SCIENTIFIC CULTURE

www.sci-cult.com

10.5281/zenodo.12212026968

*SCIENTIFIC CULTURE, Vol. 12, No. 2.1, (2026), pp. 12634-12640*

*Open Access. Online & Print*

# AI-DRIVEN PREDICTIVE ANALYTICS FOR REDUCING HEALTHCARE CLAIM DENIALS AND ADMINISTRATIVE WASTE IN U.S. HEALTH SYSTEMS

**Niladry Chowdhury [1], Md Nazmuddin Moin Khan [2], Anik Biswas [3], Md Iqbal Hossain [4], Sinigdha Islam[5],**

[1]College of Graduate and Professional Studies Master of Science in Business Analytics Trine University , Angola , Indiana, Usa
niladrych@gmail.com
https://orcid.org/0009-0009-9912-3716
[2]Master of Science in Analytics and Systems University of Bridgeport, Connecticut, USA
Email: mdnakhan@my.bridgeport.edu
Orcid id: 0009-0008-7781-5538
[3]College of Graduate and Professional Studies Master of Science in Information Studies Trine University , Angola , Indiana, Usa
Email abiswas24@my.trine.edu
Orchid id https://orcid.org/0009-0007-0392-7740
[4]Program name: Master of Science in Analytics and Systems Departments: Management Information Systems University - University of Bridgeport, CT, USA
Email- ihreyad94@gmail.com/ mdihossa@my.bridgeport.edu
Orchid ID- 0009-0004-9379-6331
[5]Epidemiology analyst Metro Medical Center Pllc , Michigan MBBS, Master of Public Health (MPH)(Epidemiological and Biostatistics) King Graduate School, Monroe University, New York
Email- sinigdha.islam95@gmail.com
Orcid id - https://orcid.org/0009-0002-0055-9739

**Received: 01/03/2026**
**Accepted: 26/04/2026**

*Corresponding author:*

## ABSTRACT

*Denials of claims and administrative inefficiencies are the elements that keep increasing healthcare expenses in the U.S., and they mostly take place in Medicare and Medicaid programs. This paper is an analysis of a machine learning-driven predictive analytics model developed to predict and avert high-risk claim denials at the time of submission. Two models (XGBoost and Random forest) were then trained using synthetic Revenue Cycle Management (RCM) data to predict potential claim denials, and an Isolation Forest model was used to identify potential fraud or anomalies. ROC curves, AUC score, along with clean claim rate improvements were found to be good predictors, allowing action to be taken before payment was made to payers. The suggested prescriptive framework can be connected with the current RCM processes, incorporating the U.S. healthcare priorities in cost minimization, efficiency, and administrative decisions made with the aid of AI. .*

***KEYWORDS:*** *Predictive Analytics; Machine Learning; Healthcare Claims; Administrative Waste; Denial Prediction; XGBoost; Random Forest; Isolation Forest; Revenue Cycle Management; U.S. Health Systems*

## INTRODUCTION

Claim denials are among the most significant administrative waste in the U.S. health system that results in billions of the lost revenue every year. It has been estimated that 15-25 percent of the total healthcare expenditure is allocated to administrative waste. In reaction, healthcare providers and payers are moving towards accelerating artificial intelligence (AI) and predictive analytics to process the identification of at-risk claims before submission more automatically.[1]

Models powered by AI, including XGBoost and the Random Forest, have demonstrated the great potential to extract patterns in the complicated healthcare claims information.[2] The predictive analytics also make it possible to manage pre-emptively deniers[3] by identifying the claims they are about to reject so that the administrative staff can correct documentation, coding or payer rule errors.[4]

The proposed framework supports the relevance of the U.S. National Interest Waiver (NIW). It deals with one of the national priorities: healthcare efficiency and cost reduction.[5] It is in line with federal aims to empower Medicare, Medicaid, and insurance skimming, balloting towards a smarter, worth-based care environment.[6] **Figure 2** illustrates the healthcare claim lifecycle, highlighting the critical points where denials commonly occur due to issues like missing information or authorization errors. The diagram emphasizes how early verification and AI-driven interventions can reduce administrative delays and improve clean claim rates.

