INSURANCE FRAUD DETECTION USING SUPERVISED LEARNING

SIT717 – Assignment 1

Abstract

Survey report on usage of supervised learning in insurance industry

Rajeshkumar Mourya
rmourya@deakin.edu.au
218615876

Title	Insurance fraud detection using supervised learning
Student Name	Rajeshkumar R Mourya
Student ID	218615876

Contents

Abstract

The survey is intended for providing information on big data in the insurance industry and supervised learning for classification of insurance claims. The insurance industry has been beneficial for various businesses and individuals as a safeguard against damages to their assets and health. However, there are several fraudulent behaviours observed when it comes to claims. Organization and individuals have claimed money by misrepresenting information to gain an unauthorized benefit from the insurers. The number of cases of such frauds is increasing every year and cases are widespread. The insurance industry has observed frauds in every sector including healthcare, auto, finances etc. With the advancement in data mining technique and an increase in its usage in different sectors, the insurance industry is also using these techniques. Supervised and unsupervised and now deep learning are being implied in the insurance industry to detect and avoid frauds in insurance claims.

Introduction

Misleading an insurance company intentionally that results in an individual or a group receiving insurance benefits illegitimately is known as insurance fraud. Financial benefits are the main reason for insurance frauds. It is estimated that the number of fraudulent claims amounts to 15% of total claims according to a recent survey. Insurance companies in the USA face losses of over 30 billion USD annually due to fraudulent healthcare insurance claims (Rawte & Anuradha 2015). The insurance claim may result in thousands to millions of dollars in claims, considering this, one undetected fraud might cause huge loss to insurers. This may disrupt the cost management and process efficiency for insurance companies. The traditional fraud detection systems instilled by these companies are proven inefficient as the number of fraudulent claims is increasing. Traditional approaches involve human intervention and might take a lot of time and money to detect fraud, impacting the profitability of the firms (Harjai, Khatri & Singh 2019).

Rawte and Anuradha (2015) discussed a broad classification of fraudulent claims. These are categorized as follows: Billing for damages or costs that never incurred by forging a signature or colluding with service providers; Providing higher cost than the actual cost of service; Quoting higher prices for assets; Duplicate claims by changing minor details in the bills; Claiming unnecessary service cost, ex. hospitals/clinics charging for services which did not apply to the patient.

Digitization has enabled these firms to switch from paper-based to digital systems. These firms are capturing customer data using the latest software and hardware capabilities resulting in the generation and processing of vast amount of data. The evolution of big data and the growth of unstructured data has enabled fraudsters to exploit the loopholes of the system. If the data is not being analyzed properly, it might result in fraud (Harjai, Khatri & Singh 2019). Data mining techniques are being used for analyzing data and detection of frauds. Data mining is associated with (a) supervised learning based on training data which has legal as well as illegal (fraud) cases and (b) unsupervised learning with data that does not have any differentiation (Bhowmik 2011). Supervised learning is the most used technique among the two. Supervised data mining methods are further discussed in the next sections.

Fraud detection

Researchers have been working on different machine learning techniques to tackle the issue of fraud detection. Semi-supervised, unsupervised, supervised and hybrid models are implemented to detect illegitimate claims by researchers to create a fraud detection system that can help insurers.

Rawte and Anuradha (2015) proposed a hybrid model using Evolving clustering method (ECM) and Support vector machine (SVM) due detect healthcare insurance frauds. Due to the dynamic nature of data acquired by insurers and new data is generated continuously, ECM was selected for dynamic data clustering while SVM is used for classification of claims. For each new data point added, ECM clusters them by changing the size and position of the cluster. Using radius and distance threshold parameters, ECM adjusts the size and radius of the clusters to accommodate new data points. SVM is a supervised learning method which uses classified data for training. This data already has legal and fraudulent claims classified for the algorithm. After the training phase, SVM can classify new claims into legitimate or fraudulent claims. Authors explained the working through a block diagram given in fig.1. After the patient has made a claim, ECM clusters the claim according to the type of disease and then SVM is applied to each of the clusters to classify those claims. These steps are repeated for every claim and depending on classification, an expert can further review the fraudulent claim and provide results to the insurer.

