International Journal of Core Engineering & Management
Volume-7, Issue-10, 2024 ISSN No: 2348-9510

EXPLORING THE PERFORMANCE OF MACHINE LEARNING MODELS FOR CLASSIFICATION AND IDENTIFICATION OF FRAUDULENT INSURANCE CLAIMS

Rajesh Goyal
FS - Insurance
IBM - USA
Glen Mill
Rajesh.nim@gmail.com

Abstract

Commercial insurers have seen an increase in fraud cases involving all kinds of claims in recent years. Fraud claims have increased significantly, which might cause serious problems. As a result, many governmental and private organizations, as well as the government, work hard to identify and halt fraudulent activities. False accident claims are a common way for the car industry to file fake insurance claims, which is one of the most common types. The goal of this research is to create a ML model that can distinguish between real and fraudulent insurance claims using datasets of claims. With an emphasis on DT, XGBoost (XGB), LR, and SVM, this research investigates how well different ML models perform in the classification and detection of fraudulent insurance claims. The XGB model demonstrated superior performance with an accuracy of 89%, excelling in both precision 91 and F1-score 89.1, indicating a well-balanced ability to identify true positives while minimizing false positives. Overall, XGB emerged as the most effective model, effectively combining high accuracy with strong precision and recall metrics. Research in the future should look into how to apply real-time fraud detection systems in real-world scenarios to boost operational efficiency in the insurance sector, and how to combine ensemble methods with advanced feature engineering techniques to make models even better.

Keywords: Fraud Detection, Insurance Fraud, machine learning, insurance claim dataset, XGBoost.

I. INTRODUCTION

Everyone needs inexpensive healthcare since it is a need. A great deal of complexity and interdependence characterize the healthcare business. The rate of expansion is rather rapid. Meanwhile, fraud is becoming an increasingly serious issue in this sector. Misuse of health insurance programs is one problem [1]. Fraud is a major issue for insurance companies. Preventing fraudulent actions is of utmost importance since defending against them could be challenging, especially in insurance companies. While just 3% of suspected fraudulent auto insurance claims end up in court, estimates place the number of such cases anywhere from 21% to 36%[2]. The first step in preventing fraudulent claims is identifying them, which is challenging and not "cost-effective" either since the arduous and drawn-out investigations might enrage the real customers [3].

It becomes more difficult to uncover fraud scenarios due to higher investigation expenditures. As a result, corporations proceed without conducting adequate investigations, which leads to many potential difficulties. Manual fraud detection is no longer cost-effective or efficient; we must examine the scam before paying the claim. Several ML and data mining approaches have shown to be effective in identifying fraud.

A branch of AI, Machine Learning (ML) primarily seeks to imitate human intellect. Building models with strong prediction ability is the main emphasis of ML [4][5]. Several unresolved issues in a variety of sectors, including banking, have been addressed with the use of operational research techniques that have lately come into prominence, including data mining, ML, and DL. The interesting thing is that these new advances have contributed to the proposal of new fraud detection systems.

1. Contribution of study

The thorough assessment of several ML models presented in this paper significantly advances the state of the art in insurance claim fraud detection. The following points summarize the key contributions:

- To utilize insurance claim dataset for model implementation.
- Addresses missing values, outliers, and class imbalance to enhance dataset quality.
- Compares multiple algorithms (DT, XGBoost, LR, SVM) for effectiveness in fraud detection.
- Assesses models using F1-score, recall, accuracy, and precision for comprehensive evaluation.
- Insurance firms can increase the reliability of fraud detection with the actionable information provided.

2. Structure of paper

Here is the outline of the rest of the paper: Some prior research on the classification and identification of Fraudulent Insurance Claims is detailed in Section II. Section III details the research process, and Section IV presents and analyses the experimental data. Lastly, Section V presents the findings.

II. LITERATURE REVIEW

In this provide the previous work on the classification and identification of Fraudulent Insurance Claims using machine and DL techniques.