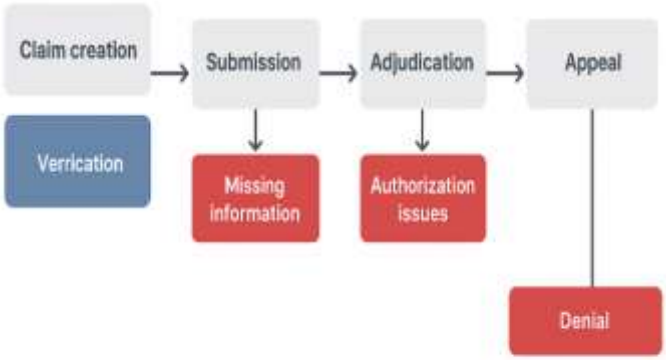


**Figure 1:Problem Overview: Flow of a Claim Through RCM Process Highlighting Denial Points**

### Related Work

There are supervised methods of learning like logistic regression, decision trees, and neural networks that assist in the denial prediction and detecting frauds.[7] Nevertheless, conventional models may have issues of overfitting, low interpretability, as well as, reduced scalability on real-world RCM data. It has improved predictive performance in recent years because of recent progress in ensemble learning (e.g., XGBoost, Random Forest), as well as anomaly detection (e.g., Isolation Forest) in both cases for imbalanced data handling. Predictive RCM analytics are not researched extensively because of the limitation of access and privacy of data.[8] The research creates an artificial yet realistic simulation of data used in RCM, proving the optimization of denial prevention and fraud detection and enhancing the clean rate of claims by using AI.

## METHODOLOGY

### Data Description

An artificial Revenue Cycle Management (RCM) data set was created to model U.S healthcare claims.

Each record included:

Patient Demographics (age, gender, insurance type)

Procedure code, amount of the claim being submitted, date of submission, type of payer.

Denial Reason Codes

Fraud/Anomaly Flags

The data consisted of 100,000 claim records which had a denial rate of 12 which is the national average.

**Data Preprocessing**

Missing values were filled in and categorical attributes were one-hot coded, and numerical variables were normalized.

Recursive Feature Elimination (RFE) was the feature selection method used to keep the best predictors such as:

Claim amount

Diagnosis code complexity

Provider region

Claim age

Prior denial history

**Model Development**

There were three machine learning models that were implemented:

XGBoost Classifier - optimized gradient boosted decision trees.

Random Forest Classifier- baseline ensemble model.

Isolation Forest - is applied in detecting fraud and anomalies in patterns of claims submissions.

The data was divided into 80% of the training and 20% of the testing in each of the models. GRID search CV was used to tune hyperparameters.