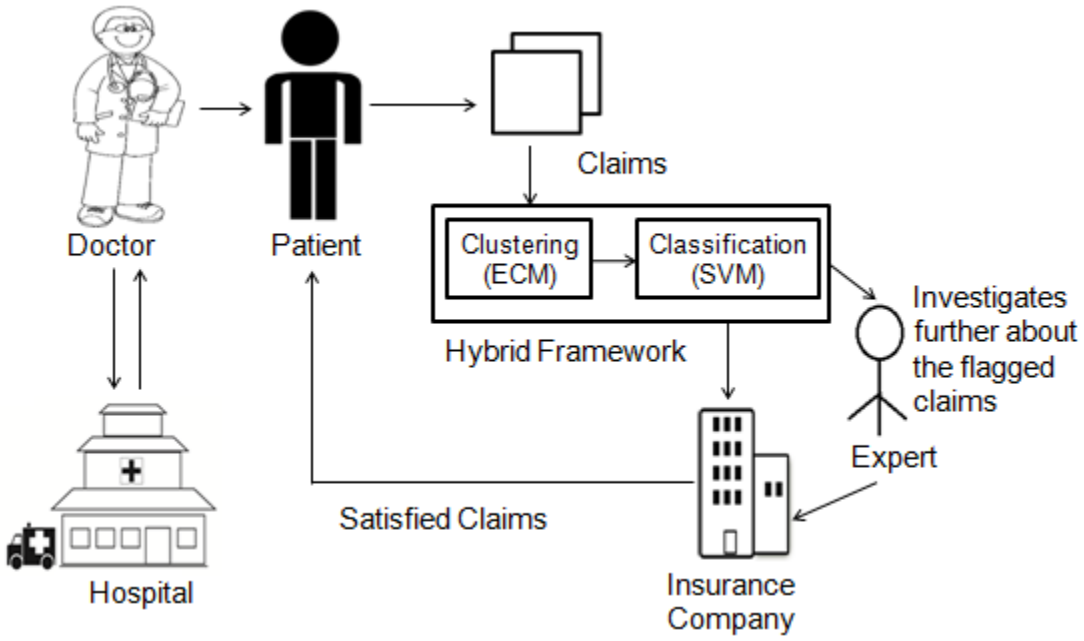

Fig 1: Hybrid model (Rawte & Anuradha 2015, fig 8, p. 4)

Bhowmik (2011) discussed auto insurance fraud detection. The author used the Naïve Bayesian classification and decision tree algorithm for predicting and analyzing fraud patterns from data. Bayesian network is used for classification of data while decision trees are used for constructing characteristics of fault. The steps utilized for crime detection are: a) classification using classifiers b) integrate classifiers c) Clustering of data using Artificial neural networks (ANN) d) Visualization of patterns. To describe the

behaviour of auto insurance, two Bayesian networks are built. One Bayesian network is constructed assuming the claim is fraudulent while another network assumes claim is legal. Fraud behaviour net is constructed with the help of experts while legal claim net is constructed using existing data of legitimate claims. By entering the data in these two networks, the probability of claim to fraudulent or legal can be computed. The author also used a decision tree for classification of the data. The basic algorithm for decision tree is starting a single N node tree representing the dataset.

- Considering outputs are legal and fraud for classification. If instances of type fraud, then the node becomes leaf and is classified as fraud.
- Otherwise, the algorithm uses entropy, classification error and Gini index to calculate the degree of impurity for selection of attributes. The selected attributes will then be used to separate the data into the above-mentioned classes.
- Entropy is the sum of the conditional probabilities of an event (p_i) times its information required for the event in subsets (b_i).

Confusion matrix and ROC graph were used as the measure of performance for these algorithms. The confusion matrix is the comparison of the predicted class of the data against the actual class of data.

Predicted/actual	Legal	Fraud
Legal	True positive (TP)	False-positive (FP)
Fraud	False-negative (FN)	True negative (TN)

Accuracy of predictions = (TP + TN) / (TP + FP + FN + TN), as well as precision and recall are also used as measure of performance. The author provided that these methods can be used effectively for the detection of frauds with an observed accuracy of over 78%.

