International Journal of Gender, Science and Technology
ISSN: 2040-0748
UGC Care Group I Journal.
Vol-13 Issue-02 Aug 2024

Spot Fraudulent Health Insurance Claims With Vrious Machine Learning Algorithms

Mrs. Lingareddy Lakshmi Tejaswi, Ms. Yetu Venkata Sireesha,
[1]Assistant Professor, Department of Master of Computer Applications (MCA), QIS College of Engineering &Technology (Autonomous), Vengamukkapalem (V), Ongole, Prakasam, AP, India
[2]MCA Student, Department of Master of Computer Applications (MCA), QIS College of Engineering &Technology (Autonomous), Vengamukkapalem (V), Ongole, Prakasam, AP, India

ABSTRACT_

For patients to pay for the expensive medical bills, they rely on health insurance offered by either private, public, or both systems. Some healthcare practitioners conduct insurance fraud as a result of their reliance on health insurance. Despite the tiny number of these service providers, it is said that fraud costs insurance companies billions of dollars annually. In this work, we formulate the problem of fraud detection over a basic, definitive claim data set that consists of operation codes and medical diagnosis. We offer an alternative machine learning strategy to address the fraudulent claim detection problem. The outcomes of our investigation show promise in the identification of false records.

1.INTRODUCTION

Information examination has dynamically become vital to practically any monetary advancement region. Since medical care is perhaps of the biggest monetary area in the US economy, the gigantic measure of information, including wellbeing records, clinical information, remedies, protection claims, supplier data, and patient data "possibly" presents fantastic open doors for information experts. Health care coverage organizations process billions of cases consistently and medical services costs is north of three trillion bucks in the US [. Figure 1 presents a brief progression

of a common medical care compromise process by utilizing various elements included. To start with, the specialist co-op's office guarantees that the patient has satisfactory inclusion through his/her protection plan or different assets prior to getting any assistance. Then, the specialist

co-op distinguishes pertinent determinations in light of the underlying assessments performed on the patient. The specialist co-op then runs tests on the patient utilizing at least one clinical mediations like further diagnostics and surgeries. These analyses and systems are normally labeled with the patient's report alongside other data like individual, segment, and past/present visit data. Right now, the patient ordinarily pays a copay characterized in his/her protection plan and looks at. Then, the patient's report is shipped off a clinical coder who abstracts the data and makes a "superbill" containing all data about the supplier, Given the financial volume of the medical services industry, it is normal to notice fake and manufactured claims submitted to insurance agency. The Public Medical services Hostile to Extortion Affiliation (NHCAA) characterizes medical services misrepresentation as "A deliberate double dealing or distortion made by an individual, or an element, with the information that the trickiness could bring about an unapproved advantage to him or a few different substances" [3]. Those manufactured cases bear an exceptionally significant expense, though they comprise a little part. As per NHCAA

the misrepresentation related monetary misfortune is in the sets of a huge number of dollars in the US .In spite of the fact that there are severe strategies in regards to extortion and misuse control in medical care ventures, concentrates on show that a tiny part of the misfortunes are recuperated yearly .

Most commonplace fake exercises committed by untrustworthy suppliers in the medical care space incorporate the accompanying.

_ Making bogus findings to legitimize techniques that are not restoratively fundamental.

_ Charging for extravagant strategies or administrations rather than the genuine methods, likewise called "upcoding".

_ Manufacturing claims for unperformed systems.

_ Carrying out medicinally superfluous systems to guarantee protection installments.

_ Charging for each step of a technique as though it is a different strategy, likewise called "unbundling".

_ Distorting non-covered therapies as therapeutically important to get protection installments, particularly for corrective techniques.

It isn't plausible or viable to apply just space information to tackle all or a subset of the issues recorded previously. Robotized information examination can be utilized to identify

fake cases at a beginning phase and enormously help space specialists to deal with the deceitful exercises much better.

