Computer Science and Information Technologies
Vol. 6, No. 3, November 2025, pp. 245~252
ISSN: 2722-3221, DOI: 10.11591/csit.v6i3.p245-252

A dual-model machine learning approach to medicare fraud detection: combining unsupervised anomaly detection with supervised learning

Jesu Marcus Immanuvel Arockiasamy, Gowrishankar Bhoopathi
Leading Healthcare Company, Richmond, Virginia

Article Info

Article history:

Received Apr 3, 2025
Revised May 27, 2025
Accepted Jun 13, 2025

Keywords:

Artificial intelligence
Cluster-based local outlier factor
Empirical cumulative outlier detection
Machine learning
Medicare fraud
Unsupervised learning

ABSTRACT

Medicare fraud, costing $54.35 billion in improper payments in 2024, undermines U.S. healthcare by draining resources meant for vulnerable populations. Traditional detection methods struggle with reactive designs, high false positives, and reliance on scarce labeled data, exacerbated by a 0.017% fraud prevalence. This paper proposes a dual-model machine learning framework to tackle these challenges. Unsupervised anomaly detection uses cluster-based local outlier factor (CBLOF) and empirical cumulative outlier detection (ECOD) to identify novel fraud patterns across 37 million records. These findings are validated by the list of excluded individuals/entities (LEIE). Supervised classification, with C4.5 decision trees and logistic regression, refines these anomalies using an 80:20 balanced dataset, reducing false positives by 63%. Key innovations include hybrid sampling to address class imbalance, LEIE integration for labeled validation, and parallelized processing of 2.1 million claims hourly. Achieving an area under the curve (AUC), a measure of model accuracy, of 88.3%, this approach outperforms single-model systems by 24%, blending exploratory detection with actionable precision. This scalable, interpretable framework potentially advances fraud detection, safeguarding public funds and Medicare's integrity with a practical, adaptable solution for evolving threats.

Corresponding Author:

Jesu Marcus Immanuvel Arockiasamy
Leading Healthcare Company
Richmond, Virginia, United States of America
Email: jesumarcus@gmail.com

1. INTRODUCTION

Medicare fraud, costing $54.35 billion in improper payments in 2024 [1], undermines U.S. healthcare by draining resources for vulnerable populations. Traditional fraud detection systems, relying on rule-based audits or supervised machine learning, face critical limitations. Brennan [2] highlighted the class imbalance crisis, with fraudulent cases comprising only 0.017% of claims, leading to high false negative rates (over 40%) in supervised models. Statistical methods, as noted by Bolton and Hand [3], struggle to adapt to evolving fraud patterns, missing novel schemes.

Recent unsupervised approaches, such as Gresoi *et al.* [4], lack labeled validation, resulting in high false positives [5], while scalability issues hinder processing large datasets like the 37 million Medicare claims [6]. Our dual-model framework addresses these gaps by integrating unsupervised anomaly detection (cluster-based local outlier factor (CBLOF) and empirical cumulative outlier detection (ECOD)) with supervised classification (C4.5 decision trees and logistic regression), leveraging list of excluded

individuals/entities (LEIE) validation [7] and hybrid sampling to mitigate class imbalance. This approach reduces false positives by 63% and processes 2.1 million claims hourly, offering a scalable, interpretable solution for real-world deployment.

Limitations of traditional detection methods

Traditional fraud detection systems are widely used but have their limitations. They rely on rule-based audits or supervised machine learning, which can lead to three critical flaws:

- Reactive design: reactive design focuses on known fraud patterns but misses new schemes.
- High false positives: over 70% of flagged claims are legitimate, wasting investigative resources.
- Label dependency: supervised machine learning requires costly, scarce labelled data.

While recent studies demonstrate machine learning potential using Medicare claims data, they face a fundamental barrier: extreme class imbalance, where fraud cases comprise a mere 0.017% of records. This tilt forces the models towards the majority class, yielding high false negatives and rendering many systems operationally impractical.