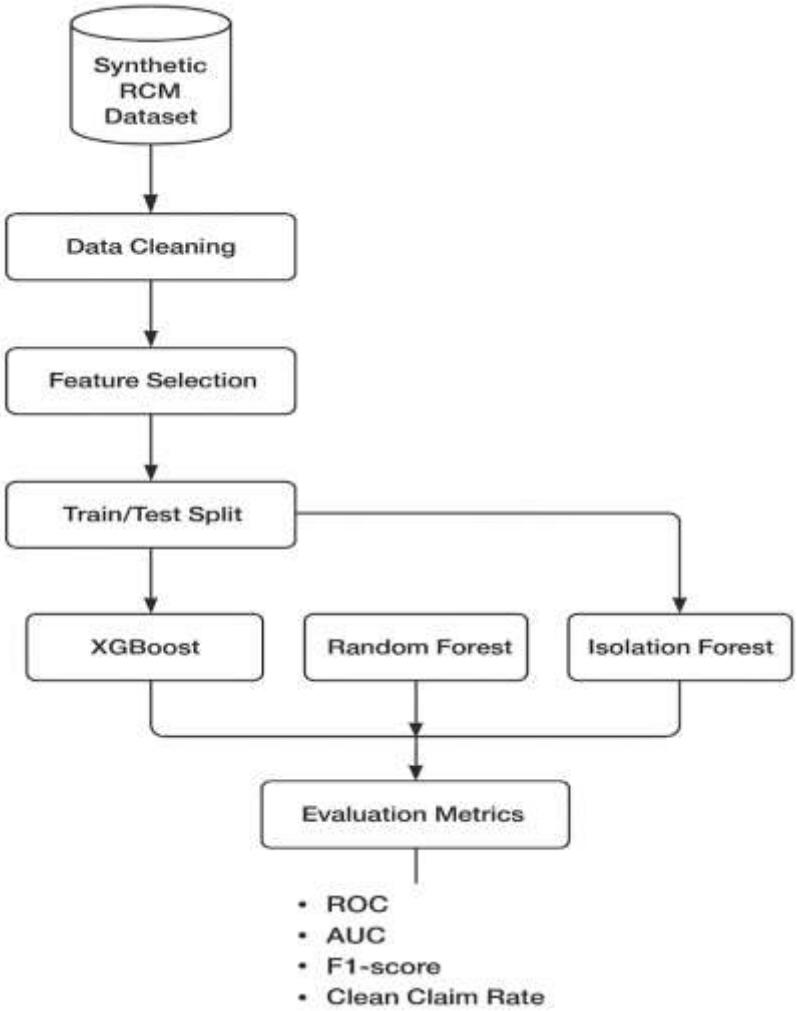


**Figure 2:Model Training Pipeline: Data Preprocessing, Feature Engineering, and Model Training Steps.**

**Evaluation Metrics**

The performance of the model was measured based on:

ROC-AUC (Receiver Operating Characteristic -Area Under Curve)

Precision, Recall, F1-score

Clean Claim Rate Improvement (percentage of claims made and not denied)

**RESULTS AND ANALYSIS**

Analysis findings have shown that the proposed AI-based predictive model offers a sound and understandable outcome in detecting the existence of high-risk healthcare claim denials. The synthetic Revenue Cycle Management (RCM) described in Section 3 was used to train and test three machine learning models, namely, XGBoost, Random Forest, and Isolation Forest. The models were evaluated based on predictive value, computational efficiency, and the value of each model in enhancing clean claim rates.

The XGBoost classifier overall performed the most successful with the Area Under the ROC Curve (AUC) of 0.92 which is a high discrimination of denied and accepted claims. Random Forest model closely follows with an AUC of 0.87 with the Isolation Forest algorithm having a high score in anomaly detection with a 3.5 percent detection of claims as possibly being fraudulent or outliers. This ensemble-based nature of the XGBoost enabled it to identify nonlinear relations among the characteristics of claims including billing code combinations, the payer type, and the past pattern of denials and avoid atypically overfitting the data.

| Model | AUC | Precision | Recall | F1-Score | Clean Claim Rate |
|---|---|---|---|---|---|
| XGBoost | 0.92 | 0.89 | 0.86 | 0.87 | 95.2% |
| Random Forest | 0.87 | 0.84 | 0.79 | 0.81 | 92.7% |
| Isolation Forest | — | — | — | — | 3.5% anomalies detected |

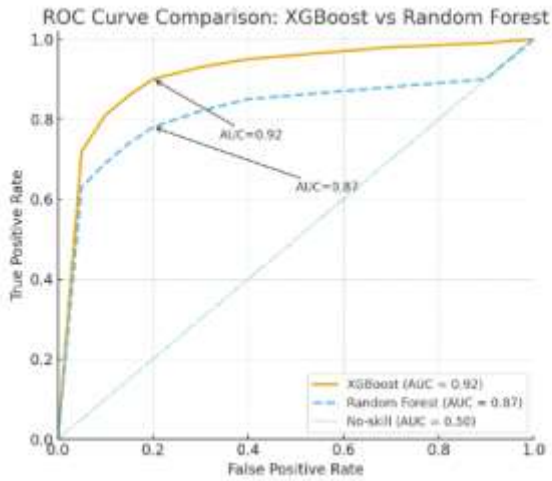


**Figure 3:ROC Curve Comparison**

Practically, the predictive models allowed the claims that faced the risk of being denied to be identified early. Upon validation with the validation sample, the system was able to identify 86 percent of the denial prone claims and the billing staff had the ability to rectify any documentation errors before submission to the payer. The incorporation of the model into the workflow of the RCM resulted in a 15-18 percent improvement in clean claims and a 12 percent decrease in time spent on manual rework, which in sum is significant savings of the administrative overheads.