This research, Garmdareh et al., (2023) proposes an innovative methodology that leverages ML algorithms based on regression to predict the overall cost of a patient's claim using their past claims data. The estimated and real amounts are then compared to determine the patient's price difference. A criterion for the absolute price difference will be used to estimate the anomalous or fraud expenses in claims. For assessment, a dataset including 99,440 records from the RASA online site is compiled. When testing, decision trees have the best MAE, but when training, deep learning is the best. Therefore, anomaly identification makes use of the decision tree, which can identify about 17% of data as aberrant with a 30% variance or more. Better than half of the reported abnormalities are approved when human experts review and accept the findings [6].

This study, Nabrawi and Alanazi, (2023) create a computational model for healthcare that can identify instances of health insurance claim fraud in Saudi Arabia. With perfect precision, the model pinpoints the most important element that leads to fraud. Three supervised DL and ML algorithms were used to the labelled imbalanced dataset. Three Saudi healthcare providers contributed to the dataset. ANNs, LR, and RF were the models that were used. They balanced the dataset using the SMOT approach. For the purpose of excluding irrelevant characteristics, Boruta object feature selection was used. F1 score, specificity, accuracy, precision, recall, and AUC were the measures used for validation. With a 98.21% accuracy rate, a 98.08% precision rate, a 100% recall rate, a 99.03% F1 score, an 80% specificity rate, and a 90.00% AUC, random forest classifiers identified age, education level, and policy type as the most important characteristics. An F1 score of 88.17%, specificity of 80%, recall of 80.39%, accuracy of 80.36 percent, precision of 97.62 percent, and AUC of 80.20% were the outcomes of the logistic regression analysis. Results from ANN showed an F1 score of 97.03%, a specificity of 80%, an AUC of 88.04%, a recall of 96.08%, and an accuracy of 94.64%. Additional research on a bigger dataset is recommended, since all three models used in this predictive analytics study produced satisfactory accuracy and validation metrics [7].

In, Fursov et al., (2022) suggested architectures for DL that can analyze insurance data, which is comprised of consecutive recordings of patients' visits and attributes. The sequential and tabular components both provide fresh perspectives on health insurance fraud detection, enhancing the model's quality. The claims management process may be greatly enhanced by our method, as shown by empirical findings obtained using pertinent data from a health insurance company, which surpass advance models. The top competitor using state-of-the-art models gets a ROC AUC value of 0.815, whereas we get 0.873. We further show that our designs are more resistant to corrupted data. Insurers are expected to have access to an increasing amount of semi-structured event sequence data in the near future. Our methods will prove useful in numerous similar applications, especially for variables with multiple categories, like those found in the ICD codes or other classification systems [8].

This research, Nur Prasasti, Dhini and Laoh, (2020) a ML-based prediction model for the detection of vehicle insurance fraud. The research was based on actual information gathered from an Indonesian motor insurer. A large number of fraudulent policyholders' records are skewed towards the other side of the dataset. This study addresses the issue of unbalanced datasets by using under sampling and the SMOTE. The supervised classifiers that have been suggested include RF, DT, and MLP. The confusion matrix, ROC curve, and metrics like sensitivity are used to assess the efficacy of models. With an accuracy rate of 98.5%, Random Forest fared better than the other classifiers tested in this study[9] develop.

In this work, Gupta et al., (2019) develop a predictive model was developed using the GBM and then applied to data from vehicle insurance claims. They used the SMOTE to fix the dataset's extreme imbalance. The outcomes were outstanding, with an accuracy rate of 99% and an F1 score of almost 98%. The actuarial model known as extreme value theory (EVT) was used for cross-validation by professionals in the field. It is possible to modify, test, and expand the predictive model offered in this article to other business domains [10].

The following Table 1 presents the background study on the identification of Fraudulent Insurance Claims using machine and deep learning techniques based on its performance.

International Journal of Core Engineering & Management
Volume-7, Issue-10, 2024 ISSN No: 2348-9510

Table 1: Summary of related work on identification of Fraudulent Insurance Claims using ML and DL methods