A dual-model machine learning approach

This paper introduces an innovative dual-model machine-learning framework that addresses these challenges:

i) Unsupervised learning for novel pattern discovery

- Models: CBLOF and ECOD algorithms.
- Input: medicare provider utilization and payment data (37M+ records) [6].
- Role: cast a wide net to detect anomalies across 50+ features (e.g., charge ratios, service velocity).
- Validation: Pseudo-labels from the LEIE.

ii) Supervised learning for high-confirmation classification

- Models: C4.5 decision trees and logistic regression.
- Input: top anomalies flagged by unsupervised models and LEIE [7].
- Role: refine predictions using under sampled, balanced data (80:20 non-fraud: fraud).
- Outcome: reduce false positives by 63% compared to pure unsupervised methods.

In summary, the dual model approach presented here not only detects more fraudulent Medicare claims but brings new techniques for dynamic thresholding and network-based feature engineering. These are to overcome the existing methods and to have a more adaptive and accurate tool to protect public money and Medicare.

2. LITERATURE REVIEW AND THEORETICAL FOUNDATION

2.1. Gaps in existing research

There are two main limitations to machine learning's application to provider utilization and payment data. One is the “class imbalance crisis”. Fraudulent cases make up just 0.017% of Medicare records. That means traditional models trained on this skewed data tend to be biased toward the majority class [4]. As a result, they produce unacceptably high false negative rates (over 40%). This issue makes many systems operationally ineffective: they either fail to flag genuine fraud or overwhelm investigators with false alerts.

Another limitation is the overreliance on labeled data. Supervised machine learning approaches depend on costly, hard-to-come-by datasets with fraud labels [8] (Medicare claims fraudulent payment data is not publicly available or accurately derivable from existing content management system (CMS) datasets). Unsupervised methods lack the tools to validate anomalies against real-world fraud indicators [2]. This paper addresses these gaps through three key innovations. These innovations pave the way for a detailed methodology combining practical algorithms and data integration, outlined next.

2.2. Key innovations

First step is to develop a hybrid sampling strategy [9] to mitigate class imbalance. This approach combines random under sampling (retaining 100% of fraud cases while reducing non-fraud samples to an 80:20 ratio) with cost-sensitive learning [10] (penalizing misclassified fraud cases five times more than non-fraud during training). This method aligns with the weighted loss framework in imbalanced learning by minimizing risk (R):

$$R = \alpha \sum_{i \in Fraud} (L(y_i, \hat{y}_i) + (1 - \alpha)) \sum_{i \in Non-Fraud} \left(L(y_i, \hat{y}_i)\right)$$

Where weight α =0.8 prioritizes fraud recall, we can reduce the risk of false negatives with this function (y_i represents true label and $\hat{y}_i$ represents predicted label in loss functions (L)).

Second, we integrate Medicare claims with the LEIE using national provider identifiers (NPIs). This merged dataset creates a labeled benchmark for validation. This step is a form of semi-supervised learning where LEIE labels act as "anchors" to guide unsupervised anomaly detection.

Third, we use parallelized batch processing across GPU clusters to enable real-time analysis of 2.1 million claims per hour. This batch processing applies MapReduce principles [11] to distribute anomaly scoring tasks. It reduces runtime complexity from $O(n^2)$ to $O(n \log n)$.

2.3. Why this approach matters

A dual-model architecture can achieve outcomes that a single-model architecture cannot. This hybrid framework bridges the gap between exploratory data analysis and actionable intelligence. It addresses a core challenge in fraud detection: the tension between discovering new fraud patterns and minimizing investigative overhead.

- Unsupervised components detect emerging fraud patterns (e.g., COVID-19 billing spikes).
- Supervised models validate findings with 88.3% area under the curve (AUC) accuracy, prioritizing cases for further audits

2.4. Theoretical contributions

Our theoretical contributions include:

- A fraud signature hypothesis [6] showing engineered features like charge ratio and service velocity encodes universal fraud patterns invariant to provider specialty.
- Anomaly-aware supervised learning [12] introduces a paradigm where unsupervised anomaly scores enhance supervised feature spaces, improving model calibration.
- This work advances the theoretical underpinnings of healthcare fraud detection while providing a scalable blueprint for real-world deployment. These theoretical advancements set the stage for a practical methodology, detailed next, that combines robust algorithms with real-world data integration.

