IJIRCCE

e-ISSN: 2320-9801 | p-ISSN: 2320-9798

INTERNATIONAL JOURNAL OF INNOVATIVE RESEARCH

IN COMPUTER & COMMUNICATION ENGINEERING

Volume 11, Issue 7, July 2023

INTERNATIONAL STANDARD SERIAL NUMBER INDIA

Impact Factor: 8.379

9940 572 462 | 6381 907 438 | ijircce@gmail.com | www.ijircce.com

International Journal of Innovative Research in Computer and Communication Engineering

| e-ISSN: 2320-9801, p-ISSN: 2320-9798| www.ijircce.com | |Impact Factor: 8.379 |

|| Volume 11, Issue 7, July 2023 ||

| DOI: 10.15680/IJIRCCE.2023.1107002 |

Medical Aid Fraud and Abuse Detection System Using Machine Learning

Faith Nyakudya

Masters Student, Dept. of Software Engineering., School of Information Sciences, Harare Institute of Technology, Zimbabwe

ABSTRACT: Healthcare fraud is a problem. Fraud is any deliberate, dishonest act committed with knowledge that it could result in an unauthorized benefit to the perpetrator or someone else. Healthcare fraud is an insurance fraud. This research aims to detect provider fraud based on payer claims. Medical fraud is a major issue. Peer providers, physicians, and patients make false claims in healthcare fraud. Zimbabwean law requires insurers to pay legitimate medical claims within 30 days. There's not enough time to investigate properly. This harms insurance companies the most. Medicare fraud has skyrocketed medical spending, according to the government. Healthcare fraud and abuse are widespread. Some common provider frauds are: Unearned fees, Double billing, misrepresenting medical care given, Billing for a complicated or expensive service not rendered and charging for uninsured services. This study's dataset includes inpatient, outpatient, beneficiary, and provider fraud labels. Based on this, key features like average claim value per doctor were created. The researcher then used 80:20 and 75:25 sampling ratios. Feature engineering was used with Logistic Regression, Decision Tree, and Random Forest. The final dataset has 360 features, 161 of which have more than 0.1% impact. Random Forest also performed best on these 161 features. Considering AUC and F1 Score, Random Forest is the best model for healthcare fraud detection. The final prototype designed is a human-decision support system for Insurers and auditors. It predicts fraudulent claims and helps auditors further investigate flagged claims. **Diagnosis Related Group (DRG)**

KEYWORDS: Hospital Discharge Charts (HDC), Diagnosis Related Group (DRG), Inpatient, Outpatient, Beneficiary, service provider, payer,

I. INTRODUCTION

Health care fraud is an international challenge and illegal act, where perpetrators create a zero-sum game for maintenance costs that affect service quality. Estimated Worldwide health care fraud costs up to 10%. Billions of dollars in health care fraud are claimed worldwide. [1]Healthcare is vital to people's lives. Human anatomy is complex. Specialist doctors are essential treat or nurse different diseases in the body. This causes various treatment procedures for patients of different specialties. [2][3]Health care aims to serve as many patients as they could possibly can. Every treatment has a charge attached to it. Doctors and pharmacists must be paid for their time and prowess. Patient prices are often unaffordable. Insurance schemes are used to spread costs across all medical patients and pay for medical staff and medical equipment used. Any insurance system can be abused or fraudulent. Providers, doctors, and beneficiaries commit healthcare fraud collaboratively. Legitimate healthcare claims must be paid within 30 days, per local regulations. Insurance companies have little time to investigate frauds. These bad practices make insurers vulnerable. According to the government, medical fraud has skyrocketed spending. Currently, a provider's claim must be manually checked for mismatched details, claim limits, and fraud. This is costly. [4]A statistical approach could help approve claims online, saving time and money. Healthcare fraud: Healthcare fraud is multifaceted. a) Unprovided services billed. b) Service claims. c) Misrepresenting the service. d) Charging more than what's provided. e) Misbilling a covered service. Corrupt pharmacists and doctors file false medical aid claims and pocket the money. Others treat patients and provide medication, but ask them to pay a huge shortfall when none exists. Corrupt institutions demand full medical aid fees after receiving "shortfall payments." For the same service, hospitals get double pay. [5]

| e-ISSN: 2320-9801, p-ISSN: 2320-9798| www.ijircce.com | |Impact Factor: 8.379 |

|| Volume 11, Issue 7, July 2023 ||

| DOI: 10.15680/IJIRCCE.2023.1107002 |

II. RELATED WORK

In healthcare anomaly detection, machine learning can be supervised or unsupervised. Unsupervised learning does not require data-labelling for training. Anomaly detection algorithms include neural network classification, genetic algorithms, SVMs, decision trees, KNN, etc.[6] Supervised learning algorithms detect fraud and abuse patterns better than unsupervised algorithms, which are partitioning, agglomerative, probabilistic, etc.