Harjai, Khatri and Singh (2019) discussed the usage of Random forest and Synthetic minority oversampling technique (SMOTE) to detect fraudulent claims. The model was proposed by the authors to help insurers to take better claim-related decisions. In this method, SMOTE was utilized to rebalance the data followed by the classification of data using random forest classifier. Steps involved in the method were:

Step 1. Preprocessing the data
   a. Data cleansing
   b. Transformation of data
   c. Visualization of data
   d. Resampling data using SMOTE filter in WEKA

Step 2. Classification of data using Random forest classifier
   e. A classification applied on data with batch size 100 and seed value 1
   f. 10 folds cross-validation applied for implementation

Step 3. Training and testing the data using 80-20 split i.e. 80% data is used for training the model and 20% for testing the predictions.

Step 4. Performance/Validation of the model

This method provided accuracy and recall of 99.9% using confusion matrix as a measure. The analysis of the data used for this research provided the following insights:

- 82% of fraud cases had old vehicles involved (6-8 years old)

- 99.6 illegitimate cases had no witnesses while 83% of legitimate cases had witnesses/

These insights are useful for insurers while making any decision against fraud cases and can be used as a pro-active measure against such cases.

Bauder and Khoshgoftaar (2017) researched Medicare fraud using supervised, unsupervised and hybrid algorithms and compared them against the performance of each model. They used Gradient boosted machine (GBM), Random forest (RF), Deep neural network (DNN) and Naïve Bayes (NB) supervised learning methods for the analysis. 2015 PUF Medicare data was used for this research and the focus was given to non-prescription data. The non-prescription data are the codes of services which were not for specific services listed on Medicare policy. 80-20 split was used for training and testing the models. The healthcare frauds impact the ability of programs such as Medicare, Old age care policy etc. to provide effective and affordable care services to individuals. Validation the methods proposed by the author in this paper proved that the machine learning technique specially supervised learning methods are highly effective in detecting frauds. This enables removal of fraudsters/ perpetrators from the system and helps insurers in reducing the cost of operations.

Rayan (2019) proposed a fraud detection engine for Medi Assist group. The engine is utilized for flagging outstanding claims and created a subset of these claims every day for the investigation team. The proposed model produces 12 binary flags against outstanding claims. The statistical rule engine is used for first 10 flags while the remaining 2 utilize decision tree for classification of outstanding claims. The proposed framework is as given below in fig.2. The first 10 flags have created the engine against every claim and the claim is evaluated against the beneficiary and claim history, hospitals and experience, disease groups, risk of policy and demographics. Decision trees were used for the implementation of the classifier. Flag 11 decision tree is constructed using fully investigated data while the flag 12 decision tree was constructed using under sampled investigation data. Each claim goes through the fraud detection engine for flagging. The claim is sent to the investigation team even if one of the 12 flags is set. The team's dashboard is updated with relevant remarks for further investigation of the claim.