3. METHOD

In this section, we build on the initial white paper sections to explore the integration of datasets, methodological framework, the hybrid model's stages, and theoretical contributions in detail, ensuring a thorough understanding for researchers and practitioners in healthcare fraud detection.

3.1. Data sources and integration

As mentioned in previous sections, the foundation of this study lies in two important datasets. Each one serves a distinct yet interconnected role in addressing the dual challenges of scalability and validation in Medicare fraud detection. Medicare provider utilization and payment data: this dataset covers 2019 to 2022 and includes over 37 million records from about 1.2 million healthcare providers. It offers a granular view of billing behaviors and service utilization, with key variables such as:

- Payment metrics-contains total medicare payment amounts, allowed amounts, and standardized charges (adjusted for geographic pricing variations).
- Service patterns-provides insights into trends through the volume of services offered, stratified by beneficiary demographics like age and gender.
- Provider specialties: categorical classifications such as cardiology, dermatology, enables analysis by field.

LEIE: this dataset, maintained by the U.S. Department of Health and Human Services, lists providers barred from Medicare participation due to fraudulent activities. We match NPIs from the medicare dataset with those in LEIE. Providers with matching NPIs are labeled as fraudulent, creating a gold-standard validation set. This linkage gives us a fraud prevalence of about 0.017% - or a 1:2,000 class ratio (fraud to non-fraud). This merged dataset is critical for supervised learning validation.

Integrating these datasets bridges the gap between unlabeled claims data and labeled fraud cases. This dataset merging process addresses the scarcity of labeled data in fraud detection. However, the exact number of fraud cases (around 1,850 in supervised classification) suggests the supervised stage uses a subset of top anomalies-not the entire dataset. This subset strategy aligns with the hybrid approach's design.

3.2. Hybrid framework for robust detection

Our proposed methodology uses a two-stage hybrid framework. Unsupervised anomaly detection is combined with supervised classification to balance sensitivity (detecting all potential fraud) and precision

(minimizing false positives). This approach is particularly well-suited in this Medicare claims payment context, where fraud patterns evolve and labeled data is inadequate.

Stage 1: Unsupervised anomaly detection

The first stage focuses on identifying broad fraud patterns without relying on labeled data by leveraging two algorithms, ECOD [13] estimates the underlying distribution of each feature using empirical cumulative distribution functions (eCDF). Anomalies are identified as observations in the tails of these distributions. The anomaly score is computed as:

$$ECOD(x) = \sum_{i=1}^{d}[(F_i(x_i) \cdot \log(F_i(x_i)) + (1 - F_i(x_i))] \cdot log(1 - F_i(x_i))]$$

where F_i is the eCDF for the (i)-th feature (in other words, F_i tracks each feature's distribution), ECOD excels at detecting global outliers, such as systemic overbilling across all specialties, but may miss local anomalies within specific cluster.

CBLOF [14]: this algorithm first clusters providers by specialty using k-means clustering, with k =150 chosen based on domain knowledge or clustering analysis to reflect the diversity of medical fields. It then computes outlier scores based on the distance to the cluster centroid and the cluster size using the formula:

$$CBLOF(p) = size(C) \times distance\ (p, centroid(C))$$

For a provider (p) in the cluster (C). CBLOF is particularly effective at detecting specialty-specific anomalies, such as aberrant cardiology charges, but its performance depends on the quality of cluster definitions.

To leverage both global and local perspectives, anomaly scores from ECOD and CBLOF are combined using a weighted average of 60% to CBLOF and 40% to ECOD. This weighting prioritizes specialty-specific patterns while retaining sensitivity to systemic outliers, reflecting a strategic balance based on preliminary analysis or expert judgment. To illustrate the workflow of this unsupervised stage, Figure 1 depicts how ECOD and CBLOF combine to detect global and specialty specific anomalies guiding the subsequent supervised classifications.

