FRAUD DETECTION IN HEALTH INSURANCE CLAIMS USING ENSEMBLE CLUSTERING

Neo Alit Cahya[1,2], Rinaldi Munir[1]

[1]Institut Teknologi Bandung, Jl. Ganesa No.10, Lb. Siliwangi, Kecamatan Coblong, Kota Bandung, Jawa Barat 40132
[2]PT PLN (Persero), Jalan Trunojoyo Blok M – I No 135 Kebayoran Baru, Jakarta 12160
Email: neo.alit.cahya@gmail.com[1,2], rinaldi-m@stei.itb.ac.id [1]

Abstract. The rapid increase in health insurance claims has emphasized the need for innovative fraud detection approaches, particularly for unlabeled datasets. Fraudulent activities in insurance claims can lead to significant financial losses and disrupt public trust in insurance systems. Existing studies have shown the effectiveness of models like Isolation Forest, K-Means, and Local Outlier Factor (LOF) for anomaly detection. However, the integration of ensemble methods for unsupervised learning in fraud detection remains underexplored.

This study proposes an ensemble-based anomaly detection framework combining K-Means, LOF, and Isolation Forest, coupled with dimensionality reduction using Principal Component Analysis (PCA). The model evaluates anomalies in health insurance claim data by integrating transactional and behavioral patterns, aiming to enhance anomaly detection accuracy. The evaluation metrics include Silhouette Score, Davies-Bouldin Index, and Calinski-Harabasz Index.

Results show that the ensemble model with a weighted voting mechanism outperforms standalone algorithms and other ensemble methods. The weighted voting ensemble also demonstrates superior performance compared to Majority Vote and Any Vote methods, confirming its robustness in detecting anomalies in unlabeled data.

This study highlights the potential of ensemble methods in unsupervised anomaly detection for health insurance claims, providing a scalable and effective solution to mitigate fraudulent activities in the insurance sector.

Keywords: *Health Insurance Claims, Anomaly Detection, Unsupervised Learning, Ensemble Methods, Principal Component Analysis (PCA).*

1 Introduction

Health insurance fraud poses significant financial and operational challenges to insurance providers, including increased claim costs and diminished public trust. Traditional fraud detection methods rely heavily on manual verification or supervised machine learning, both of which require labeled datasets. However, the availability of labeled data is often limited, especially in large-scale systems like health insurance. To address this gap, unsupervised learning methods[1] such as Isolation Forest, K-Means, and Local Outlier Factor (LOF) have been widely explored for anomaly detection[2]. Recent studies highlight the effectiveness of ensemble methods in improving model accuracy and robustness, yet their application in unsupervised anomaly detection remains underexplored. This study develops an ensemble-based framework integrating these algorithms with dimensionality reduction techniques to enhance the detection of anomalies in health insurance claims, aiming to reduce fraud effectively.

2 Materials and Methods.

The dataset used in this study consists of 21 features, including claim information (e.g., claim amount and duration), participant details (e.g., member status and division), healthcare benefits, and diagnostic codes (ICD-10). Preprocessing steps involve handling missing values, normalizing data, and applying Principal Component Analysis (PCA) for dimensionality reduction to retain the most informative features while reducing noise.

The anomaly detection framework integrates three unsupervised algorithms: K-Means for clustering, Local Outlier Factor (LOF) for density-based anomaly detection, and Isolation Forest for tree-based anomaly classification. To optimize performance, the outputs of these algorithms are combined using ensemble methods, including Majority Vote, Weighted Vote, and Any Vote mechanisms[3].

Received ________, Revised _________, Accepted for publication __________

3 Results and Discussion

The methodology involves preprocessing transaction and behavioral pattern data through cleansing, transformation, and feature selection, followed by dimensional reduction using PCA. Clustering algorithms (K-Means, LOF, Isolation Forest) are applied with parameter optimization, and their outputs are combined using ensemble methods (All Vote, Majority Vote, Any Vote, Weighted Vote). The Weighted Ensemble is used for fraud detection, leveraging its superior performance, and the results are evaluated using Silhouette Score, Davies-Bouldin Index, and Calinski-Harabasz Index to ensure effective anomaly detection.