Three earlier studies examine Medicare utilization and payment data using descriptive statistics and correlations, not machine learning. [3]Use 2012 Medicare data to find correlations between a doctor's education and practices to detect misuse or inefficiencies. The authors compared medical-related variables such as charges, number of procedures, and payments to find anomalous behaviours.[7]also uses 2012 Medicare data for Urology. High utilization variability among Urologists could save 9%. The number of patient visits was also correlated with Medicare reimbursements. Pande et al. use older Medicare data and exclusions from the LEIE database to assess Medicare fraud perpetrators and what happens to them. Authors use descriptive statistics to find Medicare fraud patterns and make recommendations. Predictive models are recommended for detecting claims fraud. Descriptive statistics, correlations, etc. are useful, but they rely on humans to extract patterns from data.

Machine learning can reduce this reliance by automatically extracting patterns to produce meaningful results, such as detecting fraudulent behavior. Thornton et al. [8]used machine learning. The authors examine outlier detection techniques using Medicaid dental claims data. Medicaid is a separate program that covers low-income Americans [9]They use univariate linear regression, box plots, and time series plots, and multivariate clustering.

In one study, [10] authors use multivariate regression to estimate Medicare payments per provider type. This baseline is used to compare actual payment amounts, with outliers flagged. Another study uses a [6]two-step approach to detect Medicare fraud by provider. Multivariate regression model residuals are the first step. These residuals are fed into a Bayesian probability model to produce fraud probabilities. Their method outperformed other outlier detection methods. Bayesian models are used to detect Medicare fraud [[11], [12]]. The final studies [[13], [3]] are exploratory studies that look for fraudulent providers by procedure count. Authors predict provider type using Multinomial Naive Bayes. If the predicted provider type doesn't match what's expected, this provider is out of normal practice and should be investigated.[[14]] Our exploratory, comparative study uses many related methods and data sources. I use these methods to test learners' ability to predict Medicare fraud. Along with learners and data, I assess class imbalance using four performance metrics.

In one study[15], authors use multivariate regression to estimate Medicare payments per provider type. This baseline is used to compare actual payment amounts, with outliers flagged. Another study [13] uses a two-step approach to detect Medicare fraud by provider. Multivariate regression model residuals are the first step. These residuals are fed into a Bayesian probability model to produce fraud probabilities. Their method outperformed other outlier detection methods. Bayesian models are used to detect Medicare fraud [10, 12].

The final studies [14] are exploratory studies that look for fraudulent providers by procedure count. Authors predict provider type using Multinomial Naive Bayes. If the predicted provider type doesn't match what's expected, this provider is out of normal practice and should be investigated. Our exploratory, comparative study uses many related methods and data sources. In this research I'm are going to use these methods to test learners' ability to predict Medicare fraud. Along with learners and data, the researcher is going to assess class imbalance using performance metrics.

III. EXPLORATORY DATA ANALYSIS

1. First, we combined the dataset using inner join (I merged Inpatient and Outpatient patient IDs).
2. Merged patient data with beneficiary ID.
3. Inner joined the provider's class labels with the merged data frame.
4. Grouped the data frame by provider id and extracted total reimbursed per provider, average reimbursed per patient per provider, per patient claims per provider, etc.