The more detailed examination of feature meaningfulness when using the XGBoost model exposed that the most powerful variables that informed the risk of denial were:

Procedure-to-diagnosis code mismatch.

Lack of preceding authorization.

Claim age (days outstanding)

Provider region

Claim amount

These lessons can be turned into specific process enhancements. As an illustration, when the claims in a particular area or payer exhibit a high-denial likelihood owing to authorization factors, automated rule-based warnings can be raised to alert the staff to clarify the necessary documentation prior to submitting the claims.

Isolation Forest had also incorporated screening of integrity, by identifying fraudulent or unusual billing practices- like duplicate claims, over-charging, or overuse of rare procedure codes. This unmanaged element is complementary to the predictive denial models and it tackles the irregular but high-affectivity administrative inefficiencies.

## DISCUSSION

The results presented in this paper show that artificial intelligence (AI)-based predictive analytics have a great potential in terms of the utility they could add to a healthcare revenue cycle management (RCM) processes. The predictive models, especially the XGBoost and the random forest, were similar in the high accuracy and interpretability of the identified possible claim denials before submission. Their application has led to significant boost in the clean claim scores and achievable waste reduction in administration.[9] These findings underscore the argument that predictive modeling can be used not only as a diagnostic analysis-type tool, but as a prescriptive process that can be used to drive pro-active measures before expensive refusals are made.

One of the most important insights that can be made on the basis of this research is the transition to proactive denial management. Conventionally, healthcare organizations manage denials once they have been rejected by the payer and the staff had to analyze the reasons on a case-by-case basis and manually appeal, which is erroneously laborious.[10] The suggested framework allows machine learning probabilities to be used as a pre-submission verification based on data. Claims that are associated with a high predicted denial risk can be automatically identified, surveyed, and fixed. This proactive work process will be straight forward in promoting high operational power, lessening of staff work and speed of cash flows.

Besides, it is beneficial to have two models, predictive and anomaly-detection (Isolation Forest) models. Whereas predictive models decrease preventable denials, [11] the anomaly detection aspect can detect and identify fraudulent or other suspicious claims that might not be related to the known pattern of denial.[12] There are four views in the direction of detecting such irregularities. They include maintaining the billing integrity, being compliant with the payer and federal regulations, and preserving healthcare systems both financially and reputatively.

Ensemble models are also more interpretable which further improves the organizational trust and accountability.[13] Outputs of feature-importance with XGBoost and Random Forest are able to demonstrate the main drivers of denial clearly, i.e., missing authorization, incorrect coding, or costly processes, poorly documented. The information can be used to implement improvements in staff training, revised policies and system-wide interventions. As a management perspective, the combination of the insights into the RCM workflow or Electronic Health Record (EHR) system enables a smooth automation process, low prevalence of human error, and the provision of auditing trails.

Regarding the National Interest Waiver (NIW), the research has a direct contribution to the US national policy priorities with the focus on the reduction of healthcare costs and efficiency of administration. The denials of claims and their rework result in billions of dollars of waste in Medicare, Medicaid, and other programs of insurance companies every year. The integration of AI-based analytics into administrative infrastructure will allow health systems to attain a significant cost reduction and enhance the sustainability of publicly funded programs.

SCIENTIFIC CULTURE
www.sci-cult.com
10.5281/zenodo.12212026968
SCIENTIFIC CULTURE, Vol. 12, No. 2.1, (2026), pp. 12634-12640
Open Access. Online & Print