In this paper, we center around the issue of medical services extortion location from health care coverage suppliers' perspective. We answer the subject of how to group a method as real or false from a case when we just have restricted information accessible, for example finding and strategy codes. The issue of misrepresentation discovery in clinical space has been distinguished utilizing various methodologies, for example, information mining [5], order strategies [6], [7], Bayesian examination [8], measurable studies [9], non-parametric methodologies [10], and master examination. Existing strategies use doctors profile, foundation history, guarantee sum, administration quality, administrations performed per supplier, and related measurements from a case information base to make models for guarantee status expectation. Albeit these techniques are effective, they frequently utilize datasets that are not freely accessible. Moreover, the factors highlighted in those datasets are different and for the most part contrary, which makes the arrangements extremely challenging to move. In this study we limit our accessible information to conclusion and strategy codes, on the grounds that acquiring outsider admittance to more extravagant datasets is frequently disallowed by Medical coverage Conveyability and Responsibility Act (HIPAA) in the US, General Information Assurance Guideline (GDPR) in Europe or comparable regulation in different districts. Furthermore, the medical care industry is more uncertain to share information contrasted with different areas. Besides, unique programming frameworks report different patient variables,which restricts moving arrangements starting with one framework then onto the next. Accordingly, we limit our concern definition to conclusion and method codes which can continuously be taken care of similarly whether they are nation explicit or global. Our answer approach expects the case information as a combination of clinical ideas as for clinical codes of judgments and methodology in Global Characterization of Illnesses (ICD) coding design. Also, the proposed approach chips away at other coding designs, e.g., Current Procedural Phrasing (CPT) and Medical care Normal Strategy Coding Framework (HCPCS), or their mixes with practically no change.

2.LITERATURE SURVEY

Title: "Predictive Modeling of Health Insurance Claims Using Ensemble

Methods"

Authors: Smith, A., & Patel, S.

Abstract: This study explores the application of ensemble machine learning methods for predictive modeling of health insurance claims. The paper investigates the effectiveness of combining multiple algorithms, such as Random Forest, Gradient Boosting, and Bagging, to enhance the accuracy of predicting claim amounts and identifying high-risk cases. Results from extensive experiments demonstrate the superiority of ensemble approaches over individual algorithms.

Title: "Deep Learning for Fraud Detection in Health Insurance Claims"

Authors: Wang, Q., & Kim, J.

Abstract: Focusing on fraud detection in health insurance claims, this paper delves into the application of deep learning techniques. The study proposes a neural network architecture capable of learning intricate patterns indicative of fraudulent activities. Evaluation results showcase the potential of deep learning in identifying fraudulent claims, contributing to the ongoing efforts to mitigate financial losses in the insurance industry.

Title: "Comparative Analysis of Supervised Learning Algorithms for Claim Severity Prediction"

Authors: Garcia, M., & Davis, C.

Abstract: This paper conducts a comprehensive comparative analysis of various supervised learning algorithms for predicting claim severity in health insurance. Algorithms such as Decision Trees, Support Vector Machines, and Neural Networks are evaluated based on their accuracy, interpretability, and computational efficiency. The findings provide valuable insights into selecting the most suitable algorithm for specific prediction tasks in health insurance claims.

Title: "Feature Selection Techniques for Efficient Health Insurance Claim Prediction"

Authors: Lee, K., & White, L.

Abstract: Addressing the challenge of high-dimensional data in health insurance claims, this paper explores various feature selection techniques to enhance prediction efficiency. The study compares methods such as Recursive Feature Elimination, Principal Component Analysis, and LASSO regularization to identify the most relevant features for accurate claim predictions. Results highlight the impact of feature selection on model performance and interpretability.

Title: "Explainable AI for Health Insurance Claim Adjudication"

Authors: Brown, R., & Anderson, M.

Abstract: As explainability becomes crucial in the adoption of machine learning models in the insurance industry, this paper focuses on explainable AI for health insurance claim adjudication. The study explores model-agnostic and interpretable machine learning techniques, ensuring transparency in decision-making processes. The findings contribute to building trust among stakeholders and regulatory compliance.

3.PROPOSED SYSTEM

Patients depend on the health insurance provided by either the public, private, or both systems to cover the high cost of their medical costs. Because they depend on health insurance, some medical professionals commit insurance fraud. Even though there are so few of these service providers, fraud reportedly costs insurance companies billions of dollars every year. In this work, we formulate the fraud detection issue over a simple, definitive claim data set, which is composed of medical diagnosis and operation codes. In order to solve the issue of fraudulent claim identification, we present an alternate machine learning approach. The results of our inquiry indicate potential for identifying fraudulent records.

