification problems. XGBoost is much more efficient and scalable; on the other hand, CatBoost works better with categorical features.

4) **Model Testing:**

The trained model was tested on the dataset to check the accuracy and prediction power of fraudulent claims.

5) **Decision:**

If the model performance is satisfactory, apply the model to new claims for fraud prediction.

If model performance is not satisfactory, optimize the model and retrain it.

XGBoost model has the highest accuracy and outperformed all other models in predicting fraudulent claims.

6)**Deployment:**

Also developed is a Streamlit API for the deployment of the trained fraud detection model. It takes in relevant user-input claim data and gives out the prediction, showing the probability that fraud may exist.

The interface is user-friendly; hence, insurance investigators can interact with it easily. The results are also very clear, and thus the user can assess if a claim has the potential of being fraudulent.

7) **Result :**

It Notifies fraud analysts whenever a claim is marked as being potentially fraudulent in red color and non fraudulent in green color.

Upon accessing the Streamlit API, users are prompted to input relevant claim data. This typically includes information such as claim amount, patient details, diagnosis codes, physician information, and other key claim attributes. These inputs are necessary for the model to evaluate the likelihood of fraud. The user-friendly interface ensures that users can easily enter the required data, and the model processes this information to return a fraud prediction. The web application provides the user interface for generating the application user to detect the whether the applicant is fraud or not by providing the required data present on the screen .

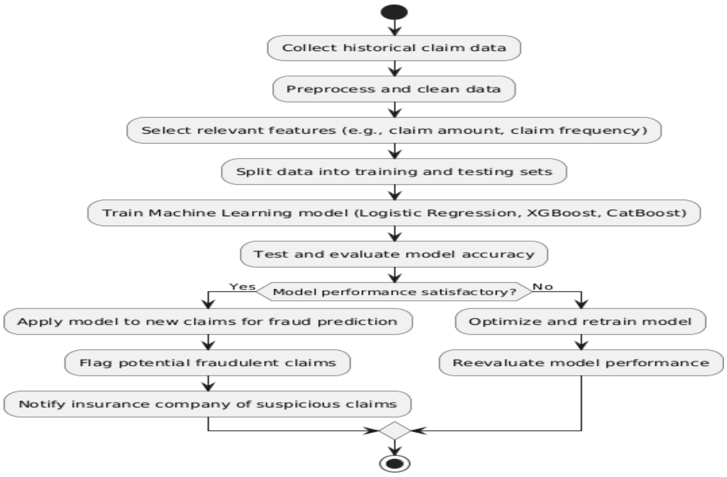
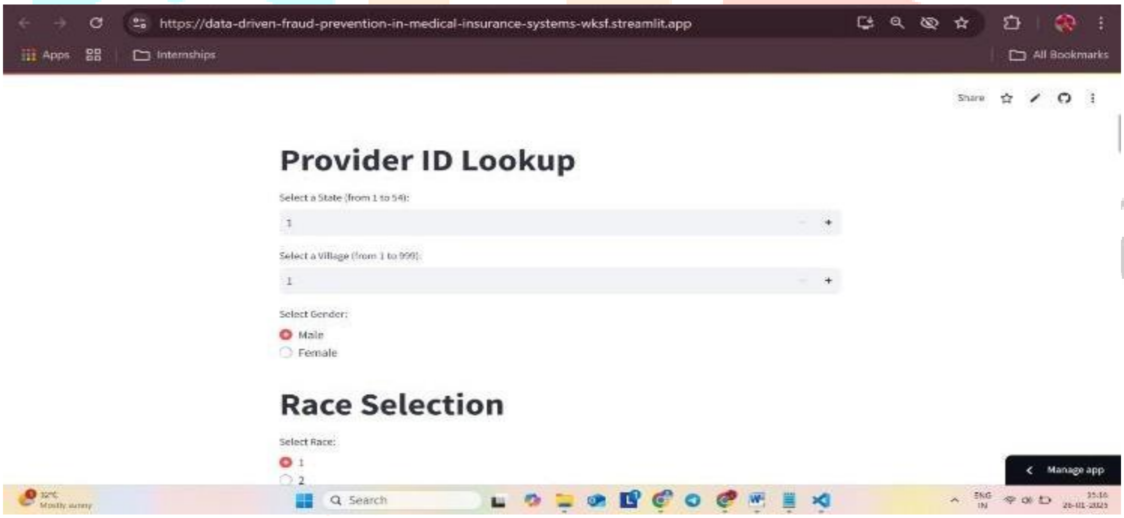

