International Journal of Artificial Intelligence in Medical Issues

Volume 4 Issue 1
ISSN 3025-4167
https://doi.org/10.56705/ijaimi.v4i1.433

Research Article

Leakage-Aware and Explainable Machine Learning for Healthcare Claim Fraud Detection Using Imbalanced Medical Insurance Data

Dian Hafidh Zulfikar [1,*]; Ery Setiyawan Jullev Atmadji [2]; Widya Wisanti [3]
[1] Universitas Islam Negeri Raden Intan Lampung, Lampung 35131, Indonesia, dianhafidhzulfikar_uin@radenintan.ac.id
[2] Politeknik Neger Jember, Kabupaten Jember, Jawa Timur 68121, Indonesia, ery@polije.ac.id
[3] Universitas Sawerigading, Kota Makassar, Sulawesi Selatan 90213, wwisanty@gmail.com

Correspondence should be addressed to Dian Hafidh Zulfikar; dianhafidhzulfikar_uin@radenintan.ac.id

Received 10 December 2025; Revised 10 January 2026; Accepted 18 April 2026; Published 30 May 2026

Abstract:
Healthcare insurance fraud is a critical challenge in health systems because fraudulent claims may cause financial losses, increase administrative burden, and reduce trust in healthcare services. This study proposes an explainable machine learning approach for detecting fraudulent healthcare insurance claims using imbalanced medical claim data. The dataset consisted of 10,000 healthcare insurance claim records with 20 attributes, including patient information, provider characteristics, claim-related financial variables, medical codes, temporal features, and fraud labels. Fraudulent claims represented only 8.29% of the dataset, indicating a clear class imbalance problem. Several machine learning models were evaluated, including Logistic Regression, Decision Tree, Random Forest, Extra Trees, and AdaBoost, under different imbalance handling strategies, namely baseline learning, class weighting, and SMOTE. In addition, two feature scenarios were compared: a full-feature scenario and a leakage-aware scenario that excluded potentially post-decision variables such as claim status and approved amount. The experimental results showed that the best full-feature model was Logistic Regression without additional imbalance handling, achieving an accuracy of 0.9900, precision of 0.9740, recall of 0.9036, F1-score of 0.9375, ROC-AUC of 0.9989, and PR-AUC of 0.9896. The model correctly detected 150 out of 166 fraudulent claims in the test set. However, the best leakage-aware model achieved a lower F1-score of 0.6983, indicating that potentially leaked variables may substantially affect model performance. Feature importance analysis showed that claim amount, approved amount, claim submission delay, claim status, and provider-related variables were among the most influential predictors. These findings demonstrate that explainable machine learning can support healthcare claim fraud detection, but careful attention must be given to class imbalance, data leakage, and operational deployment context.

Keywords: Healthcare Fraud Detection; Medical Insurance Claims; Machine Learning; Imbalanced Classification; explainable AI; Data Leakage; Health Informatics.

Dataset link: -

1. Introduction

Healthcare systems increasingly depend on large-scale digital data to support service delivery, insurance processing, clinical administration, and financial decision-making. Among these data sources, medical insurance claims represent an important component of health informatics because they contain information related to patients, providers, diagnoses, procedures, insurance types, claim amounts, and service utilization patterns [1], [2], [3]. When analyzed properly, healthcare claim data can provide valuable insights for improving operational efficiency, monitoring provider behavior, and identifying irregularities in healthcare service delivery.

Healthcare fraud remains a major challenge for insurance providers, public health programs, and healthcare institutions [1], [2], [4], [5]. Fraudulent claims may occur in various forms, including excessive billing, unnecessary procedures, incorrect coding, duplicate claims, and abnormal provider practices [2], [3], [6]. Although only a small proportion of total claims may be fraudulent, their financial and operational impact can be substantial. A recent systematic review reported that healthcare fraud may account for approximately 3% to 10% of total healthcare expenditure [1], while also emphasizing that fraudulent cases are difficult to detect because they are relatively rare and often hidden within large volumes of legitimate claims. Therefore, an automated and data-driven approach is needed to assist healthcare organizations in detecting suspicious claims more effectively.

Machine learning has become a promising approach for healthcare fraud detection because it can identify complex patterns from large and heterogeneous claim datasets [7], [8]. Previous studies have applied supervised, unsupervised, and hybrid machine learning methods to detect fraudulent or suspicious healthcare claims [9], [10], [11], [12]. However, healthcare fraud detection remains challenging due to several factors, including class imbalance, noisy data, heterogeneous claim characteristics, and evolving fraud patterns [13], [14]. In particular, the number of fraudulent claims is usually much smaller than legitimate claims, causing conventional classifiers to become biased toward the majority class. As a result, high accuracy alone may not reflect the actual ability of a model to detect fraud. Metrics such as recall, F1-score, precision-recall AUC, and balanced accuracy are therefore important in evaluating fraud detection models.

Another important issue in healthcare claim fraud detection is model interpretability [15], [16], [17]. In medical and healthcare-related decision support systems, predictive performance alone is not sufficient. Stakeholders such as auditors, insurance analysts, and healthcare administrators also need to understand why a claim is classified as suspicious. Explainable artificial intelligence can help address this problem by identifying the main factors that influence model predictions. For example, features such as claim amount, claim submission delay, provider claim volume, diagnosis-procedure patterns, and visit type may contribute to fraud risk estimation [18]. Recent studies have also highlighted the importance of interpretable and explainable models in healthcare fraud detection, especially to support transparency and practical decision-making.

