



International Journal of Bio-Inspired Computation

ISSN online: 1758-0374 - ISSN print: 1758-0366

<https://www.inderscience.com/ijbic>

Improved Dirichlet mixture model clustering algorithm for medical data anomaly detection

Lili Wu, Majid Khan Majahar Ali, Fam Pei Shan, Ying Tian, Li Tao

DOI: [10.1504/IJBIC.2024.10064803](https://doi.org/10.1504/IJBIC.2024.10064803)

Article History:

Received:	20 November 2023
Last revised:	25 December 2023
Accepted:	14 January 2024
Published online:	03 January 2025

Improved Dirichlet mixture model clustering algorithm for medical data anomaly detection

Lili Wu

Department of Computer Science,
Xinzhou Teachers University,
Xinzhou, 034000, China
and
School of Mathematical Sciences,
Universiti Sains Malaysia (USM),
Pulau Pinang, 11800, Malaysia
Email: wulili@student.usm.my

Majid Khan Majahar Ali* and Fam Pei Shan

School of Mathematical Sciences,
USM,
Pulau Pinang, 11800, Malaysia
Email: majidkhanmajaharali@usm.my
Email: fpeishan@usm.my
*Corresponding author

Ying Tian

Department of Mathematics,
Taiyuan Institute of Technology,
Taiyuan, 030024, China
Email: tian29maranda@student.usm.my

Li Tao

School of Mathematical Sciences,
USM,
Pulau Pinang, 11800, Malaysia
Email: litao@student.usm.my

Abstract: In order to address the issue of identifying over-diagnosis and anomaly expenses in the healthcare service process, a local outlier mining clustering algorithm (ILOF-DPMM) is proposed by combining the clustering-based local outlier factor (CBLOF) algorithm with Dirichlet mixture model (DPMM). By extracting the patient's hospitalisation records from the medical record homepage, the influencing factors of hospitalisation costs for different disease types are classified, and the random forest method is used to reduce the feature dimension by disease type. The feature extraction and dimensionality reduction methods adopted by this algorithm effectively cluster medical insurance expense data. When calculating the LOF value of data, using a weighted calculation method based on the similarity of discrete and continuous features can more accurately detect abnormal data points in the data set, and has the ability to detect new data in real time, thus improving detection accuracy and efficiency.

Keywords: over-diagnosis; anomaly expenses; anomaly detection; DPMM; CBLOF.

Reference to this paper should be made as follows: Wu, L., Ali, M.K.M., Shan, F.P., Tian, Y. and Tao, L. (2025) 'Improved Dirichlet mixture model clustering algorithm for medical data anomaly detection', *Int. J. Bio-Inspired Computation*, Vol. 25, No. 1, pp.11–21.

Biographical notes: Lili Wu graduated from Soochow University in Mathematics and holds a Master of Science in Mathematics from Soochow University. She is pursuing a PhD at USM. Her research interests include big data analysis and algorithm optimisation.

Majid Khan Majahar Ali is a Senior Lecturer at the School of Mathematics, USM. He holds degrees: BSc in Mathematics with Economic (UMS), and PhD in Mathematics (UMS). His topic research interests include leveraging the application of IoT, big data and simulation methods to improve model predictions of moisture loss during drying processes in controlled and uncontrolled environments.

Fam Pei Shan is a Senior Lecturer at the School of Mathematics, USM. She holds degrees in Mathematics: BSc (UM), MSc (UM), PhD in Mathematics (UM). Her research interests include categorical data analysis, and logistics regression analysis.

Ying Tian graduated from Shanxi University in Mathematics and obtained an MSc in Statistics from Shanxi University. She is currently pursuing a PhD at USM. Her research interests include missing value imputation and data pre-analysis.

Li Tao is an Associate Professor at Weifang University of Science and Technology. He holds degrees: BSc in Mathematics from Weifang University, MSc in Mathematics Teaching from Ludong University. He is currently pursuing a PhD in Applied Statistics from USM. His research interests are optimisation evaluation and business economic analysis.

1 Introduction

Health insurance is an insurance plan that utilises medical funds to compensate for medical expenses and fulfill the healthcare needs of insured individuals. Within both domestic and international social security systems, the healthcare insurance system plays a crucial role in facilitating public access to medical treatment. Health insurance fraud (Villegas-Ortega et al., 2021) has resulted in global public health funds losing approximately 260 billion US dollars, accounting for 6% of global healthcare expenditures. Current international statistical data indicates that 30% of health insurance reimbursement records exhibit potential fraudulent activities annually. In 2019, healthcare expenditures in the USA increased by 3.9%, surpassing 36 billion US dollars, with approximately 20% attributed to improper payments related to healthcare. Health insurance (Yabroff et al., 2019). In China, billions of dollars are lost each year due to healthcare insurance fraud (Tang, 2018). Therefore, in order to ensure the long-term normal operation of healthcare funds, it is necessary to strengthen the investigation and prosecution of healthcare insurance fraudulent activities.

With the rapid advancements in fields such as machine learning and pattern recognition, the domain of anomaly detection has witnessed significant progress (Kalidindi and Arrama, 2023). Currently, there are two major flaws in medical insurance anomaly detection:

- 1 rule-based detection relies heavily on the subjective judgment and experience of medical professionals, introducing a subjective element.
- 2 Sampling-based detection fails to achieve exhaustive detection, leading to incomplete review and potential oversight of anomalous data, resulting in significant losses.
- 3 New data are constantly generated, requiring real-time updates (Wu et al., 2022; Guo et al., 2023) for data analysis.