IV. ALGORITHMS USED

Logistic regression: Logit models are used for classification and predictive analytics. Logistic regression estimates the probability of an event, like voting or not voting, based on independent variables. The dependent variable is between 0

| e-ISSN: 2320-9801, p-ISSN: 2320-9798| www.ijircce.com | |Impact Factor: 8.379 |

|| Volume 11, Issue 7, July 2023 ||

| DOI: 10.15680/IJIRCCE.2023.1107002 |

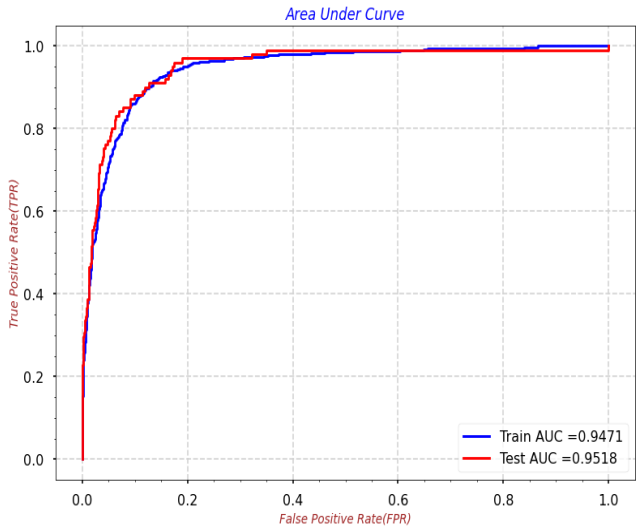
and 1 because the outcome is a probability. In logistic regression, the odds (probability of success divided by probability of failure) are logit transformed.

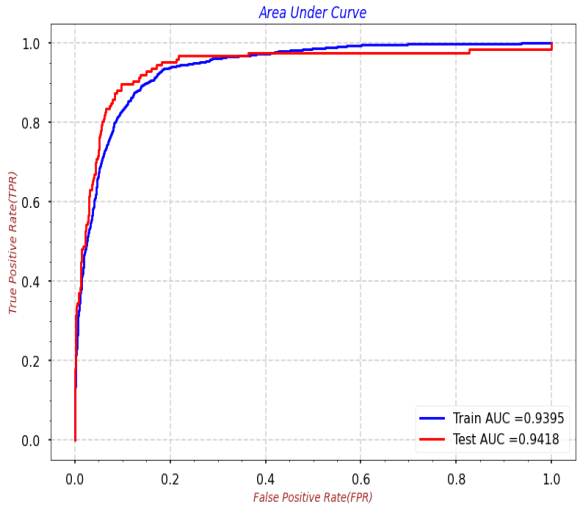
Decision Tree is a supervised learning technique used for classification and regression problems, but mostly classification. It's a tree-structured classifier where internal nodes represent dataset features, branches represent decision rules, and leaf nodes represent outcome. Decision trees have two nodes: Decision and Leaf. Decision nodes make decisions and have multiple branches, while Leaf nodes are the output and have no branches.

Random Forest uses supervised machine learning. ML Classification and Regression can use it. It uses ensemble learning to solve complex problems and improve model performance. Random Forest is a classifier that averages multiple decision trees on subsets of a dataset to improve predictive accuracy. The random forest predicts the final output based on the majority of predictions from each tree. More trees improve accuracy and prevent overfitting. Random Forest is created by combining N decision trees, and then predictions are made for each tree.