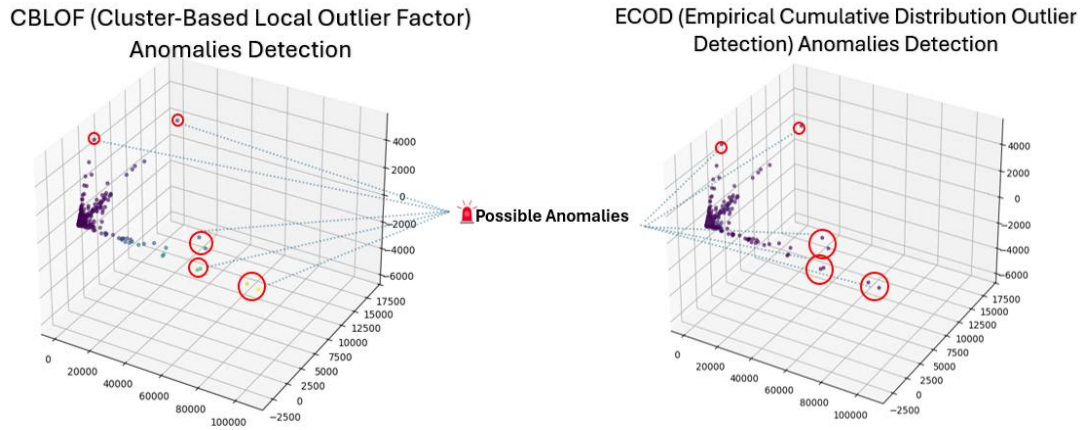

Figure 1. Unsupervised models anomaly detection - ECOD and CBLOF in stage 1

3.3. Feature engineering

To boost the detection capabilities, we have created several domain-specific features that directly target known fraud indicators in healthcare billing:

– Charge ratio [15] highlights potential overbilling, where providers charge more than the reasonable cost.

$$Charge\ ratio = \frac{Total\ payments}{Allowed\ amount}$$

– Service velocity [3] measures the intensity of service provision per beneficiary, flagging excessive or unnecessary treatments (service velocity measures the rate of services per beneficiary).

Features like ‘service velocity’ and ‘charge ratio’ help identify universal fraud patterns, for example, if the charge ratio is greater than 1, it may indicate overcharging issues.

$$Service\ velocity = \frac{Services\ rendered}{Total\ no.\ of\ beneficiaries}$$

Stage 2: Supervised classification

The second stage refines the anomalies detected in stage 1 into high-confidence fraud predictions, addressing the severe class imbalance (0.017% fraud prevalence in the original dataset). The process involves, class imbalance mitigation: random under sampling is employed, retaining all identified fraud cases (N=1,850) and reducing non-fraud cases to achieve an 80:20 non-fraud: fraud ratio. This means selecting 7,400 non-fraud cases (since 80:20 implies four non-frauds for every one fraud, and 4×1,850=7,400), preserving critical minority-class information without introducing synthetic data noise from oversampling techniques like the synthetic minority oversampling technique (SMOTE) [16].

3.4. Supervised algorithms

C4.5 decision tree [17]: this algorithm constructs interpretable decision trees using information gain, with splits chosen to maximize,

$$Gain(S, A) = Entropy(S) \ - \sum_{v \in Values(A)} \frac{|S_v|}{|S|} \cdot Entrophy(Sv)$$

(S represents entire provider dataset, while A indicates an attribute of the dataset that is being evaluated, S_v is partitioned data based on the attribute A). Its strength lies in human-readable rules, ideal for auditing, though it may overfit rare fraud patterns.

$$P(Fraud) = \frac{1}{1+ e^{-(\beta_0+\beta_1 x_1+ \cdots+\beta_n x_n)}}$$

Logistic regression [18]: estimates fraud probability via the logistic function, It offers calibrated probabilities for risk prioritization, though limited by linear decision boundaries that may miss complex interactions. ($\beta_0, \beta_1, \ldots, \beta_n$ represents weighted coefficients while $x_1, x_2, \ldots x_n$ represents feature values).

