Intelligent Decision Technologies 15 (2021) 127–139
DOI 10.3233/IDT-200052
IOS Press

An intelligent unsupervised technique for fraud detection in health care systems

Kanksha, Aman Bhaskar, Sagar Pande*, Rahul Malik and Aditya Khamparia
Department of Computer Science and Engineering, Lovely Professional University, Mumbai, India

Abstract. Healthcare is an essential part of people's lives, particularly for the elderly population, and also should be economical. Medicare is one particular healthcare plan. Claims fraud is a significant contributor to increased healthcare expenses, though the effect of it could be lessened by fraud detection. In this paper, an analysis of various machine learning techniques was done to identify Medicare fraud. The isolated forest an unsupervised machine learning algorithm which improves overall performance while detecting fraud based upon outliers. The goal of this specific paper is generally to show probable dishonest providers on the ground of their allegations. Obtained results were found more promising compared to existing techniques. Around 98.76% accuracy is obtained using an isolated forest algorithm.

Keywords: Isolated forest, fraud detection, machine learning, unsupervised learning

1. Introduction

Government departments, companies, several businesses, or present day have embraced electric transactions to boost their convenience or profitability for the selling of services or goods; of fields including charge card, banking, health insurance, automobile insurance, online auctions, etc. [1]. Table 1 list the quantity of IC3 reports received from 2011 to 2014 the ensuring dollar losses. It can be observed that the volume of harm is continuously increasing, even though the number of issues is actually declining; this is because fraud is actually leading to far more destroys than in the past [2].

Table 1
Internet crime IC3 report

Year	Complaints	Dollars used (million)
2011	3,14,246	$ 485,253,971
2012	289,874	$ 581,441,110
2013	262,813	$ 781,841,611
2012	269,422	$ 800,492,073

The fraud detection is one of Medicare's most serious problems. The government claims that overall Medicare costs have significantly inflated related to Medicare abuse allegations. Health fraud is a structured offense including peers of providers, doctors and beneficiaries who work collectively in order to produce fraudulent claims. Careful review of data from Medicaid has proven fraud for many physicians. They use methods where the most expensive treatments and medicines are used through an undefined diagnosis code. The most prominent organizations impacted by such unethical practices are insurance companies [24–26]. As a result, medical providers continue raising their policy rates and health coverage is getting costly day-by-day. The aim of this paper is to show likely dishonest providers on the grounds of their allegations. In addition, substantial parameters helpful in identifying the activities of alleged providers of fraud are also discussed. In this paper, we propose an unsupervised machine learning approach for Medicare fraud detection utilizing publicly accessible claims labels and data for known fraudulent health-related providers across almost all medical providers or specialties types (e.g., cardiology or dermatology) [27–31]. The fraud and abuse of healthcare take several ways. Many of the most common forms of provider's fraud are:

- They produce fake bills for services which are not even provided to the customer.

*Corresponding author: Sagar Pande, Department of Computer Science and Engineering Lovely Professional University, Mumbai, India. E-mail: sagarpande30@gmail.com.

- They submit duplicate claims for the same individual service.
- Wrongly presenting the services provided to the patients.
- To cast more money for facilities which are more challenging or expensive than currently offered.
- Claiming for those covered services where in practicality provider has not covered.

The main contributions of our study are as follows:

- Outlier analysis was performed for detecting the fraud occurred in health care systems.
- Isolation Forest, K Means and Local Outlier Factor algorithms were used for fraud detection on the dataset.
- For Comparative analysis five performance metrics were used.

The remaining paper is ordered as follows as Section 2 provides an explanation literature survey. Section 3 focuses on dataset description and the workflow of the implementation technique. Section 4 depicts the results obtained from the proposed model. Section 5 delivers the conclusion and future scope.

2. Literature review

With the small quantity of readily available, documented Medicare fraud cases and also the reasonably recent availability of information, a great deal of the current Medicare fraud detection research using machine learning approaches has been discussed below.

Johnson and Nagarur [3] says that a brand new multi-stage method for insurance companies to identify fraud perpetrated by clients and suppliers. Step four then includes the info received in the preceding 3 steps to the complete risk assessment. Subsequently, the Stage five choice tree based strategy measures the danger threshold values. They applied the method to 4 different specialties, as well as discovering that the methodology worked pretty well for those of them with an entire accuracy rate of 86%. Semi-supervised as well as non supervised neural community techniques.

Nsiah-Boateng et al. [4] focused on the write examined insurance compensation details from the NHIS to assess the valuation of the advantages bundle to the sensitivity and furthermore, the client of the device to the monetary needs of health service providers. They show the item of a total of 644,663 health-related statements costing Ghana cedi (GHS) 11,8 million (US$3,1 million) have been registered with the study period. The cost ratio of promises to donations received rose from 4.3 to 7.2 with the period 2011 2013 to 5.0 in 2014. The proportion of instances resolved following 90 days has improved. The study reveals a much better than expected allocation ratio of claimants, indicating the much better really worth of the NHIS advantage plan to members.

Bauder et al. [5] focused on specific place is actually the exploitation of healthcare insurance plans, including Medicare. With this particular post, we are creating a machine learning layout to identify when physicians certainly show an anomaly in the healthcare insurance statements of theirs. This particular type of results suggests it is feasible to successfully make use of machine learning in a novel way to identify doctors in their respective fields exclusively by using the remedies they bill for. This specific attempt offers a device that could classify doctors who have possibly misused Insurance schemes for even more reviews.

Bauder et al. [6] discussed a comparative analysis of controlled, unsupervised, and hybrid printer mastering methods using four performance metrics in addition to category disparity reduction utilizing oversampling and an eighty twenty under sample procedure. The results show this helpful identification of dishonest vendors is really achievable, with the eighty twenty sampling method demonizing the really best overall performance concerning learners. The launch of new Medicare samples, excluding certain LEIE Sion codes, as well as the use of various sampling methods for class imbalances will be suggested for later studies.

Abdallah et al. [7] a survey on the aim as well as fraud detection of this specific survey pa- per is really offering a systematic and detailed analysis of the issues in addition to issues that impede the excellent results of FDSs. This specific analysis article looked at state-of-the-art fraud detection mechanisms in five fields of fraud. Additionally, the tactics, and also techniques for identifying fraud, have been classified also as examined.

Fursov et al. [8] focused on that Proposed text embedding architectures through truly serious learning that assist to improve the identification of fraudulent claims compared to other machine learning strategies. They typically use a data set out of a huge international health insurance company to exhibit the methods of ours. Empirical results demonstrate that our choice outperforms many other state-of-the-art approaches and might help to create the claims management procedure a great deal much more efficiently.