In addition to class imbalance and interpretability, healthcare claim modeling must also consider the possibility of data leakage. Some variables in claim datasets may only become available after the claim has been reviewed or processed. If such features are used without careful consideration, the model may produce overly optimistic results that do not represent realistic early fraud detection scenarios. For example, claim status and approved claim amount may contain post-decision information that should not always be included in a predictive model designed for early screening. Therefore, a leakage-aware modeling strategy is required to evaluate whether model performance remains reliable when potentially leaked features are excluded.

This study proposes an explainable machine learning approach for detecting fraudulent healthcare insurance claims using imbalanced medical claim data. The study compares several machine learning models, including baseline and tree-based classifiers, under different imbalance handling strategies. Two feature scenarios are considered: a full-

feature scenario and a leakage-aware scenario that excludes features potentially unavailable during early claim screening. Furthermore, model interpretability is examined using feature importance analysis and explainable AI techniques to identify the most influential factors associated with fraudulent claims.

The main contributions of this study are as follows. First, this study develops a machine learning pipeline for healthcare claim fraud detection using structured medical insurance data. Second, it evaluates the impact of imbalanced learning strategies on fraud classification performance. Third, it compares full-feature and leakage-aware modeling scenarios to assess the effect of potentially leaked variables. Fourth, it applies explainable AI to interpret the key features influencing fraud prediction. Through these contributions, this study aims to support the development of transparent, reliable, and data-driven fraud detection systems in healthcare informatics.

2. Method

Research Design:

This study employed a supervised machine learning approach to detect fraudulent healthcare insurance claims using structured medical claim data [14]. The research design focused on three main aspects: imbalanced classification, leakage-aware feature selection, and model interpretability. Since fraudulent claims represent a minority class in real-world healthcare claim systems, this study evaluated several machine learning models under different imbalance handling strategies. In addition, two feature scenarios were developed to examine whether potentially leaked variables could influence model performance in an unrealistic manner.

The overall research workflow consisted of several stages: dataset acquisition, data understanding, preprocessing, feature engineering, feature scenario construction, model training, imbalance handling, performance evaluation, and model interpretation. The final objective was not only to identify the best-performing model, but also to explain which claim-related factors contributed most to fraud prediction

Data Description:

The dataset used in this study consists of 10,000 healthcare insurance claim records with 20 attributes, including patient information, provider information, claim characteristics, financial information, diagnosis and procedure codes, temporal variables, and the target variable. The target variable is Is_Fraud, where a value of 0 indicates a legitimate claim and a value of 1 indicates a fraudulent claim.

The dataset contains 9,171 legitimate claims and 829 fraudulent claims, meaning that the fraud rate is approximately 8.29%. This distribution indicates a clear class imbalance, where fraudulent claims are substantially less frequent than legitimate claims. Such imbalance is common in fraud detection problems and may cause conventional classifiers to favor the majority class if no appropriate handling strategy is applied.

The main attributes in the dataset include patient age, patient gender, diagnosis code, procedure code, claim amount, approved amount, insurance type, claim submission date, days between service and claim submission, monthly claim volume per provider, provider specialty, patient state, claim status, length of stay, visit type, chronic condition flag, and prior visit history.

Data Preprocessing:

Before model development, several preprocessing steps were applied to prepare the dataset for machine learning analysis. First, the target variable `Is_Fraud` was separated from the predictor variables. The identifier variable `Claim_ID` was removed because it only represents a unique claim identifier and does not provide meaningful predictive information.

Missing values were found in three variables: `Insurance_Type`, `Provider_Specialty`, and `Prior_Visits_12m`. For numerical variables, missing values were handled using median imputation. For categorical variables, missing values were handled using the most frequent category. This approach was selected to preserve all records in the dataset while reducing the potential bias caused by incomplete data.

Numerical variables were standardized using standard scaling to ensure that variables with larger numerical ranges did not dominate the learning process. Categorical variables such as gender, diagnosis code, provider specialty, patient state, claim status, visit type, and insurance type were transformed using one-hot encoding [19]. The variable `Procedure_Code` was treated as a categorical feature because it represents a medical procedure code rather than a continuous numerical measurement.

Feature Engineering:

Several additional features were generated to improve the representation of claim characteristics. The `Claim_Submission_Date` variable was converted into three temporal features: claim year, claim month, and claim day of the week. These features were used to capture possible temporal patterns in claim submission behavior.

In addition, logarithmic transformations were applied to financial variables, including `Claim_Amount` and `Approved_Amount`. The transformed variables, `Log_Claim_Amount` and `Log_Approved_Amount`, were created using the natural logarithm transformation with `log1p`. This transformation was applied because claim-related financial values commonly follow a skewed distribution, and logarithmic transformation can reduce the effect of extreme values.

Feature Scenarios:

To evaluate the effect of potentially leaked variables, this study used two feature scenarios: a full-feature scenario and a leakage-aware scenario.

In the full-feature scenario, all available variables were used except the claim identifier and the target variable. This scenario represents a general machine learning setting where most available claim attributes are included in the model.

In the leakage-aware scenario, potentially post-decision variables were excluded from the model. Specifically, `Claim_Status`, `Approved_Amount`, and `Log_Approved_Amount` were removed because these variables may only be available after claim processing or approval decisions. If such variables are used in early fraud detection,

they may introduce data leakage and produce overly optimistic model performance. Therefore, the leakage-aware scenario was designed to better reflect an early screening condition in which only pre-decision or operationally available variables are used.