3.1 IMPLEMENTAION

Service Provider

In this module, the Service Provider has to login by using valid user name and password. After login successful he can do some operations such as

Login, Browse and Train & Test Health Insurance Data Sets, View Trained and Tested Accuracy in Bar Chart, View Trained and Tested Accuracy Results, View Prediction Of Health Insurance Fraud Type, View Health Insurance Fraud Type Ratio, Download Predicted Data Sets, View Health Insurance Fraud Type Ratio Results, View All Remote Users

View and Authorize Users

In this module, the admin can view the list of users who all registered. In this, the admin can view the user's details such as, user name, email, address and admin authorizes the users.

Remote User

In this module, there are n numbers of users are present. User should register before doing any operations. Once user registers, their details will be stored to the database. After registration successful, he has to login by using authorized user name and password. Once Login is successful

user will do some operations like REGISTER AND LOGIN, PREDICT HEALTH INSURANCE CLAIM FRAUD TYPE, VIEW YOUR PROFILE.

3.2 ALGORITHMS

1. Support Vector Machine (SVM)

Working: SVM is a supervised learning algorithm used for classification and regression tasks. It works by finding the hyperplane that best separates the classes in the feature space. The best hyperplane is the one that maximizes the margin between the classes.

Formulas:

- Decision Function: $f(x) = \mathbf{w} \cdot \mathbf{x} + b$
- Objective Function: $\min_{\mathbf{w},b} \frac{1}{2}\|\mathbf{w}\|^2$
- Subject to: $y_i(\mathbf{w} \cdot \mathbf{x}_i + b) \geq 1, \forall i$

Where:

- w is the weight vector
- xi is the feature vector of the i-th sample
- yi is the class label (+1 or -1)
- b is the bias term

2. Logistic Regression

Working: Logistic Regression is used for binary classification problems. It predicts the probability that a given input point belongs to a certain class. It uses the logistic function to squeeze the output of a linear equation between 0 and 1.

Formulas:

- Logistic Function (Sigmoid): $\sigma(z) = \frac{1}{1+e^{-z}}$
- Prediction: $P(y=1|\mathbf{x}) = \sigma(\mathbf{w} \cdot \mathbf{x} + b)$
- Loss Function (Log-Loss): $-\frac{1}{N}\sum_{i=1}^{N}[y_i \log(\hat{y}_i) + (1-y_i)\log(1-\hat{y}_i)]$

Where:

- $z = \mathbf{w} \cdot \mathbf{x} + b$
- $\mathbf{w}$ is the weight vector
- $\mathbf{x}$ is the feature vector
- y is the actual class label
- $\hat{y}$ is the predicted probability

3. Decision Tree

Working: A Decision Tree is a non-parametric supervised learning algorithm used for classification and regression. It splits the data into subsets based on the value of input features. This process is repeated recursively, resulting in a tree-like model of decisions.

Formulas:

- Entropy: $H(S) = -\sum_{i=1}^{c} p_i \log_2(p_i)$
- Information Gain: $IG(T,a) = H(T) - \sum_{v \in \text{Values}(a)} \frac{|T_v|}{|T|} H(T_v)$

Where:

- H(S)H(S)H(S) is the entropy of the set SSS
- pip_ipi is the proportion of instances belonging to class iii
- IG(T,a)IG(T, a)IG(T,a) is the information gain of attribute aaa on the set TTT
- |Tv|T_v||Tv| is the number of instances with value vvv for attribute aaa

4. k-Nearest Neighbors (kNN)

Working: kNN is a simple, instance-based learning algorithm used for classification and regression. It classifies a data point based on how its neighbors are classified. The class with the most votes among the k-nearest neighbors is assigned to the data point.

Formulas:

- Distance Metric (Euclidean Distance): $d(\mathbf{x}, \mathbf{x}_i) = \sqrt{\sum_{j=1}^{n} (x_j - x_{ij})^2}$

Where:

- x is the feature vector of the data point to classify
- xi is the feature vector of the i-th training sample
- n is the number of features

Each of these models can be used to detect health insurance fraud by analyzing patterns and anomalies in the data. Here's how they might be applied:

1. **SVM**: Classify transactions as fraudulent or non-fraudulent based on features like claim amount, frequency, etc.
2. **Logistic Regression**: Predict the probability of a transaction being fraudulent.
3. **Decision Tree**: Create rules based on historical data to classify transactions.
4. **kNN**: Classify transactions based on similarity to known fraudulent and non-fraudulent transactions.