Farbmacher et al. [9] concentrated on Fraud Detection in Claims Management. The very best objective is, in fact, a statistical model for the detection of fraudulent claims as well as the automatic transaction of non fraudulent statements. Health care claims, nevertheless, have an unusual info system, that's bureaucratic and complex in range. Utilizing a sample of two million statements made by a private health insurer for Germany, they prove that the suggested models of ours outperform designs based on bag-of-words, hand-designed characteristics, and designs based on co-evolutionary neural networks. Claims administration is really a preferred way for these solutions outside of natural language processing or perhaps picture analysis which has obvious as well as quantifiable economic worth.

Kruthika and Manjunatha [10] focused that These frauds pose a threat to human dignity, resulting in financial losses. This work focuses on the link-based detection of fraud cases. Initially, the credit card fraud dataset is downloaded from Kaggle and the sorting strategies are used to pick the relevant variables from the data set. The efficiency of this classifier shall be determined using statistical parameters and cross-validation techniques. Results revealed that the SMOTE Deep algorithm obtained better fraud detection efficiency at an accuracy of 96.4 per cent. Based on these observations, the analysis of linkages has been shown to be significant in exploring the dynamics of the fraud network.

Resa et al. [11] focused on research use an electronic survey-based quantitative tool. The sample of this analysis was Indonesian civil servants. 197 respondents were chosen by random sampling and evaluated by inferential statistics. Preliminary findings – The results of this study showed that the majority of respondents (81.76 per cent) believed that fraud was the main problem in the Indonesian department and that fraud perception was strong (more than 65 per cent). This re- search leads to a more detailed understanding of the importance of legal sensitivities, internal audit and program analysis principles.

David et al. [12] described Automated rules management system for fraud detection. An integrated rules management system that evaluates the input of individual rules and optimizes the collection of active rules by means of a heuristic search and user-defined loss feature. Results show that only a minority of the original rules (50 per cent in one instance, and 20 per cent in the other) are capable of maintaining the efficiency of the original frameworks (e.g. recall or false-positive rate). In the future, they are also preparing to add a framework to match the principles and threshold for machine learning models at the same time.

Song et al. [13] proposed a novel data-driven approach to the formulation of predictive types for the detection of bulk cargo theft in ports. More exactly, many user class approaches as well as classification algorithms are actually used to choose the ideal characteristic set of relevant threat components. Subsequently, the implied Bayesian networks obtain the functions of graphically exhibiting the partnership with the risk components by fraud. Experimental results indicate which predictive airers are actually helpful, with accuracy as well as recall values greater than 0.8. Such predictive models aren't merely helpful for understanding the dependency of relevant risk factors, but additionally for advertising policy SEO in risk control.

Lucas et al. [14] discussed Machine learning and data mining methods have been used widely to diagnose credit card fraud. In this study, they suggest an HMM based function development strategy that helps us to integrate sequential information into transactions in the form of HMM-based apps. These HMM-based features allow the non-sequential classification classifier (Random Forest) to use sequential classification details. They model the real and dishonest conduct of merchants and cardholders according to two characteristics: timing and quantity of transactions.

Janbandhu et al. [15] states that this paper contains some hybrid analysis for data level and algorithm level treatment of class imbalances, which is being tested on European credit card transactions over a span of two days. The findings are related to the three major algorithms with high performance fraud detection tasks: (a) linear support vector machine, random forest, and K-NN. As a consequence, the findings have shown that advanced generative sampling approaches will be short-lived in generalizing the minority community in the face of severe social imbalances.

Tang et al. [16] discussed holding out a comprehensive review of the different kinds of optimization methods for the generality as well as performance improvement of Spark. We're introducing Spark's programming style as well as the computing process, talking about the advantages and disadvantages of Spark, as well as investigating different fixes methods in the literature. In turn, we often implement different data compilation as well as analysis technologies, machine learning algorithms, as well as frameworks run by Spark. Last but not least, we're engaged in a discussion on problems that are open as well as obstacles for Spark.

Oosterlinck et al. [17] focused on a method that strengthens present techniques by establishing semi-synthetic novelties to be able to get the information labeled for each group. The strategy was used in a real-life test situation in which the goal was identifying illegal members to a family phone program. This particular analysis shows that the two-class specialist item outperforms the one class semi-synthetic sample version. The unit was tested on a certain dataset in the following stage.

Padhi et al. [18] discussed Machine learning methodologies have proven to be the most effective approach for anonymous transactions. This paper analyzes the basic machine learning algorithms that include SVM, LDA, QDA, DT, and RF for fraud detection. At the same time, some of the latest open-source boosting machine learning algorithms, including XGBoost, LGBoost, and CatBoost, are also introduced.

Liu et al. [19] focused on model the behavioral sequences as- signed to consecutive actions to identify sequential trends, while those that deviate from norms can be identified as fraud. In order to validate the usefulness of serial behavioral embeddings, we are working with a real-world telecommunications dataset of estimation and recognition activities de- pendent on experienced embeddings. Experimental results show that learned embeddings are better able to identify fraudulent behavior.

Kundu et al. [20] states that integrate both trend identification as well as abuse detection methods. With this post, they intend to make use of a two-stage matching technique in which the Profile Analyzer (PA) initially evaluates the resemblance of the new summary of purchases on the charge card to the actual cardholder's earlier spending sequences. They recommended one algorithm called BLAH and used it for the identification of charge card fraud. The system known as BLAHFDS detects fraudulent transactions with a Profile Analyzer as well as a Variance Analyzer.

Dal Pozzolo et al. [21] prepare & test a brand new learning method that correctly repairs category disparity, concept drift, as well as latency testing. Next, in the studies of theirs, they clearly show the impact of social inequality as well as paradigm drift on a data stream of over seventy-five million transactions in real life.

Phua et al. [22] discussed a contemporary multilayer detection process complemented by 2 extra layers: collective detection (Spike detection and cd) (SD). CD thinks genuine community interactions to rising the score of skepticism and it is responsive to digital community associations. Research results help support the hypothesis that useful recognition program fraud rates are actually abrupt and cause sharp spikes in duplicates. Even though this effort is different from the identification of credit program fraud, the concept of adaptation, combined with the adaptivity as well as consistency information explored in the post, is actually typical in the design, deployment, and assessment of all detection methods.

Omair and Alturki [23] introduced a full RNN architecture to play with explicitly modifying the GRU design to take into account the unpredictable time intervals between transactions. So, instead of introducing it as a function to the software, in the expectation that the model would learn how to view it correctly, they should exploit this knowledge and explicitly insert it in a process that updates the recurrent states.

Concerning the last few preliminary studies, which also use Medicare data with fraud labels, our research is more comprehensive in the breadth and depth of experimentation and results. A comprehensive discussion of the data and the mapping of the fraud labels is provided in this paper. Three different un-supervised algorithm were imlemented on four different class distributions to assess the effects of class imbalance. Finally, results were present using several different metrics and discuss the statistical significance of the results.