The aim of this study is to propose a clustering-based anomaly detection algorithm to address the issues of overdiagnosis and excessive spending in healthcare services. The approach presented in this paper can assist healthcare institutions in identifying and addressing potential cases of overdiagnosis. It holds significant implications for enhancing healthcare service efficiency and reducing medical costs. The main contributions are as follows:

- Distinguishing between discrete and continuous attributes in the medical record data.
- Using the random forest method (Huang et al., 2023), extracting features from discrete attributes by dividing them into different cells based on the main discharge diagnosis results.
- Organising continuous cost data by financial categories within each cell, achieving attribute reduction.
- Introducing an improved DPMM (Ding and Karabatsos, 2021; Elliott and Xia, 2022) clustering algorithm combined with the LOF method (Yu et al., 2023) to partition medical insurance cost data into clusters and calculate the anomaly degree values for each data point.

The following chapters will be organised as follows: Chapter 2 will review relevant research and analyse the shortcomings of existing methods; Chapter 3 will provide a detailed introduction to the proposed ILOF-DPMM algorithm; Chapter 4 will present the experimental results; Finally, Chapter 5 will summarise the research findings and explore future research directions.

2 Related work and basic concepts

Currently, research on intelligent healthcare insurance data anomaly detection based on big data mainly focuses on two aspects: firstly, tracking and analysing medical advice and

all patient medical records. Methods such as semantic analysis (Vakulenko, 2023), association rules (Kou et al., 2023), and time series analysis (Xiao et al., 2023a) can be used to detect anomaly patient medical behaviours, doctor prescriptions, and other examinations.

2.1 Literature review of previous studies

In addressing the issue of highly imbalanced medical insurance big data, Li et al., (2023) utilised TextRank to construct a medical dictionary for segmenting surgical process texts and extracting surgical keywords, creating keyword combination patterns as the final detection rules. This model relies on a keyword dictionary and necessitates a significant amount of historical experience and rules. As the amount of data increases, the audit rules call for continuous refinement. Additionally, Xiao et al. (2023b) proposed a graph-based self-supervised LSTM sequence model for comparative learning framework in medical insurance fraud detection. However, obtaining medical records data with temporal features proves to be challenging.

However, the above method belongs to supervised learning and can only be applied to labelled datasets for fraud detection. It cannot effectively detect anomalies in unlabelled data. Therefore, the following section will introduce unsupervised anomaly detection methods. In the absence of labelled Medicare data, potential deviant sexual behaviours within the Medicare system often revolve around three key actors: patients, physicians, and healthcare providers. Kemp et al. (2022) employed sequential pattern mining to identify and group comparable treatment procedures, while also computing previously unknown patterns of abnormal claims, without focusing on overtreatment and excessive billing. Xu et al. (2022) applied dimensionality reduction and deep semi-nonnegative matrix factorisation clustering analysis to detect overtreatment and repeated healthcare in insurance data. However, the clustering results merely highlighted potential categories of abnormal data without providing specific anomaly detection. In the work by Liu et al. (2023) an evolutionary game theory model was established, taking into account multiple factors and focusing on detecting fraudulent behaviour between doctors and patients. However, the model was not validated using real data. Current methods for detecting insurance anomalies heavily rely on labelled data and manual rules, often overlook the influence of various medical expenses and medical record information on different diseases. Therefore, comprehensive detection cannot be achieved and accurate anomaly scoring cannot be provided.

Based on the above-mentioned issue, this study employs unsupervised clustering algorithms to detect anomalies within unlabelled data. Our algorithm dynamically updates the number of clusters based on the data, resulting in not only improved detection precision but also enabling real-time detection (Xin et al., 2022).

2.2 Unsupervised clustering method

DPMM is a Bayesian nonparametric model used for clustering and unsupervised learning problems. Compared to traditional clustering algorithms, DPMM can automatically determine the number of clusters without the need for prior specification. The Dirichlet process serves as the prior distribution for the data. Let X be the set of observed data. DPMM clusters the data points by assigning them to probability density functions represented by $f(\theta)$, θ represents the parameters of each cluster. DPMM can be represented by the following generative model:

$$G \sim DP(\alpha_0, G_0) \quad (1)$$

$$\theta_i | G \sim G, x_i | \theta_i \sim F(\theta_i) \quad (2)$$

This paper adopts the Chinese restaurant process (CRP) for Gibbs sampling. The specific clustering of each data point x_i is determined by an indicator variable z_i . According to the Bayes' theorem, the conditional distribution concerning can be expressed as:

$$\begin{aligned} p(z_i | x_1, \dots, x_N, Z_{\setminus i}, \lambda, \alpha_0) \\ \propto p(z_i | Z_{\setminus i}, \alpha_0) p(x_i | z_i, \dots, z_N, X_{\setminus i}, \lambda) \end{aligned} \quad (3)$$

where $Z_{\setminus i}$ represents the remaining data group when z_i is removed from $Z = \{z_1, z_2, \dots, z_n\}$, $p(z_i | Z_{\setminus i}, \alpha_0)$ can be expressed using the CRP in the Dirichlet process. If $Z_{\setminus i}$ already has K categories, with the number of observed data points in each category denoted as $n_{\setminus i}$, the first term of equation (3) can be written as:

$$\begin{aligned} p(z_i | Z_{\setminus i}, \alpha_0) \propto \sum_k^K \frac{n_{\setminus i}^k}{n - 1 + \alpha_0} \cdot \delta(z_i, k) \\ + \frac{\alpha_0}{n - 1 + \alpha_0} \cdot \delta(z_i, \bar{k}) \end{aligned} \quad (4)$$

The second term is calculated as follows: If the i^{th} observation chooses the k^{th} class, i.e., $z_i = k$, then:

$$p(x_i | z_i = k, x_{\setminus i}, \lambda) = \frac{\int f(x_i | \theta) \prod_{z_j=k, j \neq i} f(x_j | \theta) g(\theta | \lambda) d\theta}{\int \prod_{z_j=k, j \neq i} f(x_j | \theta) g(\theta | \lambda) d\theta} \quad (5)$$

If x_i chooses a new class, i.e., $z_i = \bar{k}$. Then:

$$p(x_i | z_i = \bar{k}, x_{\setminus i}, \lambda) = p(x_i | \lambda) = \int f(x_i | \theta) g(\theta | \lambda) d\theta \quad (6)$$

where $\bar{k}$ represents the $K + 1^{\text{th}}$ class.