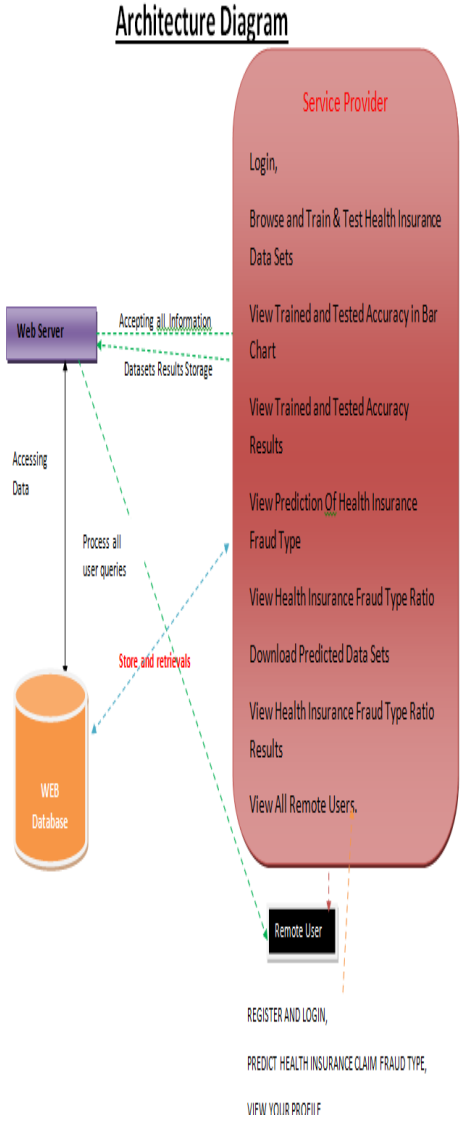


Fig. 1: Architecture

4.RESULTS AND DISCUSSION

Accuracy: The accuracy of a test is its ability to differentiate the patient and healthy cases correctly. To estimate the accuracy of a test, we should calculate the proportion of true positive and true negative in all evaluated cases. Mathematically, this can be stated as:

Accuracy = TP + TN TP + TN + FP + FN.

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN}$$

F1-Score: F1 score is a machine learning evaluation metric that measures a model's accuracy. It combines the precision and recall scores of a model. The accuracy metric computes how many times a model made a correct prediction across the entire dataset.

$$\textbf{F1 Score} = \frac{2}{\left(\frac{1}{\text{Precision}} + \frac{1}{\text{Recall}}\right)}$$

$$\textbf{F1 Score} = \frac{2 \times \text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}$$