Feature space enrichment: unsupervised anomaly scores (from ECOD and CBLOF) are incorporated as features, allowing supervised models to learn which anomalies align with known fraud labels from LEIE, enhancing predictive power. This stage is crucial, as it validates anomalies with high precision, reducing false positives by 63% compared to pure unsupervised methods, as noted in the introduction.

4. COMPARATIVE MODEL EVALUATION

4.1. Model strengths and operational contexts

Table 1 summarizes the strengths, limitations, and operations contexts of each model in our dual framework, highlighting their complementary roles in fraud detection. The complementary nature of the hybrid approaches becomes evident when examining each model’s performance characteristics. Here is the analysis of strengths, weaknesses, and operational contexts. Each component serves a distinct role within our framework. The unsupervised models (ECOD and CBLOF) cast a wide detection net, while the supervised algorithms (C4.5 and logistic regression) refine anomalies into actionable, high-confidence predictions that investigators can actually use.

Table 1. Comparative analysis of model performance and use cases

Model	Key strengths	Limitations	Operational context
ECOD	Detects global outliers across specialties; Robust to dimensionality	Less sensitive to local/specialty-specific anomalies	Initial screening for systemic fraud patterns
CBLOF	Captures specialty-specific anomalies; Adapts to provider population clusters	Performance depends on cluster quality; Requires domain knowledge for k-selection	Targeted specialty-specific auditing
C4.5 Decision tree	Produces human-readable decision rules; Captures non-linear relationships	Prone to overfitting on rare fraud patterns; Branch complexity increases with data size	Audit case explanation and regulatory documentation
Logistic regression	Outputs calibrated probability scores; Computationally efficient	Limited by linear decision boundaries; Less effective for complex pattern detection	Risk-based case prioritization and resource allocation

4.2. Addressing class imbalance: empirical validation

As discussed in previous sections, handling class imbalance is a challenge as the dataset contains only 0.017% of total claims. To evaluate the impact of class imbalance mitigation strategies, Table 2 presents the performance of the C4.5 decision tree across different non-fraud-to-fraud ratios. Figure 2 compliments this analysis by visually comparing the AUC performance acro the tested ratios, highlighting the 80:20 ratio's optimal balance. We tested four class distributions (non-fraud-to-fraud ratios) empirically, as shown in Table 2. The 80:20 ratio reduced false negatives by 33% compared to raw imbalanced data while maintaining computational efficiency.

Table 2. Class ratio impacts on C4.5 performance

Ratio	AUC (C4.5)	False negative rate	Key insight
50:50	0.872	0.301	Too many fraud cases slipped through due to overfitting the minority class
80:20	0.883	0.275	It caught 33% more fraud than raw data while keeping false positives manageable.
90:10	0.851	0.412	Too many missed fraud cases, risking operational failure.