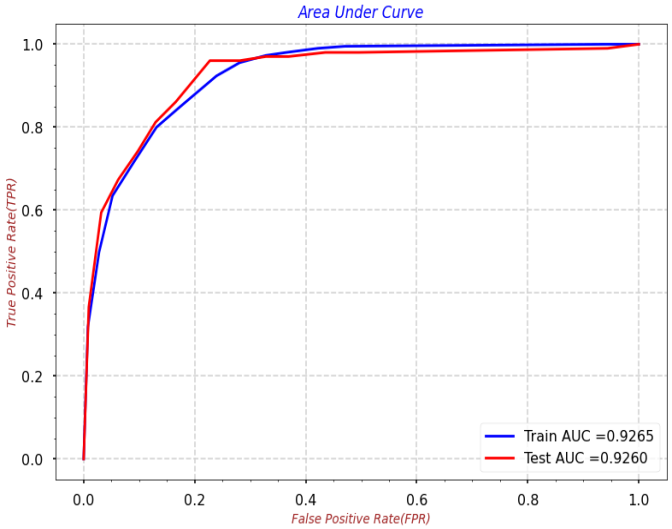
V. RESULTS OF ALGORITHM PERFORMANCES

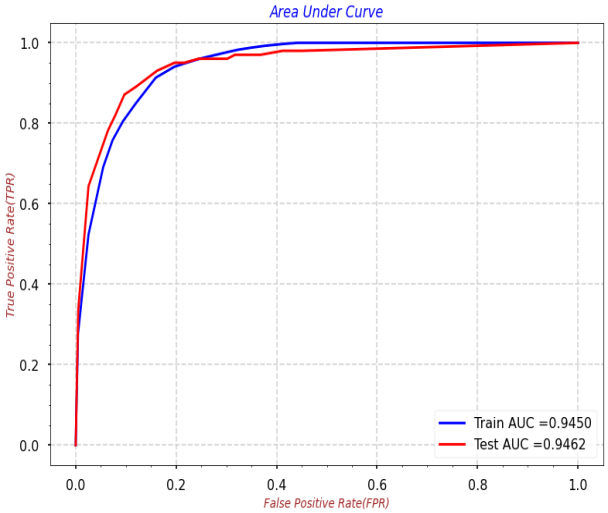
Logistic Regression Results





Decision Tree Results





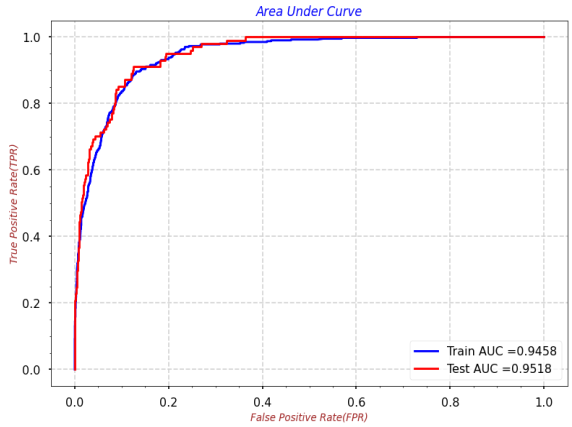
International Journal of Innovative Research in Computer and Communication Engineering

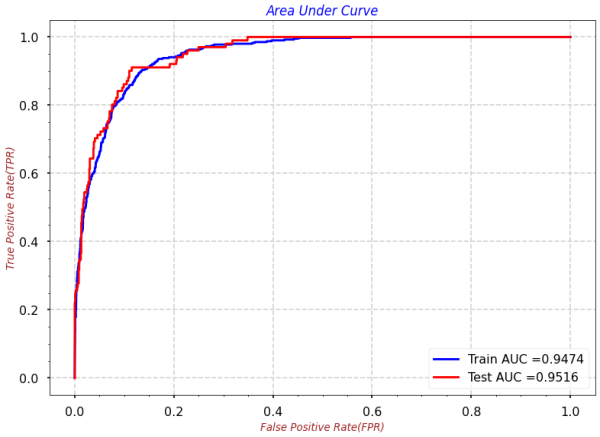
| e-ISSN: 2320-9801, p-ISSN: 2320-9798| www.ijircce.com | |Impact Factor: 8.379 |

|| Volume 11, Issue 7, July 2023 ||

| DOI: 10.15680/IJIRCCE.2023.1107002 |

Random Forest Results





VI. RESULT ANALYSIS

- Comparison Logistic Regression vs Random Forest: F1 score increased with Random Forest.
- Random Forest model outperforms the Logistic Regression model.
- False Negative (Predicted that there was no Fraud whereas there was Fraud) count is higher in Logistic Regression, which is dangerous for us.
- Random Forest outperforms Logistic Regression based on all the scores.
- Filtering important features doesn't improve LR and RF model performance.
- F1 score rises while False negative rises. False negatives are very essential as compared to false positives.
- Conclusively the model actually performs in a more efficient way when using all features are used.
- Random Forest performed as the most efficient model in detecting medical aid fraud based on AUC, F1 Score, and FNR.

VII. PROPOSED SYSTEM

The proposed system will work with the model perfectly with offline data. This model can be integrated to a system linking to a web api and ERP of different medical insurance providers and health claims systems. The model works with claim details and specific questions as input to determine fraud. This will definitely assist in approving non fraudulent claims faster without the manual functionalities. This helps the insurer focus on possibly fraudulent cases and auditors to perform further investigations.