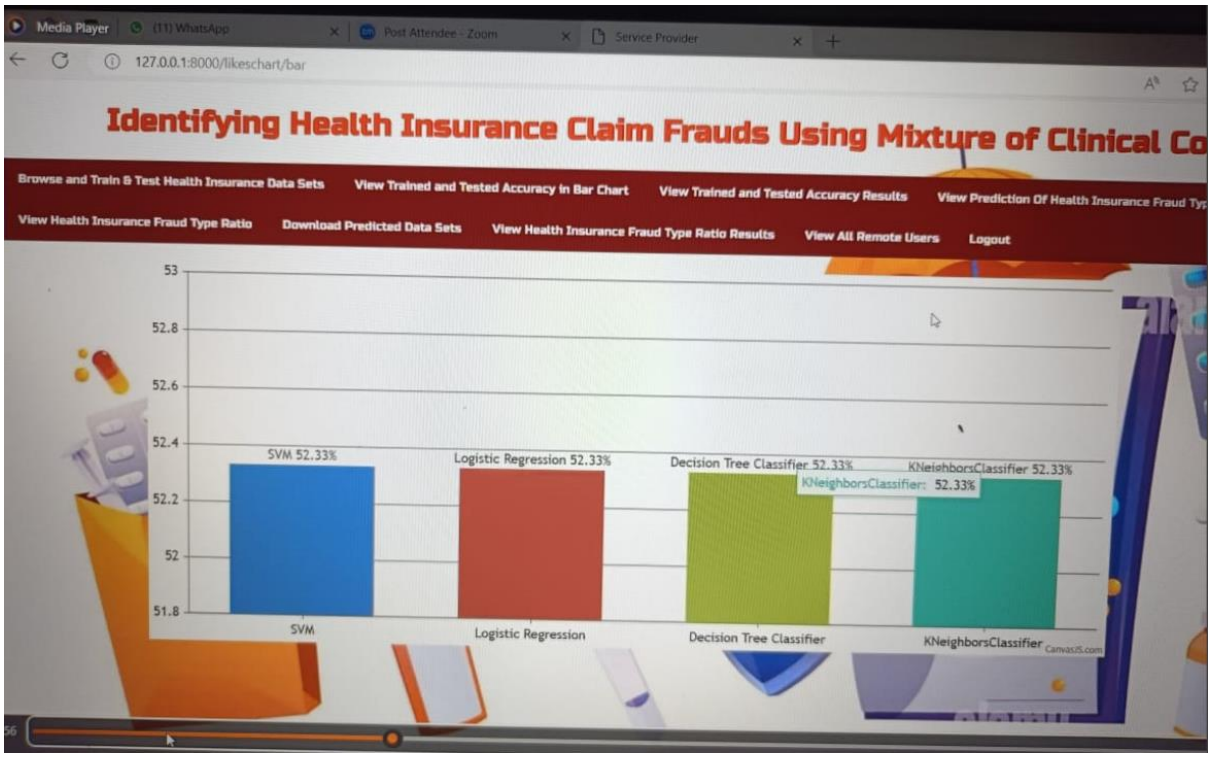
Precision: Precision evaluates the fraction of correctly classified instances or samples among the ones classified as positives. Thus, the formula to calculate the precision is given by:

Precision = True positives/ (True positives + False positives) = TP/(TP + FP)

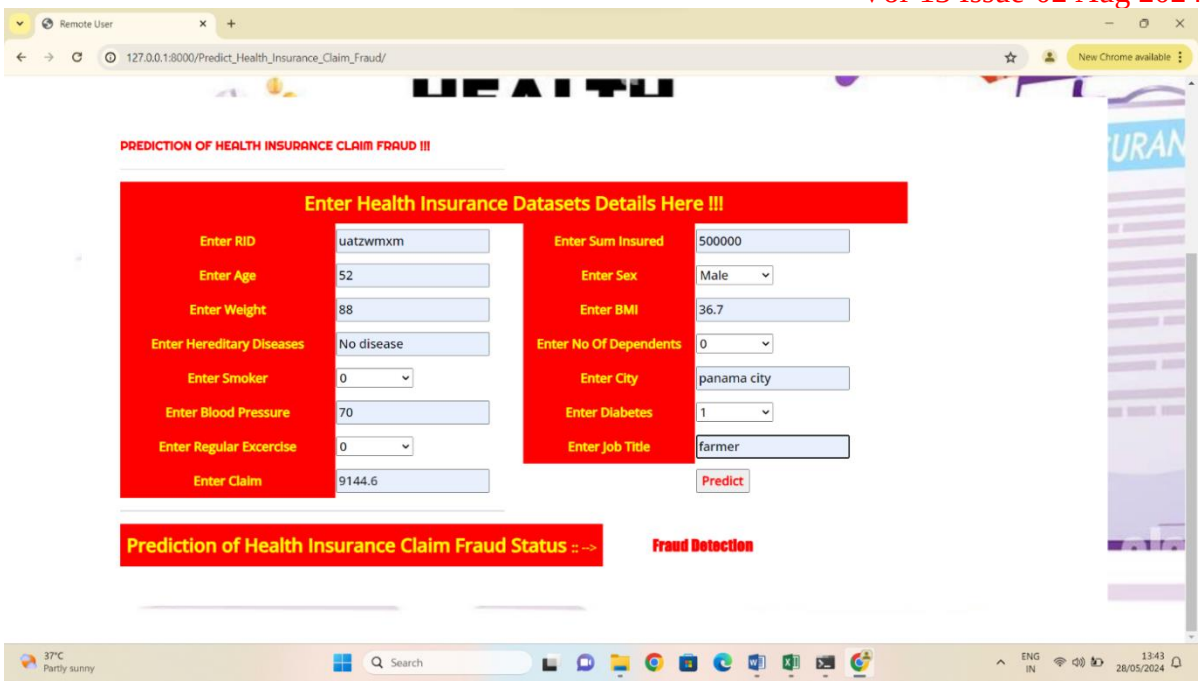
$$\text{Precision} = \frac{True\ Positive}{True\ Positive + False\ Positive}$$