3 Local outlier factor mining algorithm based on improved DPMM

Currently available anomaly detection algorithms include distance-based anomaly detection (Tayeh et al., 2020), kernel local outlier factor (Wang et al., 2021), and CBLOF detection (Premkumar et al., 2023). Propose an improved

DPMM unsupervised clustering algorithm for anomaly detection in medical insurance data. This method combines improved DPMM and CBLOF method (Pan and Zhang, 2017), aiming at data quality management, fraud detection, and prediction of abnormal behaviour. The method first applies improved DPMM for clustering medical insurance data. Subsequently, the CBLOF method is utilised to detect anomalies within the clustering results. CBLOF is a clustering-based anomaly detection method, measuring the degree of anomaly by calculating the similarity between data points and each cluster. These identified anomalies can be utilised for data quality management, fraud detection, and prediction of abnormal behaviour, providing crucial insights for data analysis and decision-making.

The aim of this study is to identify potential violations of medical insurance regulations in the field of healthcare billing information systems. The key steps involved are as follows: Importing the initial medical record data for auditing and quality control scoring. Consolidating medical records that meet the quality control score criteria into disease groups based on the discharge diagnosis and major operation/procedure codes. Reducing cost databased on the principles of medical record cost classification. Utilising DPMM to cluster the restructured data. Lastly, calculating abnormal scores based on the clustering results to identify suspicious medical records.

3.1 Data pre-processing

The method of data pre-processing depends on the characteristics of the data and the research objectives. Initially, considering that the data originates from medical records, it is necessary to assess the data quality through a quality control rating system. The quality control is designed with a minimum pass score (which can be parameterised and configured, and for this study, the pass score is set at 75). Subsequently, the medical records that meet the quality control standards are divided into three parts: patient basic information, treatment process, and medical insurance expenses. These parts are then processed separately according to the research objectives. Given that our main focus is on overdiagnosis and abnormal detection related to medical insurance expenses, it is understood that surgical decisions have a significant impact on medical insurance expenses. Therefore, disease groups are generated based on discharge diagnosis and initial surgery/procedure codes, and the data is processed and analysed according to different disease categories.

During the data pre-processing stage, the initial step involves addressing missing values. Upon examination of the selected dataset, it is observed that essential patient information and cost-related features have minimal missing values. Basic information falls under discrete features. With the scarcity of missing data, the data with missing values will be removed. On the other hand, cost information is a continuous feature with nearly no missing values, and we opt to fill in the missing values using the average cost for the same type of disease. However, there are considerable missing values in the diagnosis and treatment procedures.

We identify and remove the attributes of all missing data for the same disease. Moreover, there may still be remaining missing values in the subsequent processes. To address this issue, we have filled in the missing values in the treatment names or numbers based on the patient's basic information and the standards of the National Hospital Medical Record management and quality control centre. For data that cannot be filled in, we choose to delete it.

In the second step of feature extraction, different calculation methods are utilised for discrete and continuous features when computing anomaly scores, necessitating distinct feature extraction approaches. Since both basic information and medical procedure data possess discrete features, a random forest method is employed to combine them, with total expenditure as the target variable, extracting the primary influencing factors. The cost features processed in the first step are further subjected to dimensionality reduction (Shu et al., 2023) and clustering using an improved DPMM method. For discrete attributes, a distance calculation based on support is employed, while for continuous attributes, distance calculation is performed using the density function sampled by DPMM.

Through these pre-processing methods and distance calculation approaches, we are able to better prepare the data and provide more useful features and appropriate data representation for subsequent anomaly detection.

3.2 Improved DPMM

In dealing with high-dimensional data using DPMM, the challenges in computation and storage necessitate more efficient algorithms and techniques. In this article, we employ an improved approach known as TSNE-DPMM. Initially, we apply t-distributed stochastic neighbour embedding (t-SNE) (Meng et al., 2023) to reduce the dimensionality of the high-dimensional dataset, followed by DPMM clustering. We compare and analyse the clustering results on the iris dataset. Both DPMM and TSNE-DPMM achieve a precision of 0.97 in terms of clustering accuracy, whereas PCA-DPMM only reaches a maximum precision of 0.94. When considering different numbers of iterations, TSNE-DPMM maintains a stable accuracy ranging from 0.92 to 0.97, whereas DPMM fluctuates between 0.66 and 0.97, and PCA-DPMM fluctuates between 0.73 and 0.94. Thus, TSNE-DPMM ensures high accuracy in clustering results and demonstrates excellent stability. Moreover, compared to other algorithms, it provides superior visualisation effects with the clearest category boundaries. Therefore, even on large-scale datasets, we can enhance the computational efficiency of the algorithm by setting a lower number of iterations without compromising accuracy.