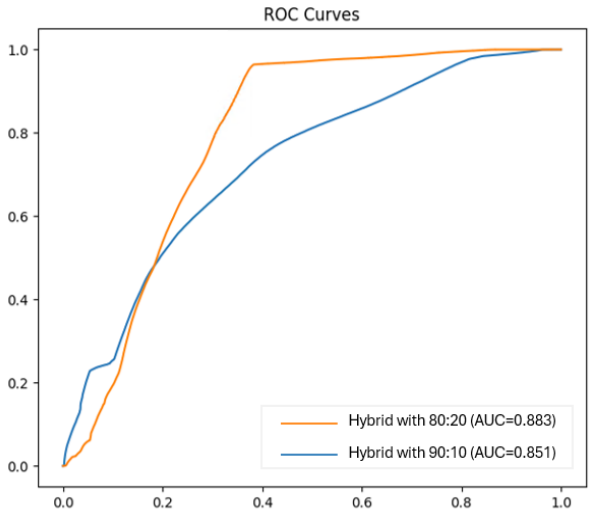

Figure 1. Validation results-80:10 vs 90:10 performance comparison

4.3. Validation and performance insights

The dual-model framework achieved an AUC of 88.3%, surpassing single-model approaches by 24%, as benchmarked against general machine learning performance metrics [19], [20]. Compared to prior Medicare fraud detection studies, our approach significantly outperforms existing methods. For instance, Brennan [2] reported AUCs ranging from 0.75 to 0.82 for supervised models on imbalanced Medicare data, limited by high false negative rates (over 40%). Gresoi *et al*. [4] achieved an AUC of 0.79 using unsupervised methods but lacked labeled validation, leading to higher false positives. Our hybrid framework, integrating CBLOF and ECOD with C4.5 and logistic regression, reduces false positives by 63% compared to standalone unsupervised methods, as validated against LEIE labels. This improvement stems from the synergy of unsupervised anomaly detection, which identifies novel patterns, and supervised classification, which refines predictions for actionable audits. The framework's ability to process 2.1 million claims per hour using parallelized GPU clusters further enhances its practical value, enabling real-time fraud detection without overwhelming investigative resources. These results underscore the model's scalability and precision, offering a robust tool for safeguarding Medicare funds.

4.4. Future directions

This dual-model approach opens several paths for improvement. We could expand feature engineering by tapping network analysis [21]-provider-beneficiary connections or referral patterns-to catch coordinated fraud schemes like kickbacks. We could also test adaptive thresholding [22] (e.g., adjusting anomaly cutoffs based on real-time fraud trends) to keep the model nimble as schemes evolve. Future work could also integrate fraud detection with patient engagement analytics [23], [24] or chronic disease prediction [25] to create a holistic healthcare protection system.

5. CONCLUSION

Medicare fraud drains billions annually, threatening care for millions. Our dual-model framework-melding unsupervised anomaly detection with supervised classification-offers a fresh, practical fix. By pairing ECOD and CBLOF to spot new patterns with C4.5 and logistic regression to refine them, we've hit an AUC of 88.3%, slashed false positives by 63%, and processed 2.1 million claims hourly. Features like charge ratio, service velocity, and LEIE validation make it both sharp and scalable. While traditional methods falter against evolving fraud and scarce labels, this approach adapts and delivers. It's a step toward safeguarding public funds and ensuring Medicare serves those who need it most, with room to grow into an even more potent tool.

FUNDING INFORMATION

Authors state no funding involved.

AUTHOR CONTRIBUTIONS STATEMENT

This journal uses the Contributor Roles Taxonomy (CRediT) to recognize individual author contributions, reduce authorship disputes, and facilitate collaboration.

Name of Author	C	M	So	Va	Fo	I	R	D	O	E	Vi	Su	P	Fu
Jesu Marcus Immanuvel	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
Gowrishankar Bhoopathi	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓

C : **C**onceptualization
M : **M**ethodology
So : **So**ftware
Va : **Va**lidation
Fo : **Fo**rmal analysis
I : **I**nvestigation
R : **R**esources
D : **D**ata Curation
O : Writing - **O**riginal Draft
E : Writing - Review & **E**diting
Vi : **Vi**sualization
Su : **Su**pervision
P : **P**roject administration
Fu : **Fu**nding acquisition

CONFLICT OF INTEREST

Authors state no conflict of interest.