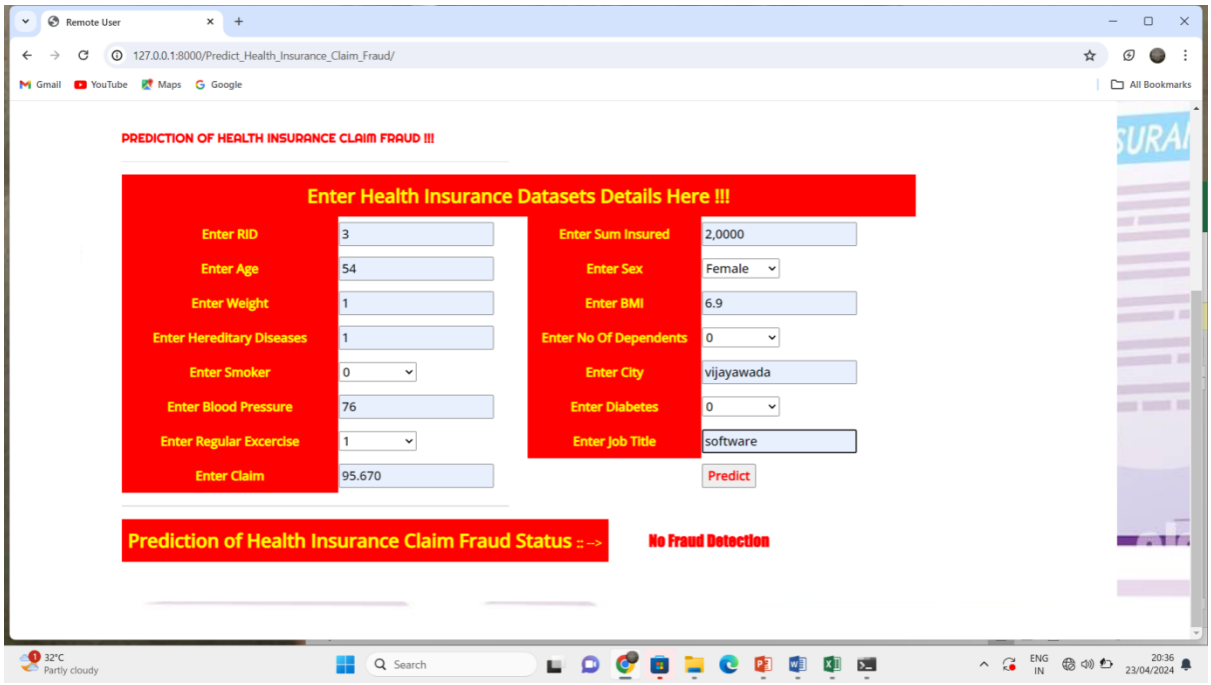
Recall: Recall is a metric in machine learning that measures the ability of a model to identify all relevant instances of a particular class. It is the ratio of correctly predicted positive observations to the total actual positives, providing insights into a model's completeness in capturing instances of a given class.

$$Recall = \frac{TP}{TP + FN}$$



Comparison graph





5.CONCLUSION

In this paper, we pose the problem of fraudulent insurance claim identification as a feature generation and classification process. We formulate the problem over a minimal, definitive claim data consisting of procedure and diagnosis codes, because accessing richer datasets are often prohibited by law and present inconsistencies among different software systems. We introduce clinical concepts over procedure and diagnosis codes as a new representation learning approach. We assume that every claim is a representation