Author	Dataset	Methodology	Performance	Limitation/contribution
Garmdareh et al. (2023) [6]	99,440 records from the RASA web portal	Regression-based models, Decision Tree for anomaly detection, Deep Learning for training	Decision Tree detected 17% of claims as abnormal with at least 30% deviation, deep learning had best training MAE	• **Contribution:** Decision Tree effective for anomaly detection. • Limitation: Only 50% of anomalies approved by expert assessors.
Nabrawi & Alanazi (2023) [7]	Health insurance data from 3 providers in Saudi Arabia	RF, LR, ANN; SMOTE for balancing; Boruta for feature selection	Random Forest: 98.21% accuracy, 100% recall, 90% AUC; ANN: 94.64% accuracy, 96.08% recall, 88.04% AUC	• **Contribution:** Identified key features (policy type, education, age) for fraud detection. • **Limitation:** Further research on larger datasets recommended.
Fursov et al. (2022) [8]	Health insurance data from an insurance company	Deep learning architectures combining sequential and tabular components	Achieved ROC AUC of 0.873, outperforming state-of-the-art models (0.815)	• **Contribution:** Improved fraud detection and robustness to data corruption. • **Limitation:** Focused primarily on sequential/tabular data integration.
Nur Prasasti et al. (2020) [9]	Real-world automobile insurance data (Indonesia)	Supervised classifiers: MLP, DT, RF; SMOTE, under sampling for balancing	Random Forest: 98.5% accuracy	• **Contribution:** Addressed imbalanced data in fraud detection. • **Limitation:** Lack of comparison with more recent models or DL techniques.
Gupta et al. (2019) [10]	Motor insurance claims data	GBM, SMOTE for imbalanced data, extreme value theory (EVT) for cross-validation	Achieved around 98% F1 score, 99% accuracy	• **Contribution:** High model accuracy and effective fraud detection. • **Limitation:** Model is highly specific to insurance claims and may need customization for other applications.

1. Research gaps

A number of knowledge gaps persist in insurance fraud detection using ML and DL models, despite substantial progress in this area.

- **Domain Performance:** Need to analyze the generalizability of fraud detection methods across various insurance industries.
- **Imbalanced Data:** Explore robust methods beyond SMOTE to handle imbalanced datasets without overfitting.
- **Explain ability:** Develop interpretable models for better stakeholder understanding of fraud detection decisions.
- **Real-Time Detection:** Focus on proactive, real-time fraud detection, especially in health insurance.

III. METHODS AND MATERIALS

The methodology for exploring a performance of ML models in the classification and identification of fraudulent insurance claims involved several systematic steps. Initially, a dataset comprising nearly 1,000 rows and 39 columns was sourced from Kaggle, focusing on various criteria relevant to insurance claims. Data preprocessing included filling missing values with mean or median imputation, removing outliers to enhance data quality, and employing SMOTE for data balancing to address class imbalance issues. Min-max normalization was applied to scale features between 0 and 1, ensuring uniformity in data representation. Feature extraction techniques were employed to

International Journal of Core Engineering & Management
Volume-7, Issue-10, 2024 ISSN No: 2348-9510

transform raw data into numerical features suitable for ML algorithms. The dataset was then divided into 80% for training and 20% for testing. Various classification algorithms, including Decision Trees (DT), XGBoost (XGB), LR, and SVM, were used to build models for identifying fraudulent claims. A confusion matrix, together with five essential metrics—accuracy, precision, recall, and F1-score—was used to assess the efficacy of each model in identifying fraud. The process of research design entails various steps and phases that are displayed in the figure 1 flowchart.

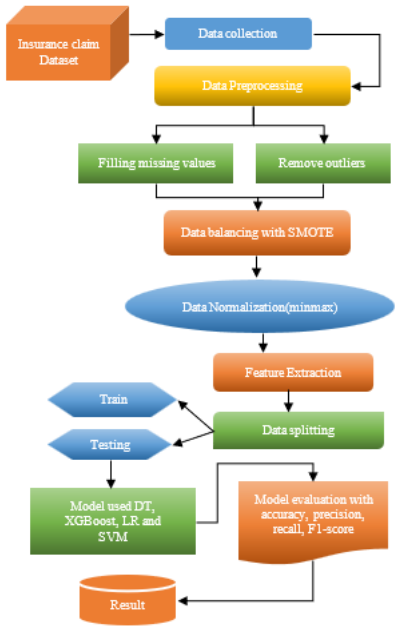

Figure 1: Flowchart for Fraud Insurance Claim

1. Data gathering

The dataset selected for this specific task was extracted from an online source named Kaggle, which contains nearly 1000 rows of historical data and 39 columns containing various criteria for the project. For the identifying fraud insurance claim, use the insurance claim dataset. The correlation matrix of data variables is visualized in below:

The following figure 2 displays the correlation heatmap that represents relationships between variables, with dark purple for strong negative correlations, light orange for strong positive correlations, and neutral colors for weak or no correlations. It helps identify patterns in the dataset.