The advantages of using the TSNE-DPMM algorithm are as follows:

- 1 it does not require a pre-specified number of clusters
- 2 it can handle high-dimensional datasets
- 3 it provides good stability and high-quality clustering results.

In this study, this method is applied to the dimensionality reduction and clustering of the continuous features in the expense portion of the medical insurance data. By partitioning the database on various expense features, it is assumed that normal data should be grouped in large clusters, while abnormal data should be in small clusters or isolated points. However, this rough partitioning only provides a general understanding and does not yield precise anomaly scores for each patient's data. In Section 3.3, the CBLOF algorithm is introduced to accurately calculate the anomaly scores for each data point.

3.3 Local outlier factor

The LOF algorithm (Boukela et al., 2021) is a density-based anomaly detection algorithm. It is based on the concept of density and calculates the anomaly degree of each sample by comparing the density of the sample with its surrounding neighbourhood. In this study, clustering is used to obtain clusters, and the similarity between data points and clusters, as well as the size of the clusters, are used to define the local outlier values of data points and detect outliers. Firstly, CBLOF is defined, which defines an outlier factor:

Let attribute $A_1, A_2, \dots, A_m$ of the record correspond to the field $D_1, D_2, \dots, D_m$ respectively. For each record x in dataset D , $x \in D_1 \times \dots \times D_m$. The clustering result of the dataset D is denoted as $C = \{C_1, C_2, \dots, C_K\}$, and $C_i \cap C_j = \Phi$, $C_1 \cup \dots \cup C_K = D$.

The number of clusters in the clustering result is denoted as K , and C can be regarded as a partition of the dataset D . After clustering the dataset into K clusters, it is necessary to distinguish which clusters are larger and which clusters are smaller.

Assuming the clusters after clustering are arranged in descending order of the number of records within each cluster, and $|C_1| \geq |C_2| \geq \dots \geq |C_K|$.

$$(|C_1| + |C_2| + \dots + |C_b|) \geq |D| * \alpha \quad (7)$$

$$\frac{|C_b|}{|C_{b+1}|} \geq \beta \quad (8)$$

The parameters α and β are used to define large clusters and small clusters. Parameter α represents the proportion of the sum of the numbers of records in large clusters to the total dataset, and parameter β represents the desired proportion of data quantity between large clusters and small clusters. b is the boundary between large clusters and small clusters. Therefore, larger clusters are defined as $LC = \{C_i | i \leq b\}$, and smaller clusters are defined as $SC = \{C_i | i > b\}$.

The CBLOF is defined as follows:

$$CBLOF(x) = \begin{cases} \frac{1}{|C_i|} * \min(sim(x, C_j)), x \in C_i, C_i \in SC, \\ \frac{1}{|C_i|} * (sim(x, C_i)), x \in C_i, C_i \in LC \end{cases} \quad C_j \in LC, j = 1, \dots, b \quad (9)$$

CBLOF measures the anomaly degree of a data point. From equation (11), it can be observed that if a data point is in a large cluster LC , the CBLOF is determined by the size of the cluster it belongs to and the similarity to the points in the cluster. If the data point is in a small cluster SC , the CBLOF is determined by the size of the cluster it belongs to and the similarity to the nearest large cluster. The larger the CBLOF value, the more likely it is an anomaly. The definition of CBLOF in this context is different from Pan, as the data contains both discrete and continuous attributes. Therefore, different similarity calculation methods are defined for different attribute values.

For a cluster C , $x(A_i) = a_i$, the support of the value a_i of the discrete attribute A_i for a data x is calculated as:

$$\sup(a_i) = \{x | x(A_i) = a_i, x \in C\} \quad (10)$$

The similarity between a data point x and the discrete attributes in cluster C is defined as:

$$sim_d(x, C) = \sum_{i=1}^m \frac{\sup(a_i)}{m|C|} \quad (11)$$

where m is the number of discrete attributes. It can be seen that the larger the value of $sim_d(x, C)$, the greater the similarity between the data point and the cluster.

The similarity measure is primarily used to calculate the similarity between basic information and treatment process characteristics in medical insurance data. It involves determining, for each patient's data, the proportion of data within its cluster that share identical basic information and treatment process characteristics.

The similarity between a data point x and cluster C is defined as:

$$sim_c(x, C) = p(x|C) \quad (12)$$

where $p(x|C)$ represents the conditional probability of data point x following the data distribution in cluster C , and represents the similarity between the data point x and the continuous attributes in cluster C . The similarity definition corresponds to the DPMM mentioned in the article for clustering medical insurance data expenses, with the Gibbs sampling process, where each data follows the probability density of the cluster it belongs to. Joint similarity is defined based on discrete feature similarity and continuous feature similarity, as equation (13):

$$sim(x, C) = \frac{1}{2}(sim_d(x, C) + sim_c(x, C)) \quad (13)$$

3.4 LOF based on improved DPMM

In order to achieve anomaly detection in medical insurance data, this paper presents an enhanced DPMM local outlier factor mining algorithm (ILOF-DPMM). Firstly, TSNE-DPMM is applied to reduce the dimensionality and cluster the cost features in the medical insurance data. These clusters are then divided into large clusters and small clusters. From a clustering perspective, edge data points and

data points in small clusters are more likely to be outliers, while cluster centres are more likely to be normal points. For data points within the clusters, the CBLOF value is calculated using equation (11), where the similarity is computed using equation (14). The joint similarity calculation considers both the discrete features of basic information and medical process in medical insurance data, as well as the continuous feature of medical insurance costs. As a result, data points in large clusters will have lower anomaly measures, while data points in small clusters will have higher anomaly measures. In other words, for the same disease category, when there is little variation in patients' basic information, it should adhere to several common treatment processes for that disease, each of which corresponds to a specific cluster. If isolated or sparse data points form a separate cluster, they are likely to be outliers. The specific degree of anomaly can be determined by the CBLOF value, and data points exceeding the threshold range can be selected as candidate outliers in the dataset.