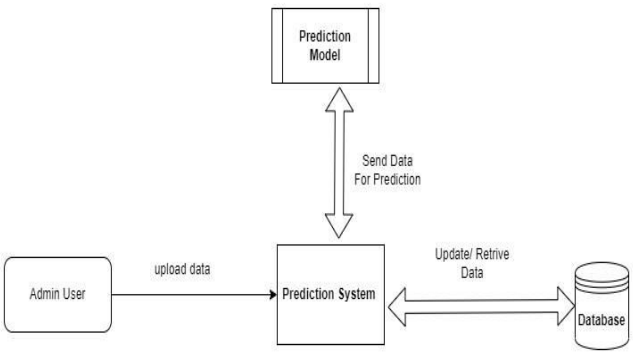


Fig.1. Context diagram of the proposed system

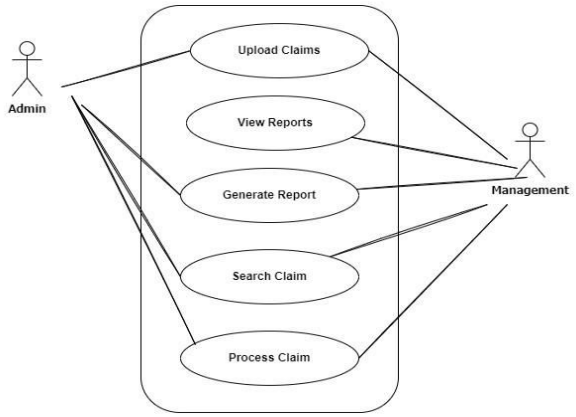


Fig. 2. Use Case Diagram

| e-ISSN: 2320-9801, p-ISSN: 2320-9798| www.ijircce.com | |Impact Factor: 8.379 |

|| Volume 11, Issue 7, July 2023 ||

| DOI: 10.15680/IJIRCCE.2023.1107002 |

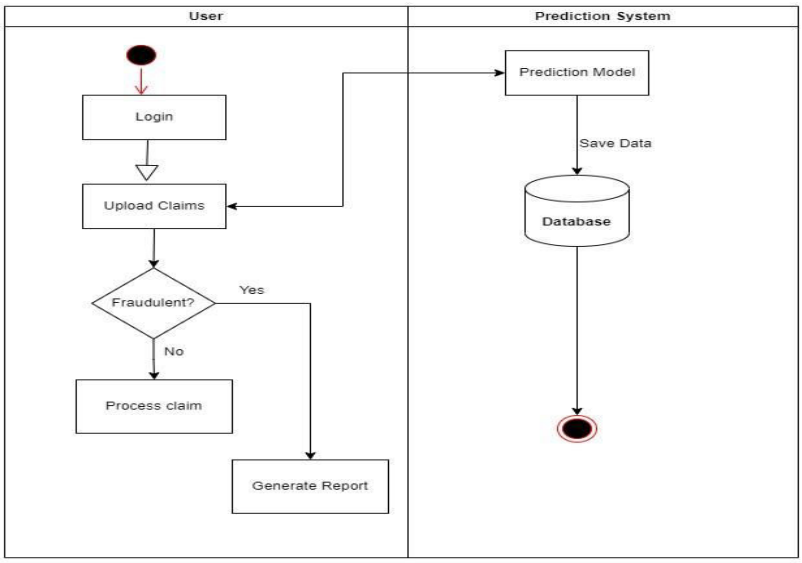


Fig. 3. UML-Activity Diagram

VIII. CONCLUSION AND FUTURE WORK

Using end-to-end multi-label prediction method for fraud correlation scores, we developed a model to detect and flag suspicious medical claims from service providers. This model requires an ordinary data type, is practical, and achieves better accuracy and recall than the traditional rule-based method, easing data analysts' work. Due to privacy issues, we couldn't collect enough training data, especially locally. In terms of ease of work for research the Government of Zimbabwe must enable researchers to have research access to government, hospital, and other institution databases is crucial to building a perfect model.

REFERENCES

[1] X. Fan, Z. Wang, and Z. Chen, "Impact of Diagnosis-Related Group Payment System on the Variation in Hospitalization Expenditures : A Case-Control Study with a Propensity-Score-Matched Control Group," pp. 1–16, 2021.

[2] S. Musau and T. Vian, "Fraud in Hospitals," *U4 ISsue*, vol. 5, no. 1, 2008.