Data Splitting:

The dataset was divided into training and testing subsets using an 80:20 stratified split [20]. Stratification was applied to preserve the original fraud and non-fraud class distribution in both subsets. This step was important because the dataset is imbalanced, and random splitting without stratification could produce an uneven distribution of fraudulent cases across the training and testing sets.

The training set was used to fit the machine learning models, while the testing set was used to evaluate model generalization on unseen data. A fixed random state was applied to ensure reproducibility of the experimental results.

Imbalance Handling Strategies:

Because fraudulent claims represent only a small proportion of the dataset, this study evaluated three imbalance handling strategies.

The first strategy was the baseline setting, where no imbalance handling technique was applied. This scenario was used to observe how each model performed under the original data distribution.

The second strategy was class weighting, where higher importance was assigned to the minority class during model training. This approach encourages the model to pay more attention to fraudulent claims and reduces the tendency to classify most observations as non-fraud.

The third strategy was Synthetic Minority Oversampling Technique, or SMOTE. SMOTE generates synthetic samples for the minority class by interpolating between existing minority samples. This technique was used to increase the representation of fraudulent claims in the training data and improve the model's ability to recognize fraud patterns [21].

Machine Learning Models:

Several machine learning models were evaluated to compare their ability to detect fraudulent healthcare claims. The selected models included both baseline and tree-based classifiers.

Logistic Regression was used as a baseline model because it is simple, interpretable, and commonly used in binary classification problems. Decision Tree was included because it can capture non-linear relationships and provides a simple tree-based classification structure. Random Forest was used as an ensemble learning model that combines multiple decision trees to improve prediction stability and reduce overfitting. Extra Trees Classifier was included as another ensemble model that introduces additional randomness in tree construction. AdaBoost was also evaluated as a boosting-based model that combines weak learners into a stronger classifier.

Each model was trained under the predefined feature scenarios and imbalance handling strategies. This design allowed the study to compare model performance across multiple experimental settings.

Evaluation Metrics:

Model performance was evaluated using several metrics suitable for imbalanced binary classification. Accuracy was reported to provide a general overview of classification performance. However, since the dataset is imbalanced, accuracy alone was not considered sufficient.

Precision was used to measure the proportion of predicted fraudulent claims that were actually fraudulent. Recall was used to measure the proportion of actual fraudulent claims that were successfully detected by the model. The F1-score was used as the harmonic mean of precision and recall, providing a balanced measure between false positives and false negatives.

In addition, the area under the receiver operating characteristic curve, or ROC-AUC, was used to measure the model's ability to distinguish between fraud and non-fraud classes. The precision-recall AUC, or PR-AUC, was also reported because it is particularly useful for imbalanced datasets. Balanced accuracy was calculated to account for both sensitivity and specificity. Cohen's Kappa was used to measure agreement between predicted and actual labels while considering chance agreement. Mean absolute error and root mean squared error were also reported to provide additional error-based evaluation.

The confusion matrix was used to analyze the number of true positives, true negatives, false positives, and false negatives. In the context of healthcare fraud detection, false negatives are particularly important because they represent fraudulent claims that were incorrectly classified as legitimate.

Explainable AI and Model Interpretation:

To improve interpretability, the best-performing model was analyzed using feature importance and explainable AI techniques. For tree-based models, feature importance was used to identify which variables contributed most to the prediction process. This analysis helped determine whether factors such as claim amount, submission delay, provider claim volume, visit type, diagnosis code, or procedure code played an important role in fraud detection.

Where applicable, SHAP analysis was used to provide a more detailed explanation of model predictions. SHAP values estimate the contribution of each feature to a specific prediction and allow both global and local interpretation of the model. In this study, SHAP was used to support transparency by explaining how different claim characteristics influenced the likelihood of a claim being classified as fraudulent.

This interpretability component is important because healthcare fraud detection systems are not only expected to produce accurate predictions, but also to provide understandable evidence that can support audit, compliance, and decision-making processes.

Experimental Procedure:

The experimental procedure was conducted as follows. First, the dataset was loaded and inspected to identify data types, missing values, and class distribution. Second, preprocessing was applied to handle missing values, encode categorical variables, and scale numerical variables. Third, feature engineering was performed by extracting temporal features and transforming financial variables. Fourth, two feature scenarios were prepared: full-feature and leakage-aware. Fifth, the dataset was split into training and testing subsets using stratified sampling. Sixth, each machine learning model was trained using the three imbalance handling strategies. Seventh, model performance was evaluated using multiple classification metrics. Finally, the best-performing model was interpreted using feature importance and explainable AI analysis.

This methodological design allows the study to evaluate not only predictive performance, but also the effects of imbalanced learning, potential data leakage, and model interpretability in healthcare claim fraud detection.

3. Result and Discussion

Dataset Characteristics:

The dataset used in this study consisted of 10,000 healthcare insurance claim records with a binary target variable, `Is_Fraud`. As shown in Figure 1, the dataset was highly imbalanced. There were 9,171 non-fraud claims, representing 91.71% of the dataset, and 829 fraud claims, representing only 8.29% of the total records. This distribution confirms that healthcare fraud detection is an imbalanced classification problem, where fraudulent cases are substantially less frequent than legitimate claims.