For each cluster C , assuming that the data points in the cluster follow a multivariate Gaussian distribution, the similarity between a data point and the cluster C , as shown in equation (14), can be written as:

$$\begin{aligned} \text{sim}_c(x, C) &= p(x|C) \\ &= \frac{1}{2\pi^{d/2} \cdot \sqrt{|\Sigma|}} \exp\left[-\frac{(x-\mu)^T \cdot \Sigma^{-1} \cdot (x-\mu)}{2}\right] \end{aligned} \quad (14)$$

where d represents the number of continuous attributes, μ and Σ respectively represent the mean and covariance of the Gaussian distribution (Zhao et al., 2023) followed by the data in cluster.

The ILOF-DPMM algorithm is divided into two parts:

- 1 clustering the dataset
- 2 calculating the LOF value for each data point.

The complexity of the first part of the algorithm is determined by the TSNE-DPMM algorithm, with a time complexity of $O(n \times n + n \times K)$ and a space complexity of $O(n \times n \times d + n \times T \times K)$, where n is the number of data points, K is the number of clusters, T is the number of iterations, and d is the dimensionality of the data. The second part calculates the LOF value of the data and only requires a single scan, with time and space complexity both being $O(n \times K)$. Typically, d and K are smaller than n , so the overall time complexity of the algorithm can be represented as $O(n \times n \times d + n \times T \times K)$, and the space complexity as $O(n \times n)$. The termination condition for the iterations is determined based on the change in the corresponding data point's outlier score between consecutive iterations, which should not exceed a specific threshold. The iterations stop when this condition is met.

The specific algorithm is shown in Table 1.

Table 1 ILOF-DPMM algorithm

<i>ILOF-DPMM algorithm</i>	
Input: Dataset $X = \{x_1, x_2, \dots, x_n\}$, parameters α, β	
Output: CBLOF value for all data	
1	Begin
2	Clustering the data X using TSNE-DPMM algorithm
3	/*Producted clusters: $C = \{C_1, C_2, \dots, C_k\}$ and $ C_1 C_2 \geq \dots \geq C_k $ */
4	get LC and SC with the 2 parameters // In order to get the largest and smallest clusters
5	for each x in the dataset do begin
6	if x belongs to C_i and C_i belongs to SC do begin
7	$CBLOF(x) = \frac{1}{ C_i } * \min(sim(x, C_j))$ // C_j belongs to LC
8	else
9	$CBLOF(x) = \frac{1}{ C_i } * sim(x, C_i)$ // C_i belongs to
10	return $CBLOF$
11	end
12	end

4 Performance results

Since the medical insurance data is unlabelled, it is not possible to directly evaluate the performance of the model. To assess the model's performance, we can initially use UCI public datasets to determine the range of the model's anomaly detection rate.

4.1 Experiment analysis of UCI datasets

In order to assess the performance of our research algorithm, we will compare it with the current mainstream anomaly detection schemes. LOF, a commonly used anomaly detection algorithm, determines the degree of anomaly of each data point by calculating the density of its neighbouring points. K-means-LOF is an anomaly detection method that combines k-means clustering and CBLOF algorithm. Deep clustering (Rao and Shi, 2023), on the other hand, is a method that utilises deep learning algorithms for clustering, and has also been applied to anomaly point detection in anomaly detection tasks. The evaluation metrics include accuracy, precision, recall, F1 score, and ROC curve. Since we need to perform anomaly detection on medical insurance data, which is unlabelled, we cannot directly evaluate the performance of the model. Therefore, we will first evaluate the model using a public dataset to determine the threshold for the detection rate of anomalous data.

4.1.1 Credit card fraud detection dataset

Comprised of 284,807 anonymous credit card transactions, this dataset contains 28 principal component features, in addition to the original features of ‘Time’ and ‘Amount’. KDD-cup99 Dataset: With over 10,000 database records, this dataset encompasses four types of attacks: denial of service (DoS), probe, remote to local (R2L), and user to root (U2R). It consists of 41 attributes and 1 target variable. Among the 41 features, 9 are discrete while the remainder are continuous. For our experimental analysis, we will solely utilise a subset of the dataset, as illustrated in Table 2.

The presented model achieved the highest scores in terms of accuracy, precision, and F1 score on the credit card fraud detection dataset, according to the table. However, K-means-LOF achieved a recall score of 1.0, followed closely by the proposed model at 0.996. The AUC values for LOF, K-means-LOF, deep clustering, and ILOF-DPMM were 0.985, 0.973, 0.982, and 0.988, respectively. Regarding the KDD-cup99 dataset, the proposed model achieved the highest accuracy of 0.99. Deep clustering obtained the highest score in terms of accuracy, while the proposed model scored slightly lower. Additionally, K-means-LOF and ILOF-DPMM achieved the highest recall score of 1.0. All four algorithms had F1 scores above 0.98, with AUC values of 0.9932, 0.9886, 0.9884, and 0.996. The experimental results demonstrate the exceptional performance of the algorithm in anomaly detection.

To determine the detection threshold for healthcare data, an analysis was conducted on the detection rates of different testing ranges using the ILOF-DPMM algorithm and other anomaly detection algorithms. The detection rate represents the proportion of detected anomalies within a certain testing range to the total number of anomalies. Tables 4 and 5 present the experimental results on two datasets. The data is sorted in descending order based on anomaly scores, as higher scores indicate higher degree of anomaly. The first column from the left in each table represents the ‘selected data volume,’ and the percentages indicate the top 5%, 15%, 20%, 25%, and 30% of anomaly scores. The subsequent columns display the number of detected anomalies and their proportions among the total number of anomalies.