3.1 Transactional Data

The results in Table 1 highlight the superiority of ensemble methods, with Weighted Vote achieving the best performance, evidenced by a high Silhouette Score (0.93) and the lowest Davies-Bouldin Index (0.555), indicating optimal cluster quality and separation. Isolation Forest also performed well individually, while K-Means demonstrated strong dispersion (Calinski-Harabasz Index 189,263) but lower compactness due to outlier sensitivity. LOF struggled with high-dimensional data, showing suboptimal clustering. Weighted Vote's ability to combine algorithm strengths effectively makes it the most robust approach for anomaly detection in transaction data.

Table 1 Transaction data evaluation scores.

No	Parameter	Evaluation Score		
		Silhouette Score	Davies-Bouldin Index	Calinski-Harabasz Index
1	Kmeans	0.515	0.96	189263
2	iForest	0.925	0.72	110094
3	LOF	0.83	1.005	2003
4	E-AllVote	0.507	0.69	37
5	E-MajorityVote	0.928	0.678	116247
6	E-AnyVote	0.498	0.89	185898
7	E- WeightedVote	0.93	0.555	140242

3.2 Behavioral Pattern

The evaluation of behavioral pattern models shows that K-Means and Isolation Forest performed best, with K-Means achieving a balanced Silhouette Score (0.681) and high Calinski-Harabasz Index (113,855), while Isolation Forest had the highest Silhouette Score (0.87) and lowest Davies-Bouldin Index (0.537). LOF performed poorly, indicating its limitations in this context. Among ensemble methods, AnyVote showed moderate success, closely matching K-Means in clustering quality, but Majority Vote and Weighted Vote produced low-quality results with poor scores across all metrics. AllVote failed to generate meaningful outcomes, forming only a single class. These results suggest that standalone algorithms, particularly K-Means and Isolation Forest, are more reliable for behavioral pattern data, while ensemble methods require refinement to improve their performance.

No	Parameter	Evaluation Score		
		Silhouette Score	Davies-Bouldin Index	Calinski-Harabasz Index
1	Kmeans	0.681	0.729	113855
2	iForest	0.87	0.537	5637
3	LOF	0.235	1.062	81
4	E-AllVote	[redacted]	[redacted]	[redacted]
5	E-MajorityVote	0.122	1.205	24
6	E-AnyVote	0.675	0.727	102926
7	E- WeightedVote	0.122	1.205	24

3.3 Fraud Detection

The evaluation results highlight the superior performance of the Weighted Voting Ensemble in fraud detection, achieving the highest Silhouette Score (0.93) and Calinski-Harabasz Index (140,242), along with the lowest Davies-Bouldin Index (0.555), indicating well-defined clusters with optimal separation and dispersion. In comparison, the Majority Vote Ensemble showed moderate performance, with a Silhouette Score of 0.883 and a Davies-Bouldin Index of 0.942, reflecting reasonable but less effective clustering. The Any Vote Ensemble performed the worst, with a very low Silhouette Score (0.203) and high Davies-Bouldin Index (1.94), indicating poor cluster compactness and separation, supported by its low Calinski-Harabasz Index (234). These results confirm that the Weighted Voting Ensemble is the most robust and accurate method for detecting anomalies in health insurance fraud.

4 Conclusions

By combining K-Means, Local Outlier Factor, and Isolation Forest with dimensionality reduction through PCA, the framework effectively leverages the strengths of each algorithm to improve anomaly detection accuracy. The results also underscore the limitations of simpler ensemble methods, such as Majority Vote and Any Vote, which perform less effectively. Overall, the proposed framework provides a robust and scalable solution for fraud detection in health insurance systems, offering significant improvements in handling unlabeled data and operational efficiency.

5 References

[1] J. Debener, V. Heinke, and J. Kriebel, “Detecting insurance fraud using supervised and unsupervised machine learning,” *Journal of Risk and Insurance*, vol. 90, no. 3, pp. 743–768, Sep. 2023, doi: 10.1111/jori.12427.

[2] Kanksha, A. Bhaskar, S. Pande, R. Malik, and A. Khamparia, “An intelligent unsupervised technique for fraud detection in health care systems,” *Intelligent Decision Technologies*, vol. 15, no. 1, pp. 127–139, Mar. 2021, doi: 10.3233/IDT-200052.

[3] M. RE and G. VALENTINI, “Ensemble Methods,” 2012. doi: 10.1201/b11822-34.