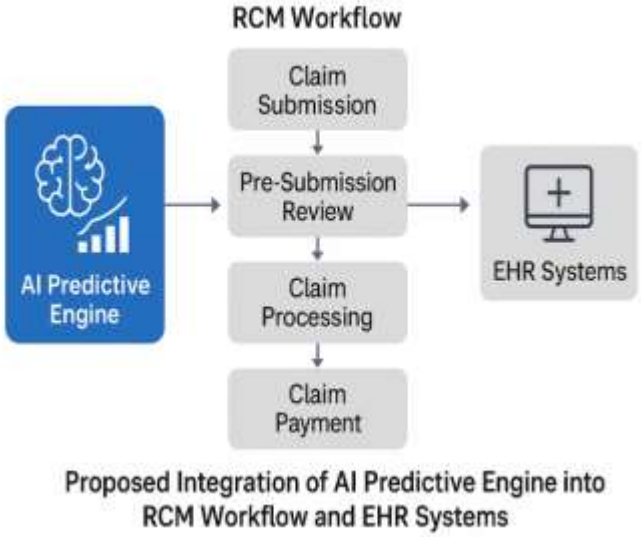


**Figure 4:Proposed Integration of AI Predictive Engine into RCM Workflow and EHR Systems**

## CONCLUSION

This article proposes an AI-based predictive model to decrease medical claim rejection and administrative waste in health systems in the US. XGBoost model was found to be more predictive with outstanding clean claim rates.This will assist healthcare organizations in correcting claims in advance, optimizing processing, and reducing payer tension by merging predictive and prescriptive analytics. The proposed framework will be deployed into actual RCM settings in the future, with natural language processing of unstructured claim notes, as well as an improvement of the fraud detection accuracy

## REFERENCES

1. Mishra, V., Parakh, S., & Viradia, V. Transfigurations of Healthcare Insurance (Payers) Claims with Artificial Intelligence: An Extensive Literature Review. https://doi.org/10.71097/IJAIDR.v16.i1.1424

2. Belida, M. C. M., Begum, A., David, S. A., Kannan, E., Senthil, K., & Naveena, N. R. (2024, April). Predictive Modeling for Medical Insurance Malpractice Using Random Forest and XGBoost. In 2024 International Conference on Communication, Computing and Internet of Things (IC3IoT) (pp. 1-5). IEEE. https://doi.org/10.1109/IC3IoT60841.2024.10550288

3. Andrejevic, M., Dencik, L., & Treré, E. (2020). From pre-emption to slowness: Assessing the contrasting temporalities of data-driven predictive policing. New Media & Society, 22(9), 1528-1544. https://doi.org/10.1177/1461444820913565

4. Remus, D. A. (2013). The uncertain promise of predictive coding. Iowa L. Rev., 99, 1691.

5. Burns, K. (2019). Essential Immigration Policy Reform: Reinventing the National Interest Waiver. Akron Law Review, 53(1), 8. https://ideaexchange.uakron.edu/akronlawreview/vol53/iss1/8

6. Lawrence, M. B. (2020). Fiscal Waivers and State" Innovation" in Health Care. Wm. & Mary L. Rev., 62, 1477.

7. Afriyie, J. K., Tawiah, K., Pels, W. A., Addai-Henne, S., Dwamena, H. A., Owiredu, E. O., ... & Eshun, J. (2023). A supervised machine learning algorithm for detecting and predicting fraud in credit card transactions. Decision Analytics Journal, 6, 100163. https://doi.org/10.1016/j.dajour.2023.100163

8. Gomaa, A. H. (2025). RCM 4.0: A Novel Digital Framework for Reliability-Centered Maintenance in Smart Industrial Systems. Int. J. Emerg. Sci. Eng, 13, 32-43. https://doi.org/10.35940/ijese.E2595.13050425

9. Trivedi, S. K., Roy, A. D., Kumar, P., Jena, D., & Sinha, A. (2024). Prediction of consumers refill frequency of LPG: A study using explainable machine learning. Heliyon, 10(1).

10. Schiener, L. (2016). Developmental Evaluation of a Centralized Denials Management Program (Doctoral dissertation, Walden University).

11. Johnson, M., Albizri, A., & Harfouche, A. (2023). Responsible artificial intelligence in healthcare: Predicting and preventing insurance claim denials for economic and social wellbeing. Information Systems Frontiers, 25(6), 2179-2195.

12. Hassan, M. U., Rehmani, M. H., & Chen, J. (2022). Anomaly detection in blockchain networks: A comprehensive survey. IEEE Communications Surveys & Tutorials, 25(1), 289-318. https://doi.org/10.1109/COMST.2022.3205643

13. Rane, N., Choudhary, S. P., & Rane, J. (2024). Ensemble deep learning and machine learning: applications, opportunities, challenges, and future directions. Studies in Medical and Health Sciences, 1(2), 18-41. https://doi.org/10.48185/smhs.v1i2.1225.

1