2. Data preprocessing

The first step in using the approach is data preparation. For this procedure, an appropriate dataset was selected from Kaggle. It defines the steps used in ML to transform raw data that is incoherent into data that can be used to train models [11]. The further pre-processing steps are as follows:

- Filling missing values: Replacing missing values with the mean or median of the column can preserve overall data distribution, especially for numerical data.
- Remove outliers: Removing outliers is another essential data preprocessing technique that enhances the quality of the dataset. Outliers can skew results and lead to misleading interpretations.

3. Data balancing with SMOTE

SMOTE has remarkable performance on tiny datasets. On the other hand, SMOTE becomes far less efficient and takes longer to generate false data points in huge datasets. Furthermore, while creating erroneous data points, there is a high probability of overlapping data points for the SMOTE minority class.

4. Data normalization (minmax)

An essential first step in preparing data for ML is data normalization. Min-max normalization, which adjusts input variables to an average with a range that only includes ones and zeros, is one of the most used normalization techniques. The lowest and maximum values of each variable are found, and the values are rescaled to fall between 0 and 1, in order to carry out min-max normalization. The following formula (1)

$$x = \frac{x - \min(x)}{\max(x) - \min(x)} \quad (1)$$

where: xscaled = scaled sample point

x = sample point

5. Feature Extraction

The term "feature extraction" describes the steps used to convert unstructured data into processable numerical features while keeping the original data intact. It outperforms just feeding the raw data into the ML algorithm.

6. Data splitting

The claim dataset was split into two parts: 20% for testing the classifier prediction and 80% for training the machine learning models.

7. Classification algorithms

For the model building of identifying fraud insurance claim use various ML models [12] like DT, XGBoost, LR, and SVM as explained in below:

A. Decision tree (DT)

Classification and regression are two applications of decision trees, which are machine learning algorithms. It builds a decision-tree model iteratively by splitting data into subsets according to a selected characteristic and continuing until a stopping requirement is satisfied. A choice based on a feature is represented by each node, and each branch represents a potential conclusion. Both numerical and categorical data are easily handled by decision trees. However, when training data is significantly changed, it may overfit and provide surprising results.

B. XGBoost

An ensemble tree approach that combines weak learners repeatedly is the XGBoost algorithm [13]. The approach employs boosting techniques to decrease the residual size and fits a diverse range of trees to the pseudo residuals, which are the real value minus the projected value. This lessens the likelihood of overfitting and ultimately leads to a classification model that performs better. A weak learner is trained to rely on the direction of the loss function's gradient in a repeated process called

International Journal of Core Engineering & Management
Volume-7, Issue-10, 2024 ISSN No: 2348-9510

"boosting," which ultimately results in a stronger classifier. The model produces projected values via (2) after fitting all trees

$$\hat{y}_i = \sum_{k=1}^{k} f_k(x_i) \quad (2)$$

If the feature vector for the ith data point is represented by xi, and fk is a classification tree k. For binary classification, the technique makes use of the LogLoss (see Equation 3 above). A regularization term regulates the model's complexity to keep it from becoming too complicated and to avoid overfitting. The XGBoost method employs the following regularization term:

$$\Omega = \gamma L + \frac{1}{2}\lambda \sum_{j=1}^{L} w_j^2 \quad (3)$$

Where L is a number of leaves, wj 2 is the score on the jth leaf, which may be transformed into probabilities using the sigmoid function, and γ and λ are the degrees of regularization. A model's goal function is produced by the regularization function and loss function combined (4):

$$obj^{(t)} = \sum_{i=1}^{n}(y_i, \hat{y}_i^{(t-1)} + f_t(x_i)) + \Omega(f_t) \quad (4)$$

Where L is the loss function and Ω is the regularization term.

C. Logistic Regression (LR)

Binary classification is accomplished by an employ of the ML method logistic regression. Sigmoid functions are used to describe the chance that an input belongs to a specific class by applying them to a linear combination of the input attributes. The coefficients that minimize the log-loss function are learnt during training. Handling both numerical and categorical input characteristics, logistic regression is easy to use and effective.