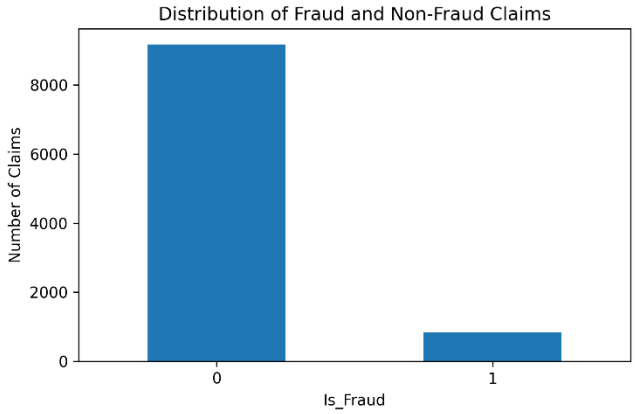

Figure 1. Distribution of fraud and non-fraud claims

The imbalance pattern is important because a model can achieve high accuracy by mostly predicting the majority class. Therefore, performance evaluation in this study did not rely only on accuracy, but also included precision, recall, F1-score, ROC-AUC, PR-AUC, balanced accuracy, Cohen's Kappa, MAE, and RMSE.

Missing values were found in three variables: `Insurance_Type`, `Provider_Specialty`, and `Prior_Visits_12m`. Each of these variables contained 350 missing values, equivalent to 3.5% of the dataset, as shown in Figure 2. The presence of missing values reflects a realistic condition in healthcare claim data, where administrative and provider-related information may not always be completely recorded.

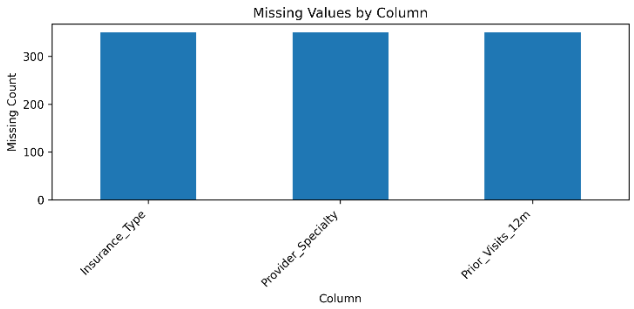

Figure 2. Missing values by column

The missing values were handled through imputation during preprocessing. Numerical missing values were imputed using the median, while categorical missing values were imputed using the most frequent category. This strategy allowed all records to be retained for model development.

Overall Model Performance:

The experimental results show that the best-performing model was obtained using the Full Features scenario without additional imbalance handling. The model was Logistic Regression, achieving an accuracy of 0.9900, precision of 0.9740, recall of 0.9036, F1-score of 0.9375, ROC-AUC of 0.9989, and PR-AUC of 0.9896.

The top overall results are presented in Table 1.

Table 1. Top model performance based on F1-score

Feature Set	Balancing Method	Model	ROC-AUC	Accuracy	Precision	Recall	F1-score	PR-AUC
Full Features	None	Logistic Regression	9,989	9,900	9,740	9,036	9,375	9,896
Full Features	SMOTE	Logistic Regression	9,990	9,885	9,040	9,639	9,329	9,901
Full Features	Class Weight	Logistic Regression	9,990	9,845	8,649	9,639	9,117	9,897
Full Features	None	AdaBoost	9,945	9,800	9,038	8,494	8,758	9,500
Full Features	None	Decision Tree	9,231	9,745	8,363	8,614	8,487	7,319

The results indicate that Logistic Regression performed strongly on this dataset. This suggests that the relationship between several predictive variables and fraud status may be sufficiently separable in the transformed feature space, especially after numerical scaling, categorical encoding, and financial feature transformation. However, this result

should be interpreted carefully because the best-performing model used the full-feature scenario, which included variables that may contain post-decision information.

Interestingly, SMOTE improved the recall of Logistic Regression from 0.9036 to 0.9639, meaning that the model detected more fraudulent claims. However, this improvement came with a decrease in precision from 0.9740 to 0.9040. This means that SMOTE reduced false negatives but increased false positives. In a healthcare fraud detection context, this trade-off is important. A higher recall is beneficial when the main objective is to detect as many suspicious claims as possible, but lower precision may increase the workload for auditors because more legitimate claims may be flagged as fraudulent.

Confusion Matrix Analysis of the Best Model:

The confusion matrix of the best-performing model is shown in **Figure 3**. The test set contained 2,000 claims, consisting of 1,834 non-fraud claims and 166 fraud claims.

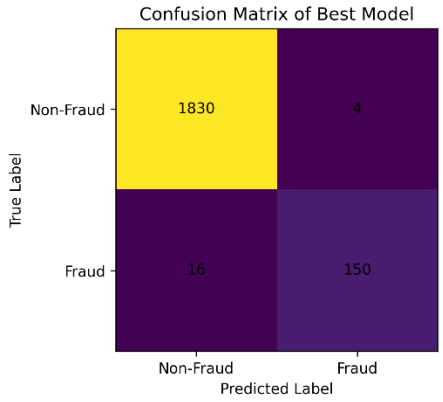

Figure 3. Confusion matrix of the best model

The best model correctly classified 1,830 non-fraud claims as non-fraud and 150 fraud claims as fraud. Only 4 non-fraud claims were incorrectly classified as fraud, while 16 fraud claims were incorrectly classified as non-fraud.

This result shows that the model had a very low false positive rate and a relatively low false negative rate. The fraud recall was 90.36%, indicating that the model successfully detected most fraudulent claims in the test set. The fraud precision was 97.40%, meaning that most claims predicted as fraud were truly fraudulent.

In practical terms, the model would be useful as an automated screening tool because it can identify suspicious claims with a high level of reliability. However, the presence of 16 false negatives remains important because these cases represent fraudulent claims that were missed by the model. In healthcare claim auditing, false negatives may lead to undetected financial leakage, while false positives may increase the burden of manual review.