3. Methodology

3.1. Dataset description

Dataset consists of Inpatient claims, Outpatient claims, and beneficiary information from each provider, which are explained below [32].

- *Impatient data*: This report offers data on the cases related individuals admitted to hospitals. Further information such as admission and release dates are given as well as the code of diagnosis.
- *Outpatient data*: This report includes clear details about the expenses for all patients who visit and are not admitted to hospitals.
- *Beneficiary details data*: This data contains details from beneficiaries of the KYC, such as health issues, their region, etc.

When the dataset is labeled, supervised learning algorithms for distinguishing regular from outlier observations can be used. But in case of fraud, labeled class depict the situation in which we already know two classes of examples, i.e., we have specified which records are not fraud and which are fraudulent. Such situations are infrequent. On the other hand, when we don't know about the labeling of the dataset, which means there is no information about who has committed the fraud or not, unsupervised methods of outlier detection can be used. Figure 1 depicts the features distribution of first five features from the training dataset.

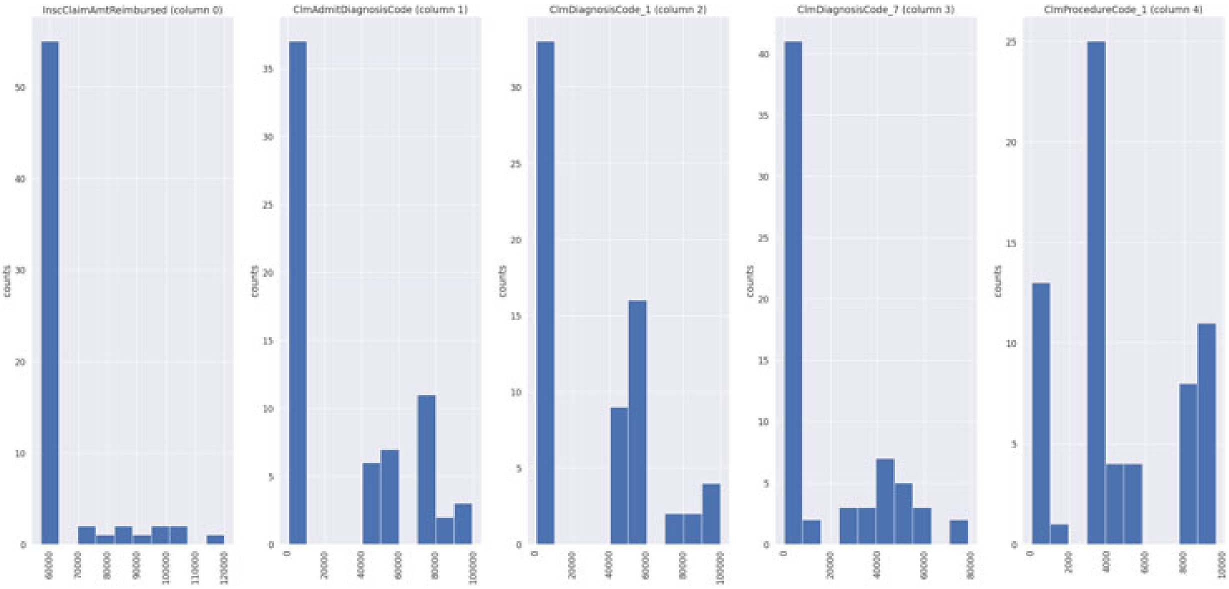

Fig. 1. Column plot of first 5 features.

3.2. Unsupervised algorithm

In this subsection, unsupervised algorithms used in the implementation process are discussed. Three different unsupervised methods: IF, K-Means, and LOF are used in the work. The key issue with this unsupervised model is that a subjective threshold needs to be determined. Limited knowledge regarding fraud cases are available, but still appropriate threshold was selected for a fraudulent event. Hence, when a forest of random trees collectively produce shorter path lengths for particular samples, they are highly likely to be anomalies. Isolated forest senses irregularities by splitting the domain space arbitrarily. Mainly, it functions like a Decision Tree Algorithm, where we start with a root node and proceed to partition the space. Just as with decision trees, where partition is centered on knowledge gain, we arbitrarily partition in isolation forest. Partitions are generated by choosing a feature randomly and then creating a branch function between the highest and lowest value feature arbitrarily. Further preceding of build the partitions is conducted until all points are isolated. In visualizing feature points in the available feature space, it was found that standard points tend to cluster more while anomaly points are far apart from each other. Hence on partitioning, the domain space anomaly will be found in fewer number partitions than a standard point. Forest isolation is an ensemble method. Many isolation trees (usually 100 trees are adequate) were built and the average of all the path lengths was taken. Finally, it was determine that whether or not a point is odd.

Figure 2 depicts the correlation matrix of the training datasets resembling 13 features. By selecting a function and randomly partitioning it an isolated tree can be created. Kmeans Algorithm is an iterative technique that attempts to segment the dataset into unique non-overlapping subgroups (bunches) where every point belongs to only one cluster. The less variety we have inside cluster, the more homogeneous (comparative) the points are inside a similar group. Working flow of the proposed methodology is demonstrated in Fig. 3.

The following equation contains anomaly value:-

$$S(x,n) = 2^{-E(h(x))/c(n)} \tag{1}$$

$$C(n) = 2H(n-1) - (2(n-1)/n) \tag{2}$$

where n – Number of data points and $c(n)$ – Average length of successful search in a BST. $E(h(x))$ – Average of path lengths from the Isolation forest.

4. Experiment and result analysis

The unsupervised algorithm was applied to the input patient's dataset. In this work, isolated forest, K-means, Local outlier algorithms were used for detection. Initially these algorithms were applied on the dataset having no outliers (after preprocessing and removing records of assumed fraud providers having high deductible claim amounts) and then training the model

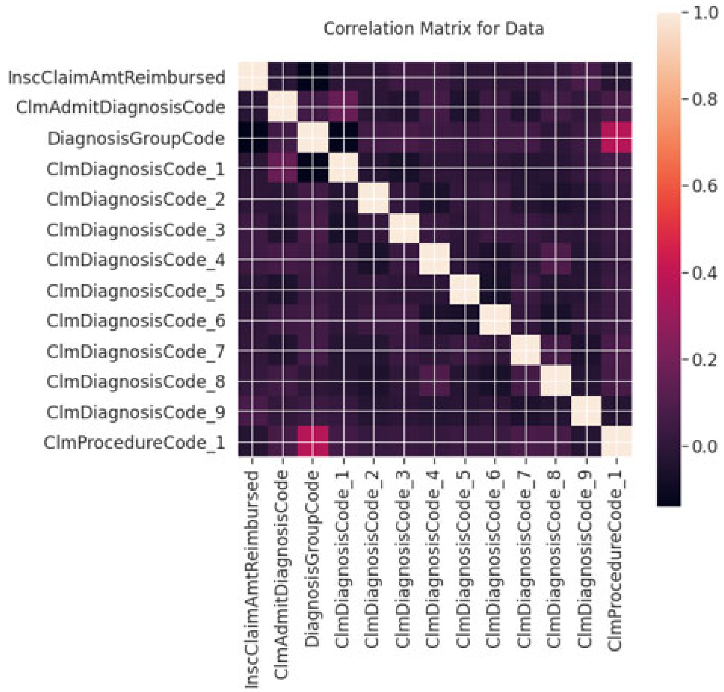

Fig. 2. Correlation matrix of training dataset.