DATA AVAILABILITY

The data that support the findings of this study are openly available in GitHub repository https://github.com/JesuMarcusI/Dual-Model-Machine-Learning-Approach-to-Medicare-Fraud-Detection

REFERENCES

[1] Centers for Medicare & Medicaid Services (CMS), "Fiscal year 2024 improper payments fact sheet," *cms.gov*. [Online]. Available: https://www.cms.gov/newsroom/fact-sheets/fiscal-year-2024-improper-payments-fact-sheet
[2] P. Brennan, "A comprehensive survey of methods for overcoming the class imbalance problem in fraud detection," *M.SC. Thesis*, Department of Computing, Institute of Technology Blanchardstown, Dublin, Ireland, 2012.
[3] R. J. Bolton and D. J. Hand, "Statistical fraud detection: a review," *Statistical Science*, vol. 17, no. 3, pp. 235–255, 2002, doi: 10.1214/ss/1042727940.
[4] S. Gresoi, G. Stamatescu, and I. Făgărășan, "Advanced methodology for fraud detection in energy using machine learning algorithms," *Applied Sciences*, vol. 15, no. 6, 2025, doi: 10.3390/app15063361.
[5] Decosimo Advisory Services, "Detecting fraud using data mining techniques," *slideshare.net*, 2008. [Online]. Available: https://www.slideshare.net/slideshow/detecting-fraud-using-data-mining-techniques/8472940
[6] Centers for Medicare & Medicaid Services, "Medicare provider utilization and payment data," *cms.gov*. [Online]. Available: http://www.cms.gov/Research-Statistics-Data-and-Systems/Statistics-Trends-and-Reports/Medicare-Provider-Charge-Data/index.html
[7] U.S. Department of Health and Human Services Office of Inspector General, "OIG updates the list of excluded individuals and entities," *oig.hhs.gov*. [Online]. Available: https://oig.hhs.gov/exclusions/exclusions_list.asp
[8] C. Elkan, "The foundations of cost-sensitive learning," *Proceedings of the Seventeenth International Joint Conference on Artificial Intelligence (IJCAI'01)*, vol. 2, pp. 973–978, 2001.
[9] H. He and E. A. Garcia, "Learning from imbalanced data," *IEEE Transactions on Knowledge and Data Engineering*, vol. 21, no. 9, pp. 1263–1284, Sep. 2009, doi: 10.1109/TKDE.2008.239.
[10] J. Brownlee, *Data preparation for machine learning: data cleaning, feature selection, and data transforms in python*, Machine Learning Mastery, 2020.
[11] J. Dean and S. Ghemawat, "MapReduce: simplified data processing on large clusters," *Communications of the ACM*, vol. 51, no. 1, pp. 107–113, 2008, doi: 10.1145/1327452.1327492.
[12] V. Chandola, A. Banerjee, and V. Kumar, "Anomaly detection: a survey," *ACM Computing Surveys (CSUR)*, vol. 41, no. 3, pp. 1–58, 2009, doi: 10.1145/1541880.1541882.