Table 2 Experimental dataset

Dataset	Number of instances	Positive instances	Outlier of instances	Number of features
Credit Card	1,000	900	100	30
KDD-cup99	1,000	900	100	41

The ultimate goal of designing models in experiments is to provide higher accuracy for anomaly detection in high-dimensional unlabelled data. According to the experimental results, Table 3 lists the anomaly detection results of four different models.

Table 3 Evaluation metrics for the three models

Comparison model	Evaluation metrics	Dataset	
		credit card	KDD-cup99
LOF	Accuracy	0.967	0.970
	Precision	0.965	0.967
	Recall	0.976	0.925
	F1 score	0.960	0.981
K-means-LOF	Accuracy	0.951	0.968
	Precision	0.957	0.954
	Recall	1.0	1.0
	F1 score	0.971	0.982
Deep clustering	Accuracy	0.960	0.962
	Precision	0.979	1.0
	Recall	0.978	0.957
	F1 score	0.982	0.978
ILOF-DPMM	Accuracy	0.982	0.990
	Precision	0.987	0.975
	Recall	0.996	1.0
	F1 score	0.989	0.991

Table 4 The results of three anomaly detection algorithms in the credit card dataset

Data volume	Anomaly detection rate (%)			
	LOF	K-means-LOF	Deep clustering	ILOF-DPMM
5% (50)	44	50	49	50
10% (100)	82	88	82	98
15% (150)	96	96	93	100
20% (200)	100	99	96	100
25% (250)	100	100	100	100

Table 5 The results of three anomaly detection algorithms in the KDD-cup9

Data volume	Anomaly detection rate (%)			
	LOF	K-means-LOF	Deep clustering	ILOF-DPMM
5% (50)	46	49	50	50
10% (100)	82	87	81	99
15% (150)	100	96	94	100
20% (200)	100	99	96	100
25% (250)	100	100	98	100
30% (300)	100	100	100	100

According to Table 4, the LOF algorithm can detect 44% anomalies when selecting the top 5% of databased on LOF values, while K-means-LOF and ILOF-DPMM can only detect anomalies within the top 5%. However, when the top 15% is selected, the ILOF-DPMM algorithm is capable of detecting all anomalies. Upon further expanding the detection range, it is observed that LOF, K-means-LOF, and

deep clustering can detect all anomalies within the top 25% of data, whereas our proposed method achieves a 100% detection rate within the top 15%.

Table 5 shows that when the top 10% of data are selected based on LOF values, the ILOF-DPMM algorithm detects 99% of anomalies, while the LOF, K-means-LOF and Deep Clustering algorithms detect 82%, 87% and 81% respectively. % exception. However, when the top 15% of data are selected based on anomaly scores, the LOF and ILOF-DPMM algorithms can detect all anomalies, while the K-means-LOF and deep clustering algorithms reach 100% at 25% and 30%, respectively.

The analysis of Table 4 and Table 5 reveals that all four algorithms achieved a 100% detection rate for the top 30% outlier scores. Thus, in the context of healthcare insurance data detection, it can be inferred that the top 30% of the data ranked by the four algorithms are considered as suspected abnormal cases. Furthermore, the experimental results of the proposed ILOF-DPMM algorithm on these two datasets demonstrate that our algorithm is capable of identifying all the abnormal data points within the top 20% with relatively fewer iterations. Therefore, in the detection of anomalies in healthcare insurance data, we adopt the approach of considering data points with outlier scores within the top 20% at the end of iterations, and with abnormal values appearing in over half of the iterations, as suspected anomalies.

Figures 1 and 2 depict the scatter plots of outlier distributions using four different methods on two datasets. A value closer to 1 indicates a higher level of outlier. It can

be observed that both the K-means-LOF method and the Deep clustering method do not exhibit clear boundaries for the division of outlier data points. The LOF method follows closely. The ILOF-DPMM method yields the best results, demonstrating a certain level of clustering effect. The number of clusters can be identified and clearly demarcated. Furthermore, calculations show that more than 99% of the outlier data are outside the range.

4.2 Medical insurance data

The study utilised comprehensive patient medical records from a hospital spanning from October 2021 to June 2022. A total of 30,920 data records were collected, rigorously adhering to the quality control standards set by the National Hospital Case Management Quality Control Centre. This dataset includes nine features of individual basic information for insured individuals (including hospitalisation frequency, medical payment methods, age, gender, occupation, nationality, marital status, race, and identity type), as well as 572 features related to diagnosis and treatment information. The information encompasses treatment processes (admission details, hospitalisation details, diagnostic information, and discharge diagnosis), as well as 30 features related to the costs incurred during the diagnostic and treatment processes. In total, there are 611 features. This analysis focuses on clustering analysis of the key cost-related features. To facilitate effective anomaly detection, statistical analysis and data pre-processing were performed on the dataset.