Collect historical claim data
Preprocess and clean data
Select relevant features (e.g., claim amount, claim frequency)
Split data into training and testing sets
Train Machine Learning model (Logistic Regression, XGBoost, CatBoost)
Test and evaluate model accuracy
Yes
Model performance satisfactory?
No
Apply model to new claims for fraud prediction
Optimize and retrain model
Flag potential fraudulent claims
Reevaluate model performance
Notify insurance company of suspicious claims

Fig-3: Flow of execution

Prediction Model:

The prediction model employs Logistic Regression, XGBoost, and CatBoost to detect fraudulent insurance claims by analyzing claim data like amounts, frequency, demographics, and diagnosis codes. It adapts to new patterns, ensuring accurate, real-time fraud detection and reducing financial losses for insurers.



Provider ID Lookup

Select a State (from 1 to 54):

1

Select a Village (from 1 to 999):

1

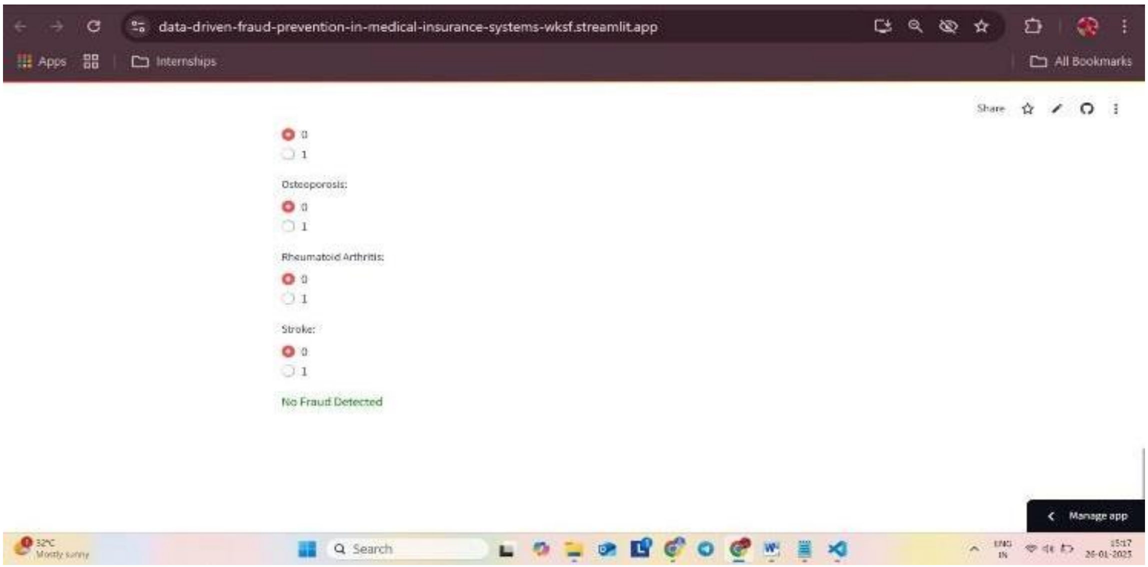
Select Gender:

● Male
○ Female

Race Selection

Select Race:

● 1
○ 2



● 0
○ 1

Osteoporosis:

● 0
○ 1

Rheumatoid Arthritis:

● 0
○ 1

Stroke:

● 0
○ 1

No Fraud Detected

FUTURE SCOPE

There is immense room for expansion in and enhancement of a machine-learning-based fraud-detection system on the claims by medical insurance; there is actually plenty of technological advancement and an increase in availability. Hereafter, key fields that can enable a huge influence and enhance future capabilities are covered:

1. Advanced Machine Learning and AI Techniques

- Deep Learning: It can use those complex neural network architectures to keep improving the capability of detecting subtle and sophisticated fraud patterns.
- Natural Language Processing: applying NLP techniques to text data coming from claims and medical records could give much richer insights into data and detect fraud based on anomalies in text descriptions**.**

2.Extension to Multi-Modal Data Analysis

- Imaging Processing: In the cases of insurance claims that involve medical imaging, like X-rays or MRIs, adding image recognition capabilities might further improve fraud detection accuracy by verifying and analyzing the images.
- Audio Analysis: Analyzing recordings of calls between clients and insurance agents with speech recognition technology might help identify discrepancies or deceitful behaviours**.**

3.Better User Experience and Accessibility

- Mobile Applications: Develop the mobile apps so that the applications give quicker access and real-time alerting of investigators and insurance staff for a better response time and for an enhanced operational efficiency.
- Best-in-class reporting and Dashboard tool: Best-quality, more usable visualization tools shall be developed; customized reporting to bring insight towards the end users in the best form, including all organizational levels.