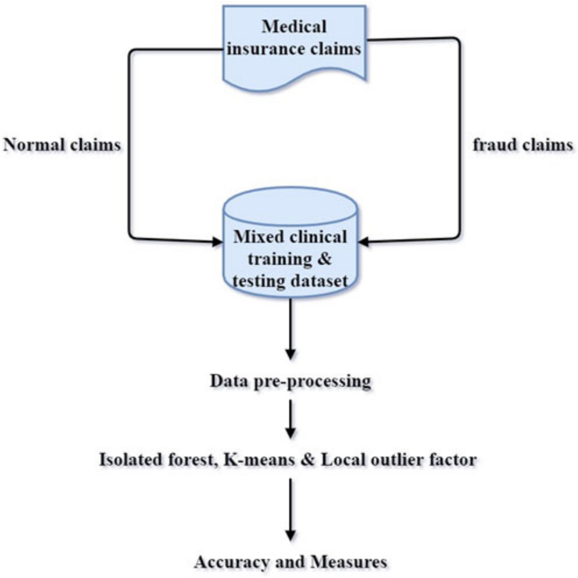

Fig. 3. Proposed methodology flowchart.

Algorithm 1: Anomaly prediction and outlier detection

```
Input: $D$ -Given dataset; $E$ - set of ensemble models;
       $G$ - Outlier generated from distribution;
Output: $P$ - Predicted anomalies on $D$;
        $Pm$ - Predicted anomalies on $D + G(M)$;
        $A$ -Accuracy in comparing $P$ and $Pm$; $R$ - roc
        accuracy score; $F$ - f1 score;
Start.
    $E$ := Isolated forest;
    $Pm$ := E.fit(M); $P$:- E.fit(D);
    $A$ := accuracy score(P m,P);
    $R$ := roc auc score(P m,P);
    $F$ := f1 score(P m,P);
    $E$ ::= K means;
    $Pm$ ::= E.fit(M); $P$:- E.fit(D);
    $A$ := accuracy score(P m,P);
    $R$ := roc auc score(P m,P);
    $F$ := f1score(Pm,P);
    $E$ := Local Outlier Factor;
    $Pm$ ::= E.fit(M); $P$:- E.fit(D);
    $A$ := accuracy score(P m,P);
    $R$ := roc auc score(P m,P);
    $F$ := f1score(P m,P);
End.
```

with the same dataset. In another case, focus was given on the assumed sample and again the same algorithms were applied and further testing was performed. The results are then compared using five evaluation metrics, i.e. Accuracy, roc_auc, f1_score, Cohen's Kappa, Average_Precision. Results show that in case of accuracy, the isolation forest algorithm performs well. Overall it is observed that isolated forest has given better results compared to other algorithms.

The system configuration used for experimenting was the Intel Core i3 processor, Windows 7 Ultimate 64-bit operating system, and 6 GB of RAM. The tool

Algorithm 2: Isolated forest

```
Algorithm 2.1 - iforest($X, t, u$)
  Input: $X$ – input data, $t$ – number of trees, $u$ – sub-sampling
         size
  Output: a set of $t$ iTrees
  Initialise Forest
  Set height limit 1 = ceiling($\log_2 u$)
  for $i = 1$ to $t$ do
      $X \Leftarrow$ sample($X, u$)
      Forest $\Leftarrow$ Forest $U$ iTrees('X', 0, l)
  end for
  return Forest
Algorithm 2.2 - iTree('X', e, l)
  Input: $X$ – input data; $e$ – current tree height, $l$ – height limit
  Output: an iTree
  if $e \geqslant l$ or $|X| \leqslant 1$ then
      return exNode$Size \Leftarrow |X|$
  else
      let $Q$ be list of attributes in $X$
      Randomly select an attributes $q \sim Q$
      Randomly select a split point $p$ from max and min
      Values of attributes $q$ in $X$
      $X_l \Leftarrow$ filter($X, q < p$)
      $X_r \Leftarrow$ fiter($X, q \geqslant p$)
      return Node Left $\Leftarrow$ iTree($X_l, e+1, l$), Right $\Leftarrow$
        iTree($X_r, e+1, l$), SplitAtt $\Leftarrow q$, SplitValue $\Leftarrow p$
  end if
Algorithm 2.3 – PathLength($x, T, e$)
  Input     : x
  – aninstance, T – an iTree, e – current path length to be
    initialized to 0
  Output: Path length of x
  if T is an external node then
      return e + c(T.size)
  end if
  a $\Leftarrow$ T.splitAtt
  if $x_a$ < T.splitValue then
      return PathLength(x.T.left, e+1)
  else
      $x_a$ >= T.splitValue
      return PathLength(x.T.right, e+1)
  end if
```

Table 2
Dataset description

Dataset	No of instances
Training dataset	6376
Testing dataset	1534
Outlier train dataset	66
Outlier test dataset	195

Table 3
Performance metrics obtained after applying isolated forest

Evaluation metrics	Result (%)
Accuracy	98.76
Roc_Auc	89.21
F1_Score	97.62
Cohen's Kappa	79.42
Average_Precision	97

Table 4
Performance metrics obtained after applying k-means

Evaluation metrics	Result (%)
Accuracy	95.56
Roc_Auc	73.06
F1_Score	51.06
Cohen's Kappa	48.72
Average_Precision	29

Table 5
Performance metrics obtained after applying local outlier factor

Evaluation metrics	Result (%)
Accuracy	98.56
Roc_Auc	94.32
F1_Score	86.41
Cohen's Kappa	85.66
Average_Precision	76

Algorithm 3: K-means

```
Initialise cluster centroide $u_1, u_2, u_3, \ldots, u_k \sim R^n$ randomly,
Repeat until convergence: {
  for every $i$, set $c^i :=$ arg $\min_j \|c^(i) - u_j\|^2$;
    For each $j$, set $u_j := \frac{\sum_{i=1}^{m} (c(i) = j)x^i}{\sum_{i=1}^{m} c(i) = j}$;
}
```