[3] C. Zhang, X. Xiao, and C. Wu, "Medical fraud and abuse detection system based on machine learning," *Int. J. Environ. Res. Public Health*, vol. 17, no. 19, pp. 1–11, 2020, doi: 10.3390/ijerph17197265.

[4] Center for Medicare and Medicaid Services (CMS), "Common Types of Health Care Fraud," pp. 1–4, 2015, [Online]. Available: https://www.cms.gov/Medicare-Medicaid-Coordination/Fraud-Prevention/MedicaidIntegrity-Education/Downloads/fwa-factsheet.pdf

[5] D. Thornton, M. Brinkhuis, C. Amrit, and R. Aly, "Categorizing and Describing the Types of Fraud in Healthcare," *Procedia Comput. Sci.*, vol. 64, no. December, pp. 713–720, 2015, doi: 10.1016/j.procs.2015.08.594.

[6] M. C. Massi, F. Ieva, and E. Lettieri, "Data mining application to healthcare fraud detection: A two-step unsupervised clustering method for outlier detection with administrative databases," *BMC Med. Inform. Decis. Mak.*, vol. 20, no. 1, 2020, doi: 10.1186/s12911-020-01143-9.

[7] M. Button and C. Leys, "Healthcare fraud in the new NHS market – a threat to patient care," 2013, [Online]. Available: http://chpi.org.uk/wp-content/uploads/2012/06/CHPI-Healthcare-Fraud-a-threat-to-patient-care1.pdf

[8] "Health care fraud and abuse enforcement: Relationship scrutiny," p. 14, 2015, [Online]. Available: http://www2.deloitte.com/content/dam/Deloitte/us/Documents/life-sciences-health-care/us-lshc-health-carefraud-abuse.pdf

| e-ISSN: 2320-9801, p-ISSN: 2320-9798| www.ijircce.com | |Impact Factor: 8.379 |

|| Volume 11, Issue 7, July 2023 ||

| DOI: 10.15680/IJIRCCE.2023.1107002 |

[9] M. Sparrow, "Fraud Control in the Health Care Industry: Assessing the State of the Art," *Natl. Inst. Justice J.*, vol. JR000235, no. 1042, p. 11, 1998, [Online]. Available: https://www.ncjrs.gov/App/Publications/abstract.aspx?ID=172841

[10] "FBI Warns Health Care Professionals of Increased Potential for Fraudulent Sales of COVID-19-Related Medical Equipment — FBI."

[11] D. Pustika Sukma, A. Sulistiyono, and W. Tresno Novianto, "Fraud in Healthcare Service," *SHS Web Conf.*, vol. 54, p. 03015, 2018, doi: 10.1051/shsconf/20185403015.

[12] M. L. Offen, "Health care fraud," *Neurologic Clinics*, vol. 17, no. 2. pp. 321–333, 1999. doi: 10.1016/S07338619(05)70135-3.

[13] G. A. Ogunbanjo and K. D. van Bogaert, "Ethics in health care: Healthcare fraud," *South African Fam. Pract.*, vol. 56, no. 1, 2014.

[14] Centers for Medicare & Medicaid Services/Medicare Learning Network., "Medicare fraud & abuse: Prevent, detect, report. ICN MLN4649244.," no. January, pp. 1–21, 2021, [Online]. Available: https://www.cms.gov/Outreach-and-Education/Medicare-Learning-Network-MLNMLNProducts/Downloads/Fraud-Abuse-MLN4649244-Print-Friendly.pdf

[15] "Health Care Fraud — FBI."

BIOGRAPHY

Faith Nyakudya is a Masters degree student in the Software Engineering Department, School of Information Sciences, Harare Institute of Technology. She received a Btech degree in 2016 from HIT, Harare, Zimbabwe. Her research interests are Digital Fraud examinations, HCI, Digital marketing etc.

INNO SPACE
SJIF Scientific Journal Impact Factor

Impact Factor: 8.379

doi cross ref

ISSN INTERNATIONAL STANDARD SERIAL NUMBER INDIA

निस्केयर NISCAIR

INTERNATIONAL JOURNAL OF INNOVATIVE RESEARCH

IN COMPUTER & COMMUNICATION ENGINEERING

9940 572 462 6381 907 438 ijircce@gmail.com

www.ijircce.com

Scan to save the contact details