D. Support Vector Machine (SVM)

Regression and classification problems may be handled with the machine learning method Support Vector Machine (SVM). In order to forecast target values for regression analysis or to effectively divide classes, it finds the optimal hyperplane. SVM employs a kernel function to transform the input data into higher dimensions in order to maximize the margin among classes or minimize the error for regression. It then optimizes a cost function. It can handle data with several dimensions and is resistant to overfitting.

8. Performance matrix

We used a suite of assessment criteria, sometimes called performance metrics, to assess how well our phishing email detection system worked. The confusion matrix is a table that shows the difference between the actual values and the values predicted by the model. Its purpose is to evaluate how well the model functions. F1-score, recall, accuracy, and precision were the five assessment metrics used to analyze the final models. To start, models are evaluated using confusion matrices according to the following: true positive (TP), false positive (FP), true negative (TN), and false negative (FN).

- An existence of any unwanted occurrence is represented by the number of records that were appropriately categorized by TP.
- The existence of any undesired occurrence is shown by the amount of records that were wrongly categorized by FP.
- The number of records that were successfully categorized is shown as typical in TN.

iJCEM
International Journal of Core Engineering & Management

International Journal of Core Engineering & Management
Volume-7, Issue-10, 2024 ISSN No: 2348-9510

- The number of records that were misclassified is shown as usual in FN.

Accuracy: A ratio of the number of accurate classifications of normal and undesirable occurrences to a total number of events in a dataset pertaining to oil wells is the accuracy. The expression is (5):

$$Accuracy = \frac{\mathrm{TN + TP}}{\mathrm{TP + TN + FP + FN}} \quad (5)$$

Precision: The word "precision" is defined as the percentage of unwanted well events that are correctly detected relative to the total number of unwanted well events. Equation (6) represents it:

$$\mathrm{Precision} = \frac{\mathrm{TP}}{\mathrm{TP+FP}} \quad (6)$$

Recall: Re The recall measures how many unwanted oil well events were properly categorized relative to a total number of unwanted events in a dataset. In representation, it is (7):

$$\mathrm{Recall} = \frac{\mathrm{TP}}{\mathrm{TP+FN}} \quad (7)$$

F1-score: The F1-score, which is formally written as (8), is the weighted average of the recall and precision.

$$\mathrm{F1} = \frac{2*(\mathrm{precision}*\mathrm{recall})}{\mathrm{precision}+\mathrm{recall}} \quad (8)$$

These matrices are used for comparative analysis of machine learning models on insurance claim dataset.

IV. RESULT ANALYSIS AND DISCUSSION

This section describes the experimental data obtained after creating the model for detecting insurance claim fraud. By comparing it to baseline models, we assess how well the intelligent ML technique performs. The following Table 2 provide the XGBoost model efficiency on the insurance claim dataset across the performance matrix.

Table 2: XGBoost model performance for fraud insurance claim on insurance claim dataset

Matrix	**XGBoost**
Accuracy	89
Precision	91
Recall	89
F1-score	89.1

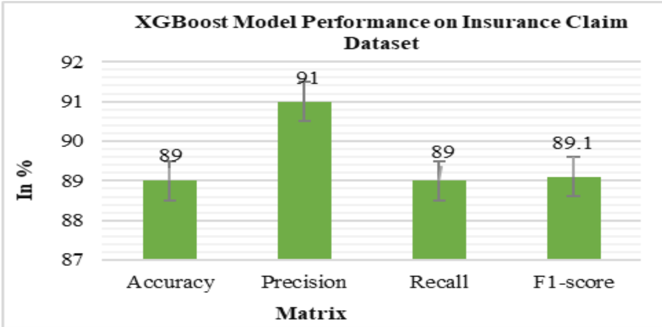

Figure 3: Bar Graph for XGBoost model

The above Table 2 and Figure 3 displayed a XGBoost model performance. A XGBoost model achieved 89% accuracy, 91% precision, 89% recall, and an 89.1% F1-score, indicating strong overall performance in classification.