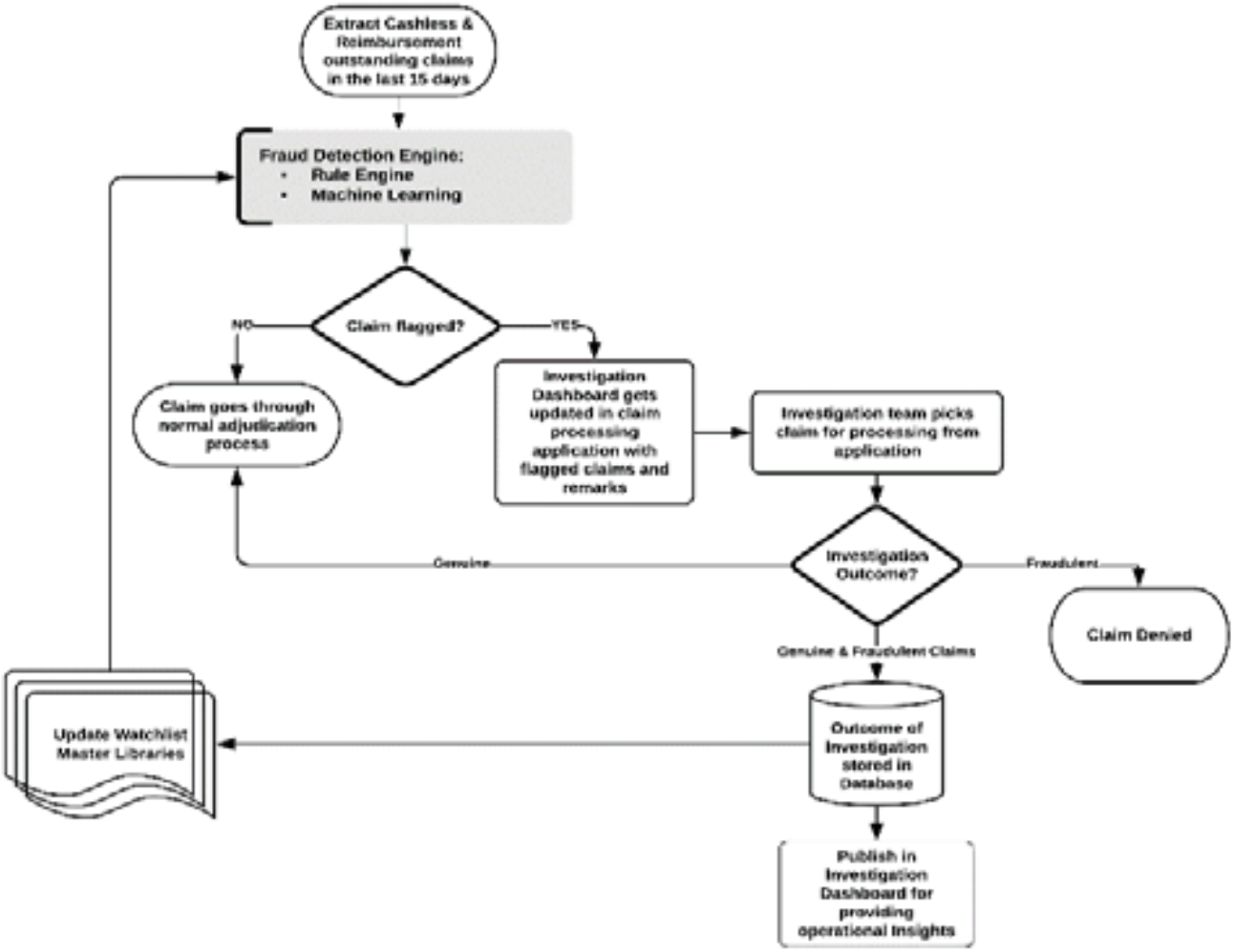

Fig 2: Fraud Analytics workflow (Rayan 2019, fig 2, p.5)

The engine was deployed in August 2018 for one of the insurers. The results of the deployed provided better hit-rate percentage to detect illegitimate claims. The results between 1 April 2017 to 25th April 2018 were as follows:

- Duration: 1st April, 2017 to 25th April, 2018
- Claims Investigated: 2,190
- Genuine: 1,999
- Fraudulent: 191
- Hit-Rate: 8.72%

Statistics post pilot:

- Duration: 25th April, 2018 to 25th April, 2019
- Triggered & Investigation: 2,827
- Investigation & Fraudulent: 763
- Hit-Rate: 26.98%

(Rayan 2019, p.7)

The results provided more than 200% increase in the hit-ratio. The model was eventually deployed to all public and private insurers for which Medi Assist was Third-Party claims administrator. As of June 2019, the model has helped save over 65 million INR for the insurers.

Analysis of the models

Rawte and Anuradha (2015) discussed a hybrid model using ECM and SVM algorithms. The classification of duplicate claim proved the efficiency of the model to be highly beneficial for the detection of fraud. The figures below show how the model can cluster the claims based on disease type and then classify the duplicate claim to be fraudulent.