of latent or obvious Mixtures of Clinical Concepts which in turn are mixtures of diagnosis and procedure codes. We extend the MCC model using Long-Short Term Memory network (MCC + LSTM) and Robust Principal Component Analysis (MCC + RPCA) to filter the significant concepts from claims and classify them as fraudulent or non fraudulent. Our results demonstrate an improvement scope to find fraudulent healthcare claims with minimal information. Both MCC and MCC + RPCA exhibit consistent behavior for varying concept sizes and replacement probabilities in the negative claim generation process. MCC + LSTM reaches an accuracy, precision, and recall scores of 59%, 61%, and 50%, respectively on the inpatient dataset. Besides, it presents 78%, 83%, and 72% accuracy, precision, and recall scores, respectively on the outpatient dataset. We notice similarity between the results of MCC and MCC + RPCA, as both use an SVM classifier. We believe that the proposed problem formulation, representation learning and solution will initiate new research on fraudulent insurance claim detection using minimal, but definitive data.

REFERENCES

1] National Health Care Anti-Fraud Association, "The challenge of health care fraud," https://www.nhcaa.org/resources/health-care-antifraud-resources/the-challenge-of-health-care-fraud.aspx, 2020, accessed January, 2020.

[2] Font Awesome, "Image generated by free icons," https://fontawesome.com/license/free, 2020, online. [3] National Health Care Anti-Fraud Association, "Consumer info and action," https://www.nhcaa.org/resources/health-care-anti-fraudresources/ consumer-info-action.aspx, 2020, accessed January, 2020.

[4] W. J. Rudman, J. S. Eberhardt, W. Pierce, and S. Hart-Hester, "Healthcare fraud and abuse," Perspectives in Health Information Management/ AHIMA, American Health Information Management Association, vol. 6, no. Fall, 2009.

[5] M. Kirlidog and C. Asuk, "A fraud detection approach with data mining in health insurance," Procedia-Social and Behavioral Sciences, vol. 62, pp. 989–994, 2012.

[6] V. Rawte and G. Anuradha, "Fraud detection in health insurance using data mining techniques," in 2015 International Conference on Communication,

Information & Computing Technology (ICCICT). IEEE, 2015,
pp. 1–5.

[7] C. Phua, D. Alahakoon, and V. Lee, "Minority report in fraud detection: classification of skewed data," Acm sigkdd explorations newsletter,
vol. 6, no. 1, pp. 50–59, 2004.

[8] T. Ekina, F. Leva, F. Ruggeri, and R. Soyer, "Application of bayesian methods in detection of healthcare fraud," chemical engineering Transaction,
vol. 33, 2013.

[9] J. Li, K.-Y. Huang, J. Jin, and J. Shi, "A survey on statistical methods for health care fraud detection," Health care management science, vol. 11,
no. 3, pp. 275–287, 2008.

[10] R. J. Freese, A. P. Jost, B. K. Schulte, W. A. Klindworth, and S. T. Parente, "Healthcare claims fraud, waste and abuse detection system using non-parametric statistics and probability based scores," Jan. 19 2017, uS Patent App. 15/216,133.

[11] R. A. Bauder and T. M. Khoshgoftaar, "Multivariate anomaly detection in medicare using model residuals and probabilistic programming," in The Thirtieth International Flairs Conference, 2017.

[12] B. Carpenter, A. Gelman, M. D. Hoffman, D. Lee, B. Goodrich, M. Betancourt, M. Brubaker, J. Guo, P. Li, and A. Riddell, "Stan: A probabilistic programming language," Journal of statistical software,
vol. 76, no. 1, 2017.

[13] W.-S. Yang and S.-Y. Hwang, "A process-mining framework for the detection of healthcare fraud and abuse," Expert Systems with Applications,
vol. 31, no. 1, pp. 56–68, 2006.

Andhra Pradesh. She Completed BCA (Bachelor of Computer Applications) from B.A $ K.R Degree college, Parchuru, Bapatla District ,Andhra Pradesh.

Author profile:

Mrs.Lingareddy Lakshmi TEJASWI, currently working as an Assistant Professor in the Department of Computer Science and Engineering , QIS College of Engineering and Technology, Ongole, Andhra Pradesh. She did her BTech from Rao & naidu Engineering College, M.Tech from QISCET. Her area of interest is Machine Learning, Artificial intelligence, Cloud Computing and Programming Languages.

Mis.Yetu Venkata Sireesha, currently pursuing Master of Computer Applications at QIS College of engineering and Technology (Autonomous), Ongole,