Algorithm 4: Local outlier factor (LOF)

```
Input: $k$ – samples; $dp$ – Data points;
Output: LOF: Predicted outliers;
kDistannce(d$p$, $pt$);
reachabilityDist($pt$);
LOF := null;
for each point pt do
    KNN := KDistance(d$p$, $k$);
    Lrd = reachabilityDist(KNN, k);
    for each p in KNN do
        Lofhold := sum(lrd[o ~ N(p)])/lrd[i]|N(p)|;
        Lof = max(LOF, lofhold);
    end for
end for
return LOF
```

used was Jupiter Notebook. Table 2 depicts the dataset distribution along with its sample instances for both training and testing datasets. Figures 4–6 represents the confusion matrix obtained after applying Isolated Forest, K-means, and Local Outlier Factor, respectively. Five metrics were used to calculate the efficiency of the proposed approach. Tables 3–5 shows the performance metrics obtained after applying Isolated Forest, K-means, and Local Outlier Factor respectively.

$$\text{Accuracy} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{TN} + \text{FP} + \text{FN}} \tag{3}$$

$$F1 = 2 \times \frac{(\text{precision} \times \text{recall})}{(\text{precision} \times \text{recall})} \tag{4}$$

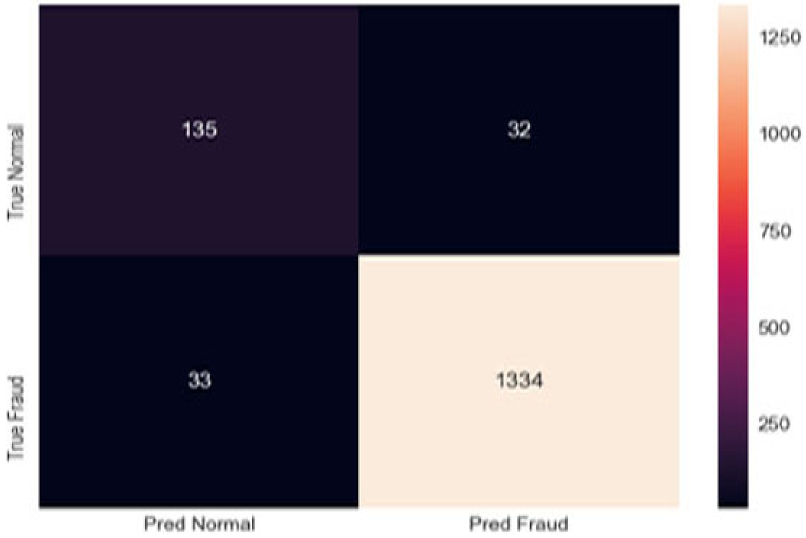

Fig. 4. Confusion matrix using isolated forest.

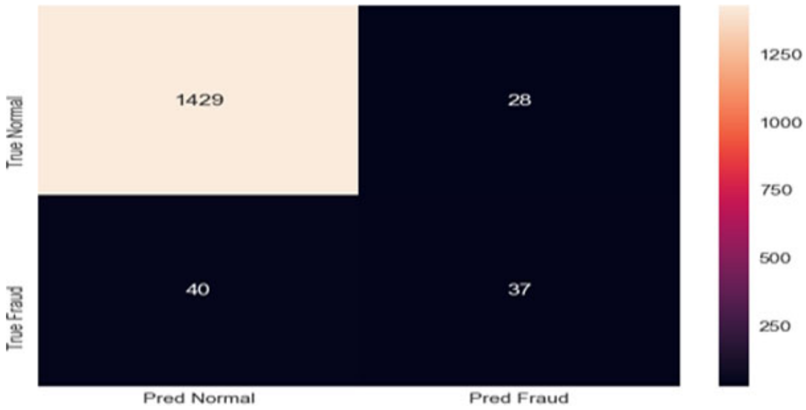

Fig. 5. Confusion matrix using k-mean algorithm.

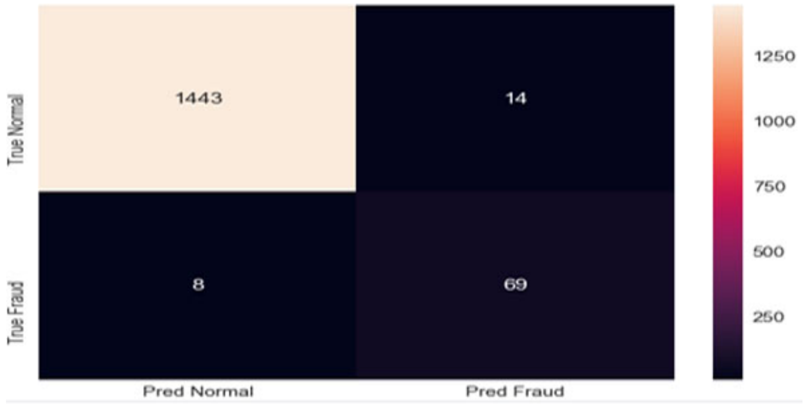

Fig. 6. Confusion matrix using local outlier factor.

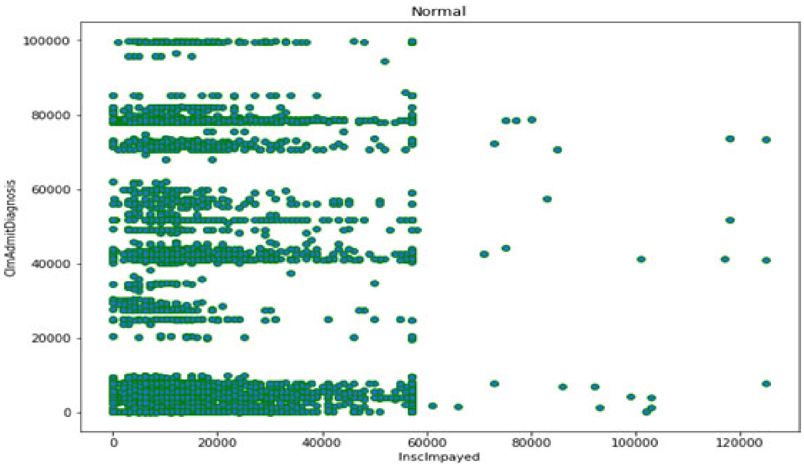

Fig. 7. Distribution of normal popuation.