4.More Security and Compliance

- Compliance Tools: The development of tools that automatically ensure adherence to new regulations and standards eases the burden of providers and guarantees that the system is updated with the changes in laws.
- Advanced Security Features: State-of-the-art cybersecurity technologies can be deployed to assure sensitive data and trust in the integrity of the system.

CONCLUSION

A Machine Learning-Based Fraud Detection System for medical insurance claims would use an algorithm like Logistic Regression, XGBoost, or CatBoost in analyzing historical data to make predictions on fraudulent claims. It would be more precise and efficient to consider such characteristics of the claims as their frequency and amount, as well as demographics and diagnosis codes

The main advantages include automated fraud detection, decreased operational costs, real-time analysis, and adaptability to evolving fraud techniques. The system reduces financial losses, increases policyholder confidence, and preserves the integrity of the insurance industry—a giant leap in fighting fraud

REFERENCES

[1] S. G. Patel, P. R. Sharma, A robust fraud detection model for medical insurance claims using XGBoost, Journal of Healthcare Informatics Research, 2021, 25(3), 212–220.

[2] R. S. Kumar, P. Yadav, A. Verma, Machine learning-based detection of fraudulent medical claims using logistic regression, Journal of Artificial Intelligence in Medicine, 2020, 32(5), 110–120.

[3] M. Sharma, V. R. Gupta, Fraudulent claim detection using machine learning and statistical models, in

Proceedings of the International Conference on Data Science and Machine Learning, 2021, 94–102.

[4] A. Singh, R. Ghosh, P. Mehta, Medical fraud detection using CatBoost classifier: A comparative study, Journal of Healthcare Technology, 2021, 8(4), 134–142.

[5] B. K. Thakur, R. P. Kumar, Fraud detection in insurance claims using data mining techniques, in Proc. of the International Conference on Business Analytics, 2020, 60–70.

[6] L. P. Joshi, M. V. Pathak, Machine learning models for fraud detection in health insurance claims, International Journal of Machine Learning, 2022, 14(2), 75–85.

[7] A. V. Raj, R. P. K. Mishra, Fraudulent claim detection using advanced data mining techniques, International Journal of Applied Computer Science, 2021, 7(3), 101–110.

[8] N. S. Arora, P. R. Yadav, XGBoost and logistic regression for detecting medical insurance fraud, in Proc. of the International Conference on Artificial Intelligence in Healthcare, 2020, 122–128.

[9] T. G. Kumar, R. K. Rao, A hybrid approach for fraud detection in health insurance using machine learning, International Journal of Computational Intelligence, 2021, 9(2), 91–100.

[10] M. J. Bhatia, S. A. Desai, Fraudulent claims detection system for health insurance using data mining algorithms, Journal of Data Science & Analytics, 2021, 5(3), 159–168.

[11] M. K. Mishra, P. Kumar, Fraud detection in insurance claim management using decision trees and random forests, in Proc. of the International Symposium on Data Engineering, 2020, 97–105.

[12] S. S. Pandey, A. R. Kumar, Machine learning techniques for fraud detection in health insurance claims, Journal of Financial Engineering and Risk Management, 2022, 10(1), 48–57.

[13] L. R. Gupta, R. Singh, Prediction of fraudulent activities in insurance using CatBoost algorithm, Journal of Business Analytics, 2021, 17(4), 220–229.

[14] V. N. Tripathi, R. P. Yadav, Identifying fraudulent claims in health insurance with advanced machine learning models, International Journal of Financial Risk, 2022, 15(6), 213–220.

[15] K. M. Singh, J. A. Sharma, Fraud detection in medical insurance using feature selection techniques, Journal of Artificial Intelligence & Insurance, 2020, 9(3), 72–80.

[16] A. Kumar, M. S. Bansal, Advanced fraud detection techniques for health insurance claims using neural networks, International Journal of Fraud Prevention, 2021, 19(2), 80–90.

[17] S. R. Mehta, V. N. Verma, Identifying fraud in medical insurance claims using predictive modeling, International Journal of Machine Learning in Healthcare, 2021, 5(2), 50–60.

[18] P. Y. Sharma, M. V. Bhatia, Detection of fraud in health insurance using random forests and decision trees, IEEE Trans. on Machine Learning and Healthcare, 2020, 10(7), 1500–1510