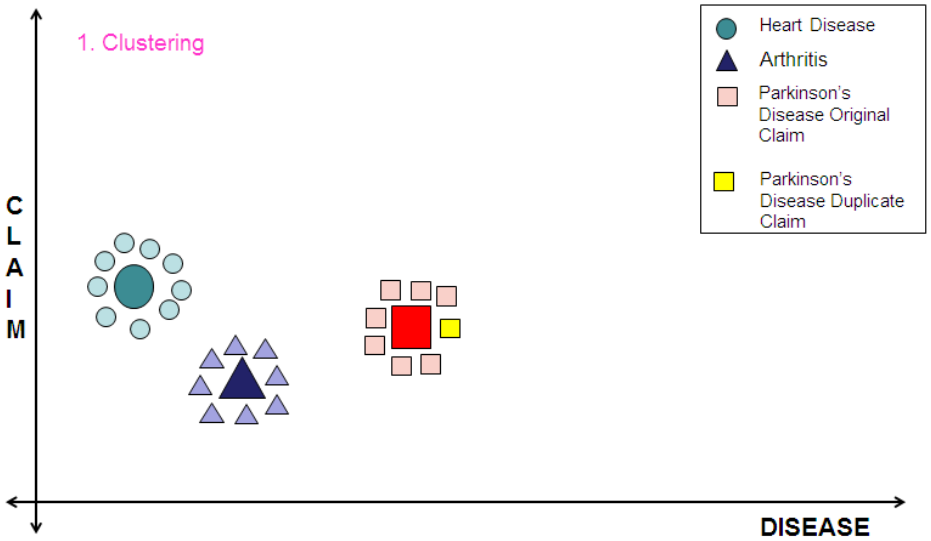

Fig. 3: Clusters created according to disease type (Rawte & Anuradha 2015, fig 9, p.5)

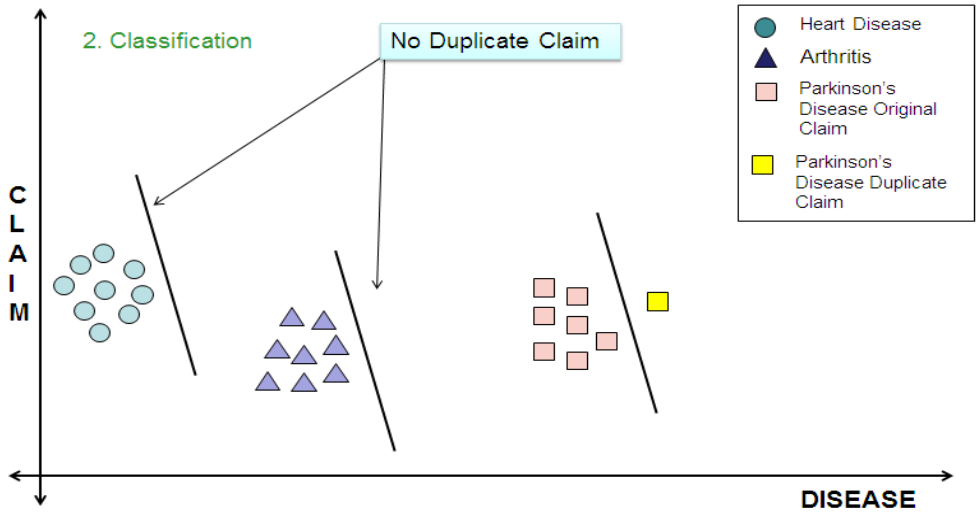

Fig. 4: Detection of a duplicate claim (Rawte & Anuradha 2015, fig 10, p.5)

This model overcomes the disadvantage associated with unsupervised learning to detect duplicate claims. ECM provides clustering for incoming data as there is a need to cluster the dynamic data. SVM is

generalized and much easier to train than the neural networks and it provides scalability and usability for good data mining.

The performance Naïve Bayesian and Decision tree algorithm (Bhowmik 2011) of measured using confusion matrix and different performance metrics were established. The model provided an accuracy of 78% for test data.

Table 2b. Confusion matrix of a model applied to test dataset

		Observed		
		legal	fraud	accuracy: 0.78
predicted	legal	3100	1125	recall: 0.86
	fraud	395	2380	precision: 0.70

(Bhowmik 2011, p. 5)

The accuracy of the result is promising but not accurate enough as the cost involved with the investigation cost may increase significantly if the number of false alarms is higher.

fraud	legal
True Positive(Hit) cost = number of hits * average cost per investigation	False Positive(False alarm) cost =number of false alarms * (Average cost per investigation + average cost per claim)
False Negative(miss) cost = number of misses * average cost per claim	True Negative(correct rejection) cost = number of correct rejection claims * average cost per claim

(Bhowmik 2011, p. 5)

The model based on Random forest classifier and SMOTE is highly accurate (Harjai, Khatri & Singh 2019). The performance of the model was compared against SMOTE-SVM, Decision tree and Multilayer perception (MLP) and the parity was astounding. The model provided an accuracy of 94% and recall of 99.9% which outperformed other models. The time taken to build the model by the system was 1.43 seconds. The model works on any real-time data as the data is transformed before it is fed to the classifier.