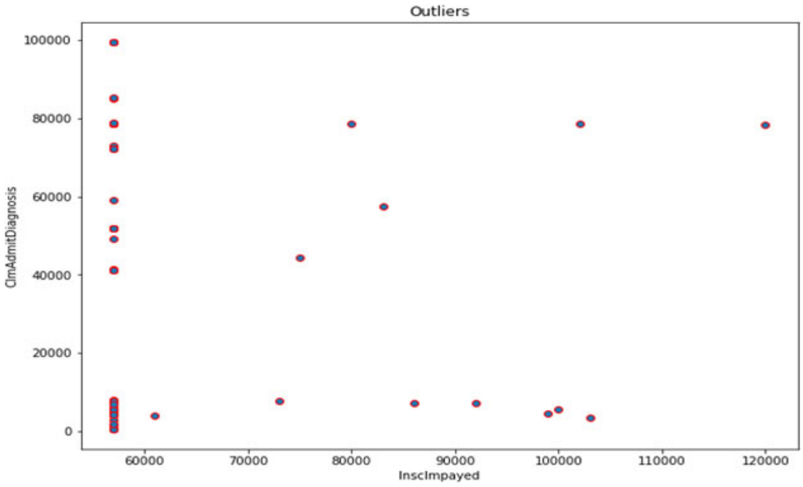

Fig. 8. Distribution of outliers.

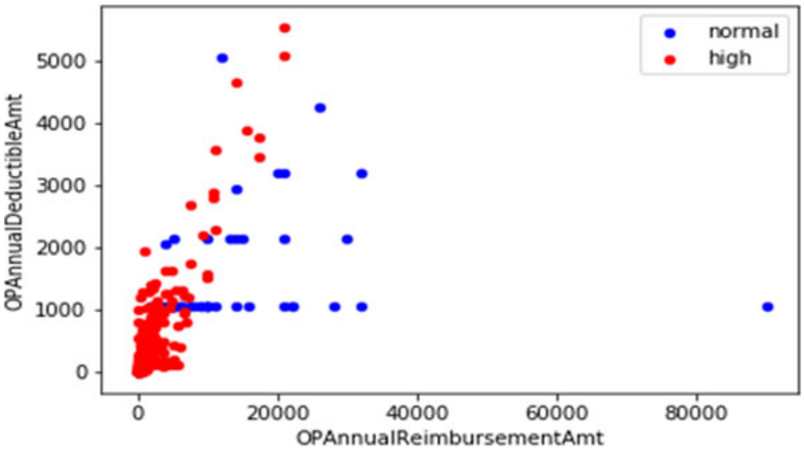

Fig. 9. Sample population of beneficiary details dataset plot.

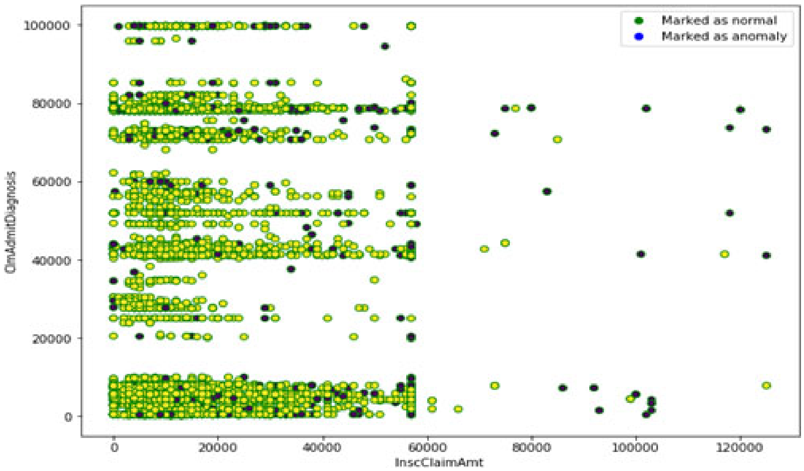

Fig. 10. Represent the population using a normal v/s anomaly plot after implementing the isolated forest unsupervised algorithm.

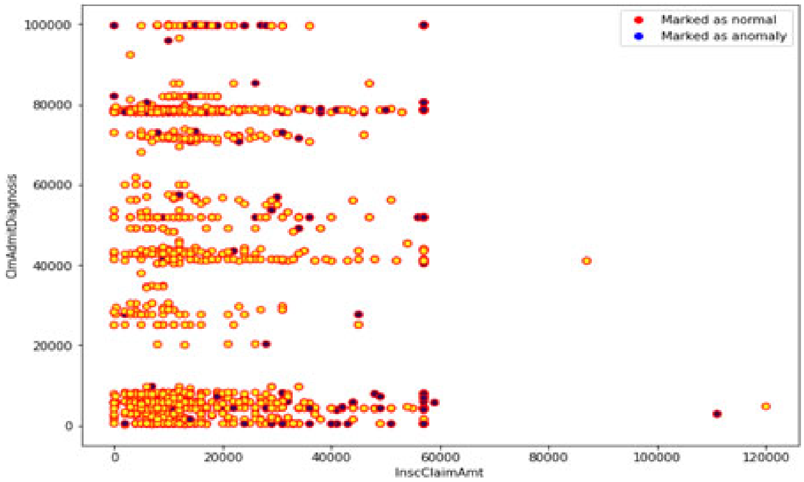

Fig. 11. Normal v/s anomaly plot on testing dataset.

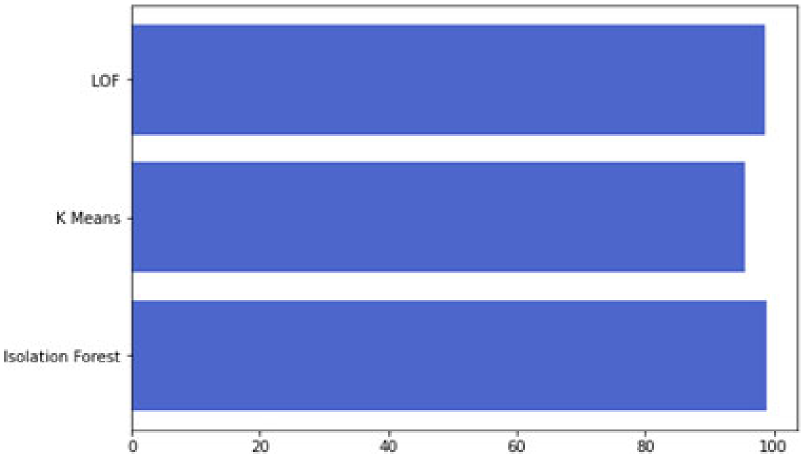

Fig. 12. Accuracy comparison.

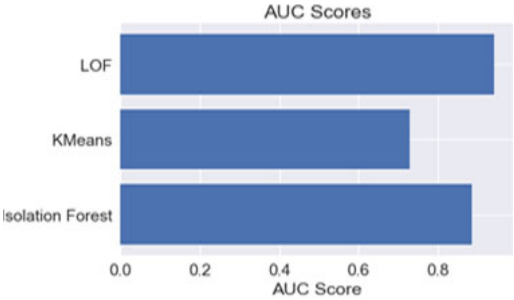

Fig. 13. AUC scores comparison.