ROC and Precision-Recall Curve Analysis:

The ROC curve of the best model is shown in **Figure 4**. The curve is positioned very close to the upper-left corner, indicating excellent discrimination between fraud and non-fraud claims. The ROC-AUC score of the best model was 0.9989, suggesting that the model had a very strong ability to rank fraudulent claims higher than legitimate claims.

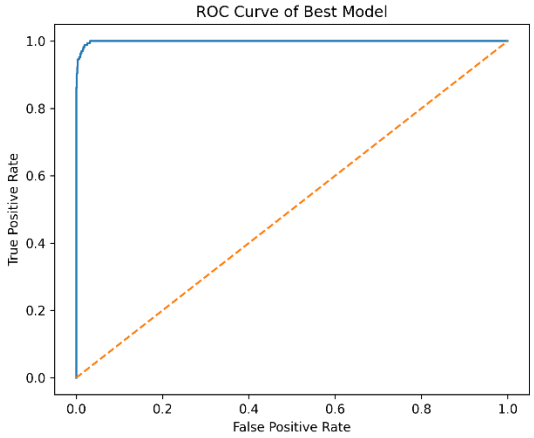

Figure 4. ROC curve of the best model

However, because the dataset is imbalanced, the Precision-Recall curve is especially important. As shown in **Figure 5**, the model maintained high precision across a wide range of recall values. The PR-AUC score was 0.9896, which confirms that the model performed well in detecting the minority fraud class.

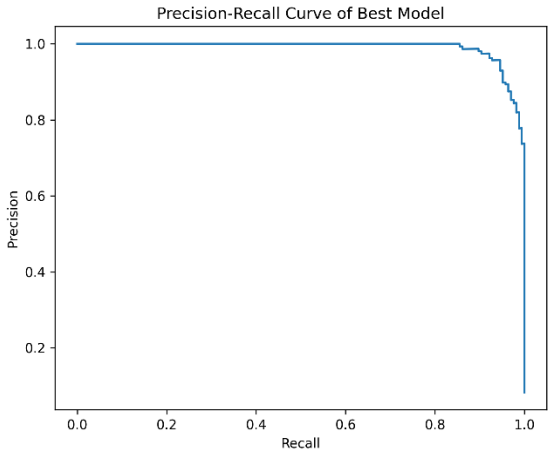

Figure 5. Precision-recall curve of the best model

The strong PR-AUC result is particularly meaningful because PR-AUC is more informative than ROC-AUC in imbalanced classification settings. In this study, the high PR-AUC indicates that the model was able to maintain a favorable balance between detecting fraudulent claims and minimizing false alarms.

Leakage-Aware Model Performance:

To evaluate whether the model performance was affected by potentially leaked variables, this study also tested a leakage-aware feature scenario. In this scenario, `Claim_Status`, `Approved_Amount`, and `Log_Approved_Amount` were excluded because these variables may not be available during early fraud screening.

The best leakage-aware model was AdaBoost without imbalance handling, with an accuracy of 0.9555, precision of 0.7984, recall of 0.6205, F1-score of 0.6983, ROC-AUC of 0.9725, and PR-AUC of 0.7875.

Table 2. Best leakage-aware model performance

Feature Set	Balancing Method	Model	Accuracy	Precision	Recall	F1-score	ROC-AUC	PR-AUC
Leakage-Aware	None	AdaBoost	9,555	8,495	6,205	6,983	9,725	7,875
Leakage-Aware	SMOTE	Decision Tree	9,305	5,652	7,048	6,273	8,279	4,229
Leakage-Aware	None	Decision Tree	9,345	5,978	6,446	6,203	8,027	4,148
Leakage-Aware	None	Logistic Regression	9,395	6,531	5,783	6,134	9,602	7,123
Leakage-Aware	None	Random Forest	9,495	8,495	4,759	6,100	9,609	7,447

Compared with the best full-feature model, the best leakage-aware model showed a clear decrease in F1-score from 0.9375 to 0.6983. This represents an absolute reduction of approximately 0.2392. Recall also decreased from 0.9036 to 0.6205, and PR-AUC decreased from 0.9896 to 0.7875.

This performance gap suggests that post-decision or highly informative administrative variables may have contributed strongly to the near-perfect performance of the full-feature model. Therefore, the leakage-aware scenario provides a more conservative and operationally realistic estimate of model performance for early fraud detection.

This finding is important for healthcare AI implementation. If the objective is to detect fraud after the claim has already been processed, full-feature modeling may be acceptable. However, if the objective is early fraud screening before claim approval, leakage-aware modeling is more appropriate.

Effect of Imbalance Handling:

The results show that imbalance handling had different effects depending on the model and feature scenario. In the full-feature setting, SMOTE improved the recall of Logistic Regression from 0.9036 to 0.9639, but precision decreased from 0.9740 to 0.9040. Similarly, class weighting produced the same recall of 0.9639, but precision further decreased to 0.8649.

This pattern indicates that imbalance handling can help detect more fraudulent claims, but it may also increase the number of false positives. In healthcare claim auditing, this trade-off should be aligned with organizational priorities. If the primary objective is to minimize undetected fraud, higher recall may be preferred. If the objective is to reduce unnecessary investigations, higher precision may be more important.

In the leakage-aware setting, class weighting and SMOTE also increased recall for some models. For example, Logistic Regression with class weighting achieved a recall of 0.8976, detecting 149 out of 166 fraud cases. However, it also produced 220 false positives, resulting in a lower precision of 0.4038 and F1-score of 0.5570. This indicates that although the model became more sensitive to fraud, it also flagged many legitimate claims as suspicious.