Figure 1 Scatter plot of four methods in the credit card fraud detection (see online version for colours)

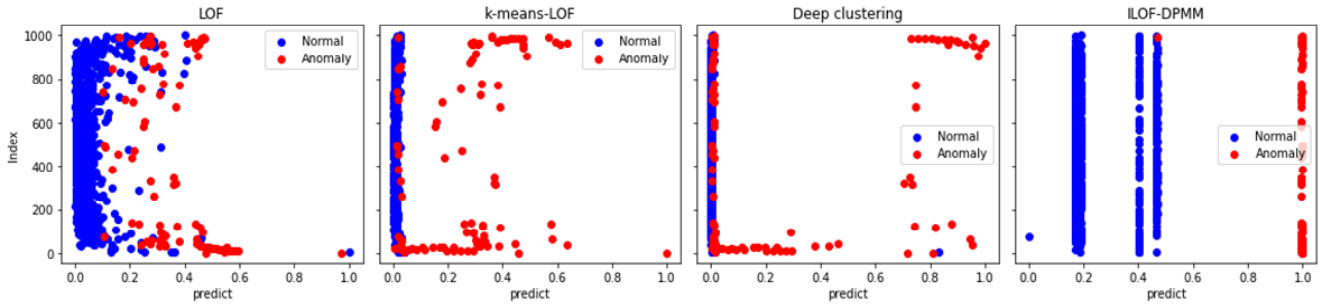
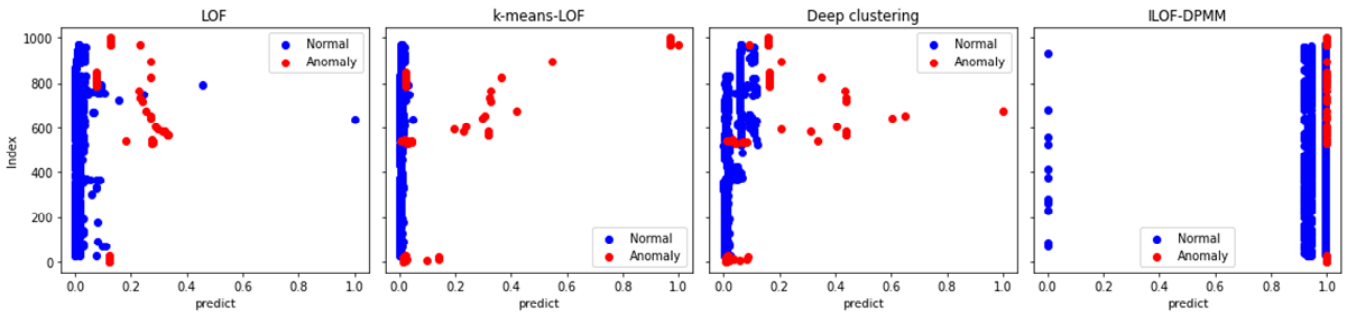


Figure 2 Scatter plot of four methods in the KDD-cup99 (see online version for colours)



In order to analyse anomalies in medical insurance, we have chosen to utilise medical insurance claim data. This subset consists of 22,886 data records that adhere to the quality control standards, accounting for 74% of the total data. Since different diseases require different diagnoses and treatments, the resulting costs can vary. Thus, we have classified diseases into two categories based on the initial diagnosis at discharge: bronchopneumonia and multiple cerebral infarctions. To address missing values, we have utilised similar diagnostic and treatment procedure data combined with nine fundamental information features. Out of the 572 features during the inpatient treatment process, 24 features were extracted by retaining numeric values and eliminating all missing attributes. As a result, a total of 33 discrete features were extracted based on patients' basic information and the inpatient process. Regarding cost features, the average value for the same disease is used to fill in a small number of missing values. Table 6 specifically lists the volume of data on whether these two diseases required surgery.

Table 6 Statistics by disease type

Cell	Bronchopneumonia	Multiple cerebral infarction	Total
No surgery	11,705	6,770	18,475
Surgery	2,736	1,675	4,411
Total	14,441	8,445	22,886

This study focused on analysing Medicare data for patients who did not require surgery. Each disease is regarded as a unit, and the random forest method is used to extract the discrete characteristics of each disease type, which are the main influencing factors on the total hospitalisation costs.

Figure 3 Factors influencing total hospitalisation expenditure for bronchopneumonia (see online version for colours)

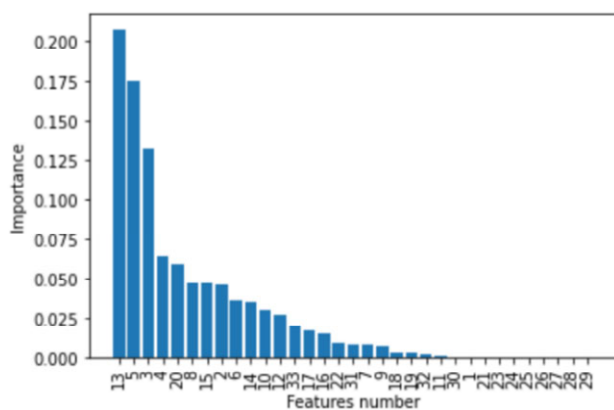


Figure 3 presents that the primary factors influencing the total cost of hospitalisation for bronchopneumonia include the top 15 features, encompassing hospitalisation frequency, healthcare payment methods, gender, age, ethnicity, occupation, identification document type, admission path, actual length of stay, admission status, initial discharge time, allergy history, ABO blood type, secondary nursing

days, and discharge mode. Age and actual length of stay are regarded as continuous attributes and will be subject to subsequent clustering analysis along with various medical insurance cost features. The remaining 13 features will be treated as discrete attributes to determine similarity and detect outliers.