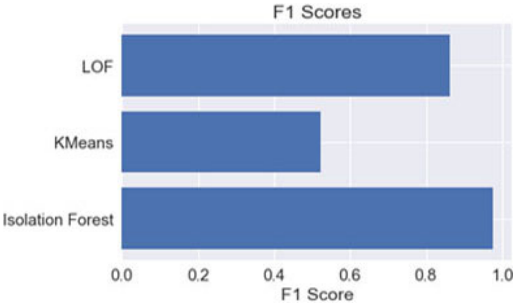

Fig. 14. F1 score comparison.

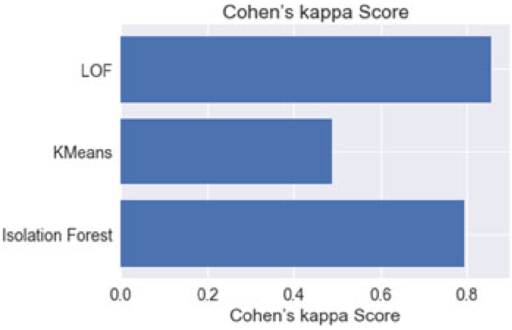

Fig. 15. Cohen's Kappa score comparison.

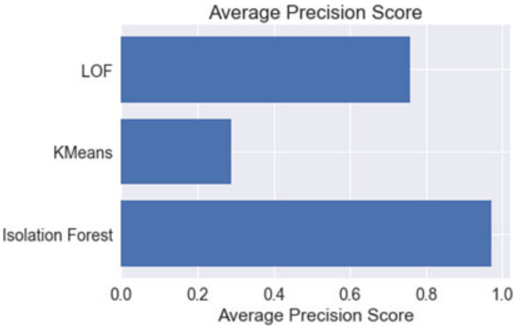

Fig. 16. Average precision score comparison.

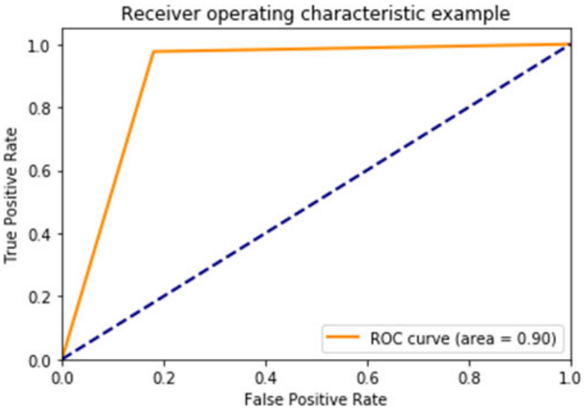

Fig. 17. ROC curve using isolated forest.

Distribution of training sample is depicted in Figs 7 and 8. Represents the outliers available in the training dataset. Figure 7 consists of scatter plot of health insurance patients who have claim the amount in the year of in the year of 2012–14 as given in the dataset. The plot shows the distribution of insurance claim given to claims of admitted patients under diagnosis. Clearly plot is denser in initial region showing normal patients claim and with some outliers having higher pay of claims. Figure 8 consists of plot representing potential outliers having higher insurance claim amount. Sample population of beneficiary details in dataset was plotted in Fig. 9. Plot shows the randomly distribution of data and the normal claims with suspects of fraudulent claims. The following plot is plotted with the help of target class Renal Disease Indicator. Blue dots are showing patients not having disease and redone are those which are positive. It is only to indicate patients having deductible amounts high and are not positive with disease may be case of fraud. Figure 10 represent the population using after implementing the isolated forest unsupervised algorithm on training the mixture. Training the algorithm with the following data set will give the predicted outliers in multi-class problem which then be used for calculating accuracy with the prediction from normal patient's outliers. Blue dots as legend says are marked as anomaly and green dots are normal ones.

Figure 11 shows then plot the normal v/s anomaly plot on the testing dataset to show the efficiently algorithm work. The following plot shows testing estimates of the isolated forest algorithm, and the red dots are predicted normal patients and red dots are predicted fraudulent once. Clearly, blue dots are in small in quantity which shows algorithm is working efficiently. Comparative analysis of all the 3 algorithms was done

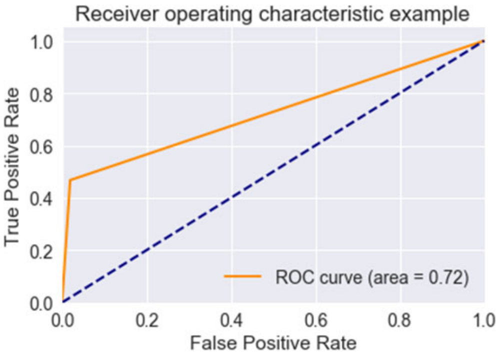

Fig. 18. ROC curve using k-means.

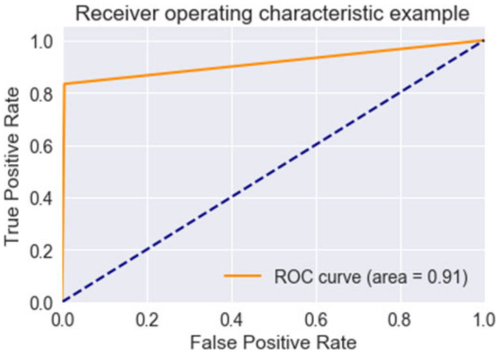

Fig. 19. ROC curve using LOF.

based upon Accuracy, AUC Score and F1-Score, Cohen's Kappa, Average Precision in the Figs 12–16 respectively. ROC Curve of all the three algorithms were plotted in Figs 17–19 respectively.

5. Conclusion and future work

Medical healthcare is very crucial for our society, but lot of fraud cases take place in this field. Existing algorithms which can identify these fraud cases need to be modified for enhancing the results so that the equal and deserving patients can utilize the health care benefits. The proposed system tries to identify the outliers available in the dataset using three different algorithms. Among all of the implemented algorithms, Isolated Forest has given the most promising results. In Future, various hybrid or ensemble algorithms can be applied for improving the accuracy.

References

[1] Aisha A, Mohd Aizaini M, Zainal A. Fraud detection system: A survey. Journal of Network and Computer Applications in Elsevier. 2016.

[2] Khamparia A, Pande S, Gupta D, Khanna A, Sangaiah AK. Multi-level framework for anomaly detection in social networking. LHT. 2020; 38(2): 350-366. doi: 10.1108/lht-01-2019-0023.

[3] Johnson ME, Nagarur N. Multi-stage methodology to detect health insurance claim fraud. Health Care Manag Sci. 2015; 19(3): 249-260. doi: 10.1007/s10729-015-9317-3.