Therefore, imbalance handling should not be evaluated only based on recall. A balanced evaluation using F1-score, PR-AUC, and confusion matrix analysis is necessary to determine whether the resulting model is practical for real-world fraud screening.

Feature Importance Analysis:

The feature importance analysis of the best model is shown in **Figure 6**. Since the best model was Logistic Regression, the importance values represent the magnitude of model coefficients after preprocessing. The most influential features included `Log_Claim_Amount`, `Log_Approved_Amount`, `Claim_Amount`, `Approved_Amount`, and `Days_Between_Service_and_Claim`.

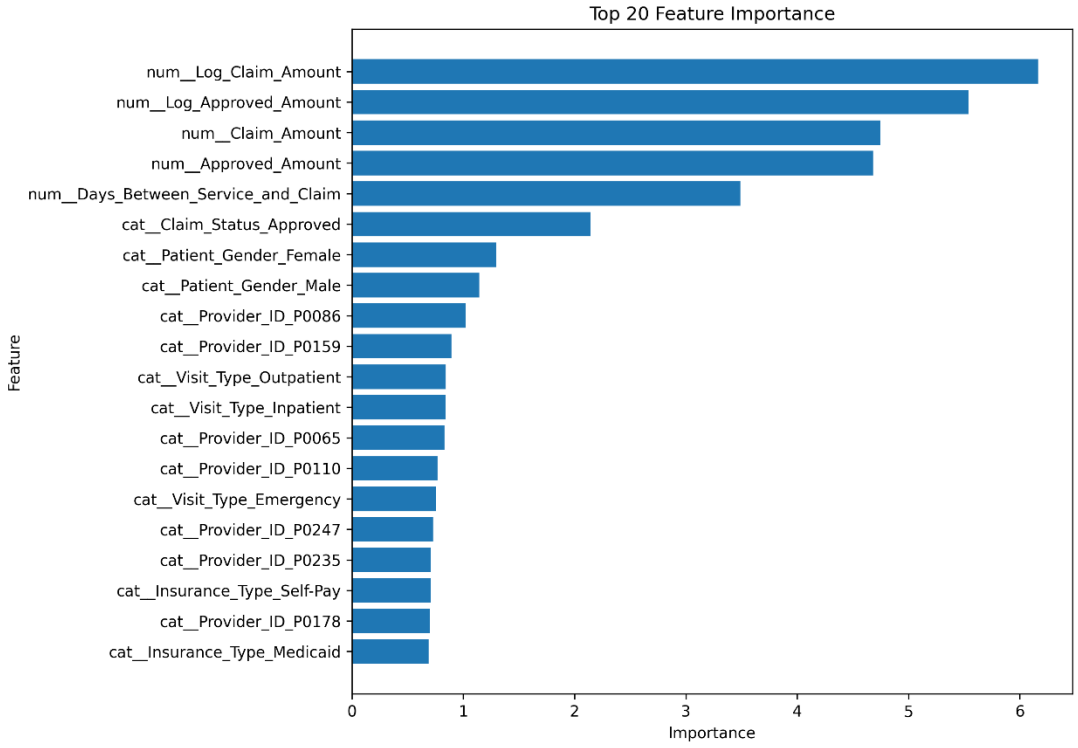

Figure 6. Top 20 feature importance of the best model

The dominance of claim-related financial variables suggests that claim value plays a central role in fraud prediction. Higher claim amounts may indicate abnormal billing behavior or unusual service patterns. The importance of `Days_Between_Service_and_Claim` also supports the assumption that claim submission timing is relevant, where suspicious claims may follow different submission patterns compared with legitimate claims.

However, the importance of `Log_Approved_Amount`, `Approved_Amount`, and `Claim_Status_Approved` also supports the concern regarding potential data leakage. These variables may contain information that is only available after claim review or approval. Their presence among the most important features explains why the full-feature model achieved very high performance. This reinforces the importance of including the leakage-aware scenario in the study.

Several provider-related features, such as specific `Provider_ID` categories, also appeared among the top predictors. This indicates that provider-level behavior may contribute to fraud risk estimation. Nevertheless, such features should be interpreted carefully because the model may learn provider-specific patterns from the dataset. If the model is later applied to new providers not present in the training data, its performance may differ.

The appearance of patient gender categories among the important features should also be interpreted as model association rather than causal evidence. Since categorical variables were transformed using one-hot encoding, the importance of such variables reflects their statistical contribution within the dataset, not necessarily a direct causal relationship with fraud.

Discussion:

The findings of this study demonstrate that machine learning can effectively detect fraudulent healthcare insurance claims, particularly when informative claim-related features are available. The best full-feature model achieved high predictive performance, with an F1-score of 0.9375, ROC-AUC of 0.9989, and PR-AUC of 0.9896. These results show that structured healthcare claim data contain useful patterns for distinguishing fraudulent and legitimate claims.

However, the comparison between full-feature and leakage-aware scenarios reveals an important methodological issue. The substantial decline in performance after removing potentially leaked variables suggests that some features may provide information that would not be available during early claim screening. Therefore, the full-feature model may overestimate real-world early detection performance.

From a practical perspective, this study suggests that healthcare fraud detection systems should be designed according to their intended operational use. If the system is intended for post-payment audit, variables such as approved amount and claim status may be useful. However, if the system is intended for early fraud detection before claim approval, these variables should be excluded to avoid data leakage.