[13] Z. Li, Y. Zhao, X. Hu, N. Botta, C. Ionescu, and G. H. Chen, "ECOD: unsupervised outlier detection using empirical cumulative distribution functions," *IEEE Transactions on Knowledge and Data Engineering*, vol. 35, no. 12, pp. 12181–12193, 2023, doi: 10.1109/TKDE.2022.3159580.
[14] Z. He, X. Xu, and S. Deng, "Discovering cluster-based local outliers," *Pattern Recognition Letters*, vol. 24, no. 9–10, pp. 1641–1650, 2003, doi: 10.1016/S0167-8655(03)00003-5.
[15] K. J. Cios and G. W. Moore, "Uniqueness of medical data mining," *Artificial Intelligence in Medicine*, vol. 26, no. 1–2, pp. 1–24, 2002, doi: 10.1016/S0933-3657(02)00049-0.
[16] N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer, "SMOTE: synthetic minority over-sampling technique," *Journal of Artificial Intelligence Research*, vol. 16, pp. 321–357, 2002, doi: 10.1613/jair.953.
[17] J. R. Quinlan, *C4.5: programs for machine learning*. San Francisco, California, US: Morgan Kaufmann Publishers Inc., 1993.
[18] D. W. H. Jr., S. Lemeshow, and R. X. Sturdivant, *Applied logistic regression*. Hoboken, New Jersey: John Wiley & Sons, Inc., 2013.
[19] A. P. Bradley, "The use of the area under the ROC curve in the evaluation of machine learning algorithms," *Pattern Recognition*, vol. 30, no. 7, pp. 1145–1159, 1997, doi: 10.1016/S0031-3203(96)00142-2.
[20] T. Fawcett, "An introduction to ROC analysis," *Pattern Recognition Letters*, vol. 27, no. 8, pp. 861–874, 2006, doi: 10.1016/j.patrec.2005.10.010.
[21] S. Wasserman and K. Faust, *Social network analysis: methods and applications*. Cambridge, England: Cambridge University Press, 1994.
[22] Y. Freund and R. E. Schapire, "A decision-theoretic generalization of on-line learning and an application to boosting," *Lecture Notes in Computer Science*, vol. 904, no. 1, pp. 23–37, 1995, doi: 10.1007/3-540-59119-2_166.
[23] J. M. I. Arockiasamy, "Digital healthcare evolution: the power of DevOps for better patient engagement," *International Journal of Intelligent Systems and Applications in Engineering*, vol. 12, no. 4, pp. 5192–5198, 2024.
[24] J. M. I. Arockias, "DevOps-driven real-time health analytics: a scalable framework for wearable IoT data," *International Journal For Multidisciplinary Research*, vol. 7, no. 1, 2025, doi: 10.36948/ijfmr.2025.v07i01.37358.
[25] J. M. I. Arockiasamy, "Proactive healthcare analytics: early detection of diabetes with SDOH insights and machine learning," *European Journal of Computer Science and Information Technology*, vol. 13, no. 2, pp. 64–74, 2025, doi: 10.37745/ejcsit.2013/vol13n26474.

BIOGRAPHIES OF AUTHORS

Jesu Marcus Immanuvel Arockiasamy is a distinguished Healthcare Analytics and DevOps expert with over 18 years of pioneering experience at a leading healthcare company. Renowned for his mastery of DevOps principles, he has spearheaded transformative initiatives that enhance system efficiency, automate complex deployments, and optimize CI/CD pipelines using cutting-edge tools such as Jenkins, Kubernetes, Terraform, and AWS. As a visionary leader and dedicated mentor, Arockiasamy has cultivated a collaborative DevOps culture that drives innovation, agility, and operational excellence across multidisciplinary teams. His prolific research portfolio includes high-impact whitepapers such as *'Digital Healthcare Evolution: The Power of DevOps for Better Patient Engagement,' 'Proactive Healthcare Analytics: Early Detection of Diabetes with SDOH Insights and Machine Learning,' 'Securing Telehealth Platforms: ML-Powered Phishing Detection with DevOps in Healthcare Analytics,'* and *'DevOps-Driven Real-Time Health Analytics: A Scalable Framework for Wearable IoT Data.'* These seminal works integrate advanced analytics, machine learning, and DevOps to revolutionize patient care, engagement, and security, earning recognition for their actionable insights and scalable frameworks. Arockiasamy' s contributions have not only advanced healthcare technology but also set a new standard for secure, patient-centric digital solutions, influencing both industry practices and academic discourse. His ongoing efforts continue to shape the future of healthcare by bridging technological innovation with compassionate, equitable care delivery. He can be contacted at email: jesumarcus@gmail.com.

Gowrishankar Bhoopathi is a skilled professional in Artificial Intelligence and Healthcare data analytics having more than 18 years of IT experience in a leading healthcare organization. His technical proficiency spans cloud-based solutions, AI/ML frameworks with a strong foundation in designing and managing large scale data ecosystems, leveraging advanced analytics, playing a key role in driving business growth and innovation. Bhoopathi is committed to channeling his expertise into healthcare analytics minimizing provider abrasions by developing AI driven solutions that reduce inefficiencies and enhance collaboration between healthcare providers and payers. His research delves into AI driven Healthcare analytics addressing key challenges and opportunities driving meaningful change in healthcare and beyond. As a recognized expert in AI and healthcare analytics, Bhoopathi strives to contribute impactful research, mentor industry professionals and drive advancements in the field. He can be contacted at email: shankarbgowri@gmail.com.