[4] Nsiah-Boateng E, Aikins M, Asenso-Boadi F, Andoh-Adjei FX. Value and service quality assessment of the national health insurance scheme in Ghana: Evidence from ashiedu keteke district. Value in Health Regional Issues. 2016; 10: 7-13. doi: 10.1016/j.vhri.2016.03.003.

[5] Bauder RA, Khoshgoftaar TM. Medicare fraud detection using machine learning methods. in: 2017 16th IEEE International Conference on Machine Learning and Applications (ICMLA); 2017. 858-865.

[6] Bauder RA, Khoshgoftaar TM. The detection of medicare fraud using machine learning methods with excluded provider labels. Artificial Intelligence Research Society Conference. 2018.

[7] Abdallah A, Maarof MA, Zainal A. Fraud detection system: A survey. Journal of Network and Computer Applications. 2016; 68: 90-113. Available from: https://www.sciencedirect.com/science/article/pii/S1084804516300571.

[8] Fursova I, Zaytseva A, Khasyanova R, Spindlerb M, Burnaeva E. Sequence embeddings help to identify fraudulent cases in healthcare insurance. Preprint submitted to Journal of Econometrics. 2018.

[9] Farbmacher H, Löw L, Spindler M. An explainable attention network for fraud detection in claims management. Journal of Econometrics. 2020; Available from: https://www.sciencedirect.com/science/article/pii/S0304407620302852.

[10] Kruthika S, Manjunatha S. A survey on SMOTE deep: Novel link based classifier for fraud detection. International Journal of Computer Science Engineering Techniques. 2020; 5.

[11] Resa A, Hariman B. Fraud awareness in indonesian governmental Sector: Multi-agency responses. Fraud Awareness in Indonesian Governmental Sector: Multi-Agency Responses. 2020.

[12] David A, Ricardo B, João B, João TA, Pedro B. ARMS: Automated rules management system for fraud detection. Computer Science. 2020.

[13] Song R, Huang L, Cui W, Óskarsdóttir M, Vanthienen J. Fraud detection of bulk cargo theft in port using bayesian network models. Applied Sciences. 2020; 10: 1056.

[14] Lucas Y, Portier PE, Laporte L, He-Guelton L, Caelen O, Granitzer M, et al. Towards automated feature engineering for credit card fraud detection using multi-perspective HMMs. Future Generation Computer Systems. 2020; 102: 393-402. Available from: https://www.sciencedirect.com/science/article/pii/S0167739X19300664.

[15] Janbandhu R, Begum S, Ramasubramanian N. Credit card fraud detection. in: Iyer B, Deshpande PS, Sharma SC, Shiurkar U. Computing in Engineering and Technology. Singapore: Springer Singapore. 2020; 225-238.

[16] Shanjiang T, Bingsheng H, Ce Y, Yusen L, Kun L. A survey on spark ecosystem: Big data processing in- frastructure, machine learning, and applications. IEEE Transactions on Knowledge and Data Engineering.

[17] Oosterlinck D, Benoit DF, Baecke P. From one-class to two-class classification by incorporating expert knowledge: Novelty detection in human behaviour. Eur J Oper Res. 2020; 282(3): 1011-1024. doi: 10.1016/j.ejor.2019.10.015.

[18] Padhi B, Chakravarty S, Biswal B. Anonymized credit card transaction using machine learning techniques. Advances in Intelligent Computing and Communication Lecture Notes in Networks and Systems. 2020; 109.

[19] Liu G, Guo J, Zuo Y, Wu J, Guo RY. Fraud detection via behavioral sequence embedding. Knowl Inf Syst. 2020; 62(7): 2685-2708. doi: 10.1007/s10115-019-01433-3.

[20] Kundu A, Panigrahi S, Sural S, Majumdar AK. BLAST-SSAHA hybridization for credit card fraud detection. IEEE Trans Dependable and Secure Comput. 2009; 6(4): 309-315. doi: 10.1109/tdsc.2009.11.

[21] Dal Pozzolo A, Boracchi G, Caelen O, Alippi C, Bontempi G. Credit card fraud detection: A realistic modeling and a novel learning strategy. IEEE Transactions on Neural Networks and Learning Systems. 2018; 29(8): 3784-3797.

[22] Phua C, Smith-Miles K, Lee V, Gayler R. Resilient identity crime detection. IEEE Trans Knowl Data Eng. 2012; 24(3): 533-546. doi: 10.1109/tkde.2010.262.

[23] Omair B, Alturki A. A Systematic literature review of fraud detection metrics in business processes. IEEE Access. 2020; 8: 26893-26903.

[24] Agrawal U, Arora J, Singh R, Gupta D, Khanna A, Khamparia A. Hybrid wolf-bat algorithm for optimization of connection weights in multi-layer perceptron. ACM Transactions on Multimedia Computing, Communications, and Applications. 2020; 16(1s): 1-20. doi: 10.1145/3350532.

[25] Khamparia A, Gupta D, de Albuquerque VHC, Sangaiah AK, Jhaveri RH. Internet of health things-driven deep learning system for detection and classification of cervical cells using transfer learning. The Journal of Supercomputing. 2020; 76(11): 8590-8608. doi: 10.1007/s11227-020-03159-4.

[26] Khamparia A, Pandey B. Threat driven modeling framework using petri nets for e-learning system. Springer Plus. 2016; 5(1). doi: 10.1186/s40064-016-2101-0.

[27] Khamparia A, Singh KM. A systematic review on deep learning architectures and applications. Expert Systems. 2019; 36(3): e12400. doi: 10.1111/exsy.12400.

[28] Khamparia A, Saini G, Pandey B, Tiwari S, Gupta D, Khanna A. KDSAE: Chronic kidney disease classification with multimedia data learning using deep stacked autoencoder network. Multimed Tools Appl. 2019; 79(47-48): 35425-35440. doi: 10.1007/s11042-019-07839-z.

[29] Chouhan V, Singh SK, Khamparia A, Gupta D, Tiwari P, Moreira C, et al. A novel transfer learning based approach for pneumonia detection in chest X-ray images. Applied Sciences. 2020; 10(2): 559. doi: 10.3390/app10020559.

[30] Khamparia A, Singh A, Anand D, Gupta D, Khanna A, Kumar NA, et al. A novel deep learning-based multi-model ensemble method for the prediction of neuromuscular disorders. Neural Computing and Applications. 2018; 32(15): 11083-11095. doi: 10.1007/s00521-018-3896-0.

[31] Pande S. An information security scheme for cloud based environment using 3DES encryption algorithm. International Journal of Recent Development in Engineering and Technology. 2014; 2: 65-68.

[32] Available from: https://www.kaggle.com/rohitrox/healthcare-provider-fraud-detection-analysis.