The results also show that imbalance handling should be applied carefully. SMOTE and class weighting can improve the detection of fraudulent claims, but they may increase false positives. In fraud detection, this trade-off is not merely technical; it has operational consequences. A high-recall model may reduce missed fraud cases, but it can also increase the workload of audit teams. Therefore, model selection should consider both predictive performance and practical audit capacity.

Overall, the proposed approach provides a useful framework for healthcare claim fraud detection by combining machine learning, imbalance-aware evaluation, leakage-aware modeling, and interpretability analysis. The results

support the potential use of explainable machine learning as a decision-support tool for healthcare informatics, insurance auditing, and claim integrity monitoring.

Summary of Findings:

The main findings of this study can be summarized as follows. First, the dataset was highly imbalanced, with fraud cases representing only 8.29% of all claims. Second, the best full-feature model was Logistic Regression, which achieved an F1-score of 0.9375 and PR-AUC of 0.9896. Third, SMOTE and class weighting improved fraud recall but reduced precision, indicating a trade-off between fraud detection sensitivity and false positive control. Fourth, the leakage-aware scenario produced lower but more realistic performance, with the best F1-score of 0.6983. Fifth, feature importance analysis showed that claim amount, approved amount, claim submission delay, claim status, and provider-related variables were important predictors of fraud [22].

These findings indicate that explainable machine learning can support healthcare claim fraud detection, but careful attention must be given to class imbalance, data leakage, and the operational context in which the model will be deployed.

4. Conclusion

This study presented an explainable machine learning approach for detecting fraudulent healthcare insurance claims using imbalanced medical insurance data. The dataset consisted of 10,000 healthcare claim records, with fraudulent claims representing only 8.29% of the total data. This class distribution confirmed that healthcare fraud detection is an imbalanced classification problem that requires careful evaluation beyond accuracy alone.

The experimental results showed that the best full-feature model was Logistic Regression without additional imbalance handling, achieving an accuracy of 0.9900, precision of 0.9740, recall of 0.9036, F1-score of 0.9375, ROC-AUC of 0.9989, and PR-AUC of 0.9896. The confusion matrix further showed that the model correctly detected 150 out of 166 fraudulent claims in the test set, with only 4 false positives and 16 false negatives. These results indicate that structured healthcare claim data contain strong predictive patterns for distinguishing fraudulent and legitimate claims.

However, the comparison between full-feature and leakage-aware scenarios revealed an important methodological issue. When potentially post-decision variables such as claim status and approved amount were removed, the best leakage-aware model achieved a lower F1-score of 0.6983. This decline suggests that some variables in the full-feature model may contain information that is not available during early claim screening. Therefore, while the full-feature model provides strong predictive performance, the leakage-aware model offers a more realistic estimate for early fraud detection applications.

The study also found that imbalance handling techniques such as SMOTE and class weighting improved fraud recall in several scenarios, but often reduced precision. This indicates a trade-off between detecting more fraudulent claims and increasing false positives. In practical healthcare claim auditing, this trade-off should be considered

carefully because false positives may increase manual review workload, while false negatives may allow fraudulent claims to remain undetected.

The feature importance analysis showed that claim-related financial variables, including claim amount and approved amount, were among the most influential predictors. In addition, days between service and claim submission, claim status, visit type, insurance type, and provider-related variables also contributed to fraud prediction. These findings support the value of explainable machine learning because they help identify the main factors influencing model decisions and provide more transparent insights for healthcare auditors and claim analysts.

Overall, this study demonstrates that machine learning can support healthcare fraud detection as part of AI-driven health informatics and medical insurance analytics. The proposed approach combines predictive modeling, imbalance-aware evaluation, leakage-aware feature analysis, and explainable AI to provide a more transparent and operationally relevant fraud detection framework. Future studies may extend this work by using real-world healthcare claim data, applying advanced gradient boosting models, incorporating temporal provider behavior analysis, and validating the model in prospective claim screening environments [23], [6].

References:

[1] A. Du Preez, S. Bhattacharya, P. Beling, and E. Bowen, "Fraud detection in healthcare claims using machine learning: A systematic review," *Artif. Intell. Med.*, vol. 160, p. 103061, Feb. 2025, doi: 10.1016/j.artmed.2024.103061.

[2] Z. Hamid, F. Khalique, S. Mahmood, A. Daud, A. Bukhari, and B. Alshemaimri, "Healthcare insurance fraud detection using data mining," *BMC Med. Inform. Decis. Mak.*, vol. 24, no. 1, p. 112, Apr. 2024, doi: 10.1186/s12911-024-02512-4.

[3] E. Nabrawi and A. Alanazi, "Fraud Detection in Healthcare Insurance Claims Using Machine Learning," *Risks*, vol. 11, no. 9, p. 160, Sep. 2023, doi: 10.3390/risks11090160.

[4] Z. Wang, X. Chen, Y. Wu, L. Jiang, S. Lin, and G. Qiu, "A robust and interpretable ensemble machine learning model for predicting healthcare insurance fraud," *Sci. Rep.*, vol. 15, no. 1, p. 218, Jan. 2025, doi: 10.1038/s41598-024-82062-x.

[5] K. Razzaq and M. Shah, "Next-Generation Machine Learning in Healthcare Fraud Detection: Current Trends, Challenges, and Future Research Directions," *Information*, vol. 16, no. 9, p. 730, Aug. 2025, doi: 10.3390/info16090730.