TABLE V: COMPARISON OF VARIOUS MODELS WITH THE PROPOSED APPROACH [5]

Performance Metrics (in %)	Support Vector Machine (SVM)	Decision Tree	Multi-layer Perceptron (MLP)	Proposed Mo
Accuracy	58.41	57.39	74.98	94.33
Sensitivity (or Recall value)	90.53	86.94	47.83	**99.9**
Specificity	36.86	38.14	18.75	45.1

(Harjai, Khatri & Singh 2019, p. 6)

Bauder and Khoshgoftaar (2017) compared performance based on data sampling and split of data for unsupervised, supervised and hybrid algorithms. With oversampling, these algorithms performed poorly in terms of MCC, F-measure and G-measure scores. However, with 80-20 split, supervised learning methods performed better than the other two and there was a distinct difference in the performance. Authors also concluded than the general providers of Medicare find it difficult to detect frauds than specialized providers.

The fraud detection engine provides a calibration than ensures the rate of flagged claims does not exceed the rate of investigation (Rayan 2019). This ensures the workload of investigation officers is considered to handle additional claims triggered by the engine. The engine proved effective during the case study, increasing the hit rate of fraud claims by over 200% and later deployed to all insurers under Medi Assist. The study also compared different two-class classifier models and identified that the average perceptron outperformed all other models and took a fraction of time compared to the neural network. This added scope for changing the externalized classifier used by the authors to be changed to average perception. The results of the comparison are shown below

Table I Summary Statistics for Neural Network

True Positive	False Negative	Accuracy	Precision	Threshold
97,207	771	0.973	0.980	0.5
False Positive	**True Negative**	**Recall**	**F1 Score**	**AUC**
1,943	796	0.992	0.986	**0.48**

Table II Summary Statistics for Averaged Perceptron

True Positive	False Negative	Accuracy	Precision	Threshold
1,07,653	3,589	0.946	0.963	0.5
False Positive	**True Negative**	**Recall**	**F1 Score**	**AUC**
4,098	27,929	0.968	0.966	**0.982**

(Rayan 2019, p. 7)

Conclusion

In this literature review, supervised learning methods and use of these techniques for data mining in the insurance industry to identify fraud claims were reviewed. We studied existing fraud detection systems and their application in the insurance industry. Fraud is harder to recognize with the ever-growing size of data and changes to the dynamics of the industry. We may not be able to eliminate the fraud although, with the use of data mining, we can identify and reduce the number of fraudulent cases. Data mining helps to identify the hidden patterns in the data using supervised learning and hybrid models which were proven effective. Data analytics are required to be performed in the current era else there are high chances of fraudsters using the system to their benefits. Any data mining model to detect such frauds is required to be validated to identify its usage in any system. A model can be evaluated based on performance metrics, confusion matrix etc. and the quality of data plays a crucial role in the performance of models. Compared to unsupervised data mining, supervised methods are highly effective when it comes to the classification of the data and to create a system to provide real-time analytics for investigation of fraud. The selection of algorithm and performance metrics are important for model evaluation. It can affect the result of the proposed methods. Current results are promising and with the advancement of machine learning and artificial intelligence as well as deep learning methodologies, there is scope to improve existing systems and create new models to provide highly accurate and faster systems to identify fraud. Insurers, government authorities can use the data generated through these systems to make better decisions and to blacklist the beneficiaries, hospitals, clinics and pharmaceuticals, preventing them from taking advantage of the system.

References

Bauder, RA & Khoshgoftaar, TM 2017, 'Medicare Fraud Detection Using Machine Learning Methods', in *2017 16th IEEE International Conference on Machine Learning and Applications (ICMLA)*, pp. 858-65.

Bhowmik, R 2011, 'Detecting auto insurance fraud by data mining techniques', *Journal of Emerging Trends in Computing and Information Sciences*, vol. 2, no. 4, pp. 156-62.

Harjai, S, Khatri, SK & Singh, G 2019, 'Detecting Fraudulent Insurance Claims Using Random Forests and Synthetic Minority Oversampling Technique', in *2019 4th International Conference on Information Systems and Computer Networks (ISCON)*, pp. 123-8.

Rawte, V & Anuradha, G 2015, 'Fraud detection in health insurance using data mining techniques', in *2015 International Conference on Communication, Information & Computing Technology (ICCICT)*, pp. 1-5.

Rayan, N 2019, 'Framework for Analysis and Detection of Fraud in Health Insurance', in *2019 IEEE 6th International Conference on Cloud Computing and Intelligence Systems (CCIS)*, pp. 47-56.