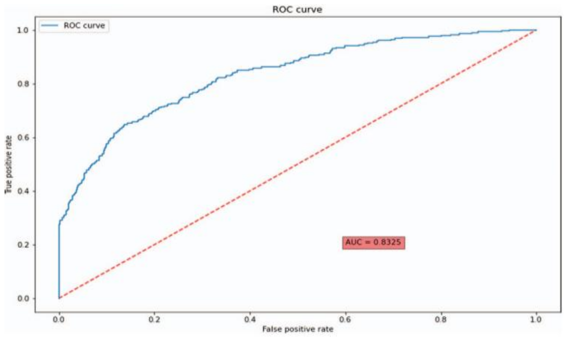

Figure 4. ROC curve for XGBoost model

Figure 4 shows the XGBoost model's ROC curve, which shows the TP/FP trade-off at various levels. The blue line represents the ROC curve, with an AUC of 0.8325, indicating that the model has an 83.25% chance of correctly distinguishing between positive and negative classes, reflecting good performance.

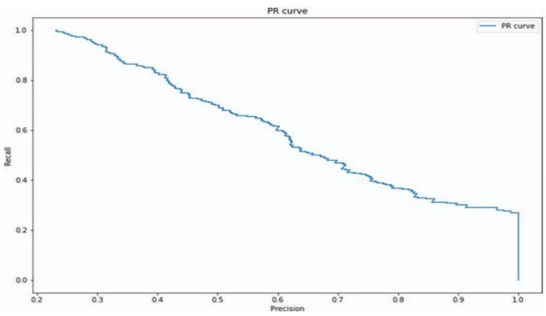

Figure 5. PR curve for XGBoost model

The image shows a Precision-Recall (PR) curve for an XGBoost model, illustrating the trade-off between precision (the ratio of TP to total forecasted positives) and recall (the ratio of TP to actual positives) at various thresholds. The horizontal axis represents precision, and the vertical axis represents recall. The curve typically starts with high precision and recall at the top-left and moves

International Journal of Core Engineering & Management
Volume-7, Issue-10, 2024 ISSN No: 2348-9510

towards lower values at the bottom-right, reflecting the model's performance across different decision thresholds.

1. Comparative Analysis

This section presents a comparison study of insurance fraud claims that used ML methods. The comparison between ML models like DT [14], LR[15], XGBoost, and SVM [16] on the insurance claim dataset.

Table 3: comparative analysis for fraud insurance claim detection

Models	Accuracy	Precision	Recall	F1-score
DT	63.6	65	56	60
LR	72	65	88	76
SVM	86.50	67	86	75
XGB	**89**	**91**	**89**	**89.1**

Table 3 provides the comparative analysis between model performance. In comparing SVM achieved the highest accuracy at 86.50%, though its precision of 67 and F1-score of 75 were relatively moderate. The XGB model was closely followed with an accuracy of 89%, excelling in both precision 91 and F1-score 89.1, indicating a balanced performance in identifying TP while minimizing false positives. The LR model demonstrated a strong recall 88 but had lower overall accuracy 72 and precision 65. In contrast, the DT model performed the least effectively, with the lowest accuracy of 63.6 and an F1-score of 60, highlighting its struggles in both precision and recall. Overall, XGB emerged as the most effective model, combining high accuracy with strong precision and recall metrics.

V. CONCLUSION AND FUTURE SCOPE

The insurance industry confronts the significant problem of insurance claim fraud. It costs insurance companies money and raises policyholders' rates. Machine learning has become a potential method for insurance claim fraud investigation and detection in recent years. ML algorithms—like data mining and DL techniques—have been effectively applied to identify trends and abnormalities in insurance claim data that point to fraudulent activity. This study's overarching goal is to lessen the impact of class imbalances in ML by using preprocessing sampling approaches to more evenly distribute datasets. In this study, a performance of various ML models—Decision Trees (DT), XGB, LR, and SVM—was evaluated for the classification and identification of fraudulent insurance claims. The results revealed that XGBoost achieved the highest accuracy at 89%, with precision at 91% and an F1-score of 89.1%, demonstrating its effectiveness in fraud detection. Despite these promising results, the study has limitations. Although the dataset is pertinent, its small size could restrict how broadly the results can be applied. The efficacy and dependability of fraud detection methods will be greatly enhanced by this continuing study.