[6] A. A. Amponsah, A. F. Adekoya, and B. A. Weyori, "A novel fraud detection and prevention method for healthcare claim processing using machine learning and blockchain technology," *Decis. Anal. J.*, vol. 4, p. 100122, Sep. 2022, doi: 10.1016/j.dajour.2022.100122.

[7] O. Cherkaoui, H. Anoun, and A. Maizate, "A benchmark of health insurance fraud detection using machine learning techniques," *IAES Int. J. Artif. Intell. IJ-AI*, vol. 13, no. 2, p. 1925, Jun. 2024, doi: 10.11591/ijai.v13.i2.pp1925-1934.

[8] J. M. Johnson and T. M. Khoshgoftaar, "Data-Centric AI for Healthcare Fraud Detection," *SN Comput. Sci.*, vol. 4, no. 4, p. 389, May 2023, doi: 10.1007/s42979-023-01809-x.

[9] J. Lu, K. Lin, R. Chen, M. Lin, X. Chen, and P. Lu, "Health insurance fraud detection by using an attributed heterogeneous information network with a hierarchical attention mechanism," *BMC Med. Inform. Decis. Mak.*, vol. 23, no. 1, p. 62, Apr. 2023, doi: 10.1186/s12911-023-02152-0.

[10] S. Mardani and H. Moradi, "Using Graph Attention Networks in Healthcare Provider Fraud Detection," *IEEE Access*, vol. 12, pp. 132786–132800, 2024, doi: 10.1109/ACCESS.2024.3425892.

[11] L. Settipalli and G. R. Gangadharan, "WMTDBC: An unsupervised multivariate analysis model for fraud detection in health insurance claims," *Expert Syst. Appl.*, vol. 215, p. 119259, Apr. 2023, doi: 10.1016/j.eswa.2022.119259.

[12] Y. Yoo, J. Shin, and S. Kyeong, "Medicare Fraud Detection Using Graph Analysis: A Comparative Study of Machine Learning and Graph Neural Networks," *IEEE Access*, vol. 11, pp. 88278–88294, 2023, doi: 10.1109/ACCESS.2023.3305962.

[13] R. Bounab, K. Zarour, B. Guelib, and N. Khlifa, "Enhancing Medicare Fraud Detection Through Machine Learning: Addressing Class Imbalance With SMOTE-ENN," *IEEE Access*, vol. 12, pp. 54382–54396, 2024, doi: 10.1109/ACCESS.2024.3385781.

[14] D. Farahmandazad, K. Danesh, and H. F. N. Abadi, "Application of Standard Machine Learning Models for Medicare Fraud Detection with Imbalanced Data," *Risks*, vol. 13, no. 10, p. 198, Oct. 2025, doi: 10.3390/risks13100198.

[15] M. Balayet Hossain Sakil *et al.*, "Enhancing Medicare Fraud Detection With a CNN-Transformer-XGBoost Framework and Explainable AI," *IEEE Access*, vol. 13, pp. 79609–79622, 2025, doi: 10.1109/ACCESS.2025.3562577.

[16] J. T. Hancock, R. A. Bauder, H. Wang, and T. M. Khoshgoftaar, "Explainable machine learning models for Medicare fraud detection," *J. Big Data*, vol. 10, no. 1, p. 154, Oct. 2023, doi: 10.1186/s40537-023-00821-5.

[17] R. Muhammad *et al.*, "Fraud detection and explanation in medical claims using GNN architectures," *Sci. Rep.*, vol. 15, no. 1, p. 41734, Nov. 2025, doi: 10.1038/s41598-025-22910-6.

[18] J. Kemp, C. Barker, N. Good, and M. Bain, "Sequential pattern detection for identifying courses of treatment and anomalous claim behaviour in medical insurance," in *2022 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, Las Vegas, NV, USA: IEEE, Dec. 2022, pp. 3039–3046. doi: 10.1109/BIBM55620.2022.9995541.

[19] Y. Chen, C. Zhao, and C. Nie, "Health Insurance Fraud Detection: The Role of Feature Engineering and Preprocessing Techniques," in *Proceedings of the 2nd Guangdong-Hong Kong-Macao Greater Bay Area International Conference on Digital Economy and Artificial Intelligence*, Dongguan China: ACM, Mar. 2025, pp. 858–862. doi: 10.1145/3745238.3745373.

[20] D. Dash, M. Kumar, S. Patra, A. Kumar, and A. Ganguly, "Healthcare Fraud Detection Using an Integrated ML Approach with SMOTE," *Procedia Comput. Sci.*, vol. 258, pp. 800–810, 2025, doi: 10.1016/j.procs.2025.04.312.

[21] H. Shi, M. A. Tayebi, J. Pei, and J. Cao, "Cost-Sensitive Learning for Medical Insurance Fraud Detection With Temporal Information," *IEEE Trans. Knowl. Data Eng.*, vol. 35, no. 10, pp. 10451–10463, Oct. 2023, doi: 10.1109/TKDE.2023.3240431.

[22] B. Hong, P. Lu, H. Xu, J. Lu, K. Lin, and F. Yang, "Health insurance fraud detection based on multi-channel heterogeneous graph structure learning," *Heliyon*, vol. 10, no. 9, p. e30045, May 2024, doi: 10.1016/j.heliyon.2024.e30045.

[23] M. A. Mohammed, M. Boujelben, and M. Abid, "A Novel Approach for Fraud Detection in Blockchain-Based Healthcare Networks Using Machine Learning," *Future Internet*, vol. 15, no. 8, p. 250, Jul. 2023, doi: 10.3390/fi15080250.