REFERENCES

1. S. S. Waghade and A. M. Karandikar, "A Comprehensive Study of Healthcare Fraud Detection based on Machine Learning," Int. J. Appl. Eng. Res., vol. 13, no. 6, pp. 4175–4178, 2018.
2. K. Nian, H. Zhang, A. Tayal, T. Coleman, and Y. Li, "Auto insurance fraud detection using unsupervised spectral ranking for anomaly," J. Financ. Data Sci., 2016, doi: 10.1016/j.jfds.2016.03.001.

3. M. Kirlidog and C. Asuk, “A Fraud Detection Approach with Data Mining in Health Insurance,” Procedia - Soc. Behav. Sci., 2012, doi: 10.1016/j.sbspro.2012.09.168.
4. S. K. R. Anumandla, V. K. Yarlagadda, S. C. R. Vennapusa, and K. R. V Kothapalli, “Unveiling the Influence of Artificial Intelligence on Resource Management and Sustainable Development: A Comprehensive Investigation,” Technol. \& Manag. Rev., vol. 5, no. 1, pp. 45–65, 2020.
5. L. Settipalli and G. R. Gangadharan, “WMTDBC: An unsupervised multivariate analysis model for fraud detection in health insurance claims,” Expert Syst. Appl., 2023, doi: 10.1016/j.eswa.2022.119259.
6. M. S. Garmdareh, B. S. Neysiani, M. Z. Nogorani, and M. Bahramizadegan, “A Machine Learning-based Approach for Medical Insurance Anomaly Detection by Predicting Indirect Outpatients’ Claim Price,” in 2023 9th International Conference on Web Research, ICWR 2023, 2023. doi: 10.1109/ICWR57742.2023.10139290.
7. E. Nabrawi and A. Alanazi, “Fraud Detection in Healthcare Insurance Claims Using Machine Learning,” Risks, 2023, doi: 10.3390/risks11090160.
8. I. Fursov et al., “Sequence Embeddings Help Detect Insurance Fraud,” IEEE Access, 2022, doi: 10.1109/ACCESS.2022.3149480.
9. I. M. Nur Prasasti, A. Dhini, and E. Laoh, “Automobile Insurance Fraud Detection using Supervised Classifiers,” in 2020 International Workshop on Big Data and Information Security, IWBIS 2020, 2020. doi: 10.1109/IWBIS50925.2020.9255426.
10. R. Y. Gupta, S. Sai Mudigonda, P. K. Kandala, and P. K. Baruah, “Implementation of a Predictive Model for Fraud Detection in Motor Insurance using Gradient Boosting Method and Validation with Actuarial Models,” in 2019 International Conference on Clean Energy and Energy Efficient Electronics Circuit for Sustainable Development, INCCES 2019, 2019. doi: 10.1109/INCCES47820.2019.9167733.
11. A. Rahman, “Statistics-based data preprocessing methods and machine learning algorithms for big data analysis,” Int. J. Artif. Intell., 2019.
12. R. Tandon, “The Machine Learning Based Regression Models Analysis For House Price Prediction,” Int. J. Res. Anal. Rev., vol. 11, no. 3, pp. 296–305, 2024.
13. M. Chen, Q. Liu, S. Chen, Y. Liu, C. H. Zhang, and R. Liu, “XGBoost-Based Algorithm Interpretation and Application on Post-Fault Transient Stability Status Prediction of Power System,” IEEE Access, 2019, doi: 10.1109/ACCESS.2019.2893448.
14. B. Ambrose Muchangi Njeru and E. A. Miriti, “Detection of Fraudulent Vehicle Insurance Claims Using Machine Learning,” no. November, 2022.
15. S. Agarwal, “An Intelligent Machine Learning Approach for Fraud Detection in Medical Claim Insurance: A Comprehensive Study,” Sch. J. Eng. Technol., 2023, doi: 10.36347/sjet.2023.v11i09.003.
16. A. Urunkar, A. Khot, R. Bhat, and N. Mudegol, “Fraud Detection and Analysis for Insurance Claim using Machine Learning,” in 2022 IEEE International Conference on Signal Processing, Informatics, Communication and Energy Systems (SPICES), IEEE, Mar. 2022, pp. 406–411. doi: 10.1109/SPICES52834.2022.9774071.