Figure 4 Factors influencing total hospitalisation expenditure for multiple cerebral infarction (see online version for colours)

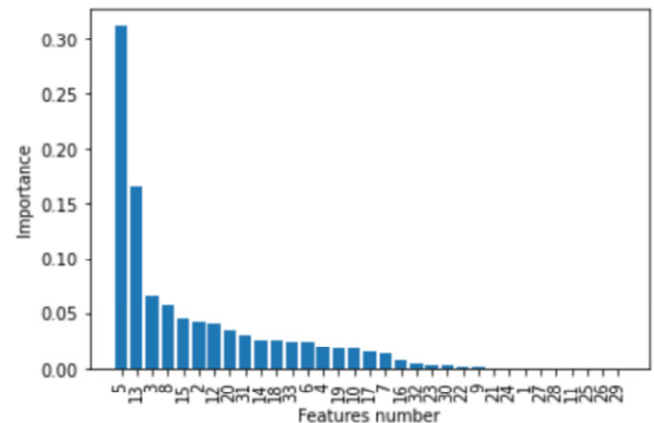


Figure 4 depicts that there are a total of 13 main factors influencing the overall hospitalisation costs of multiple cerebral infarctions. Similarly, age and actual length of stay will be regarded as continuous attributes. The remaining 11 features will be considered as discrete attributes, utilised for similarity calculation and subsequent anomaly detection. Based on the characteristics of 30 medical insurance expenses, using the classification principles of medical guidance expenditure, they are classified into ten major financial categories.

Tables 7 and 8 present a statistical analysis of the scoring methodology for specific disease healthcare data abnormalities. They delineate the quantities and proportions of outlier values, ranked within the top 5%, 10%, 15%, 20%, 25%, and 30%, as well as occurrences of 3 and 4 instances, with a fourfold increase in proportions.

The statistical results of Tables 7 and 8 indicate that as the detection range increases from 5% to 30% in the four methods, the number of abnormal data points continuously increases, suggesting a consistent distribution of outliers.

Table 7 Suspected abnormal data for bronchopneumonia

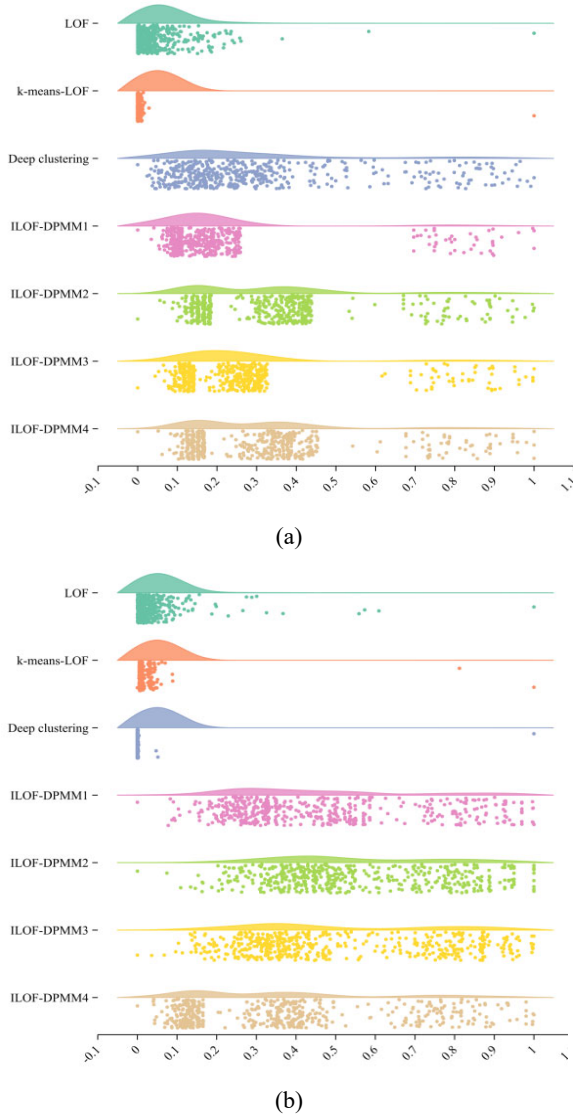
Data volume	Number of occurrences in 4 methods		
	4	3	Total
585 (5%)	141	112	253 (43.25%)
1,170 (10%)	383	307	690 (58.97%)
1,755 (15%)	690	418	1,108 (63.13%)
2,341 (20%)	1,009	510	1,519 (64.89%)
2,926 (25%)	1,409	510	1,919 (65.58%)
3,511 (30%)	1,827	556	2,383 (67.87%)

Figure 5 displays the rain-cloud plots of 500 randomly selected data points for each disease, where values closer to 0 represent normal observations, while values closer to 1 indicate abnormality. From the graph, it is evident that the suspected outliers are the same for all four methods. Furthermore, the results of different iterations of ILOF-DPMM demonstrate the stability of the method.

Table 8 Suspected abnormal data for multiple cerebral infarction

Data volume	Number of occurrences in 4 methods		
	4	3	Total
339 (5%)	165	95	260 (76.92%)
677 (10%)	291	181	472 (69.72%)
1,016 (15%)	517	306	823 (81.08%)
1,354 (20%)	822	392	1,214 (89.66%)
1,692 (25%)	1067	452	1,519 (89.76%)
2,031 (30%)	1301	514	1,815 (89.36%)

Figure 5 Rain-cloud plot of predicted value distribution (a) bronchopneumonia (b) multiple cerebral infarction (see online version for colours)



5 Conclusions

The primary aim of this study is to identify anomalous data in medical insurance billing, thereby reducing overdiagnosis and overcharging, and proposing the ILOF-DPMM algorithm. Patient data obtained from medical records, comprising both discrete and continuous features, is employed, and similarity is calculated using joint distance. On the credit card fraud detection dataset, the ILOF-DPMM algorithm exhibits accuracy rates that are 0.031, 0.067, and 0.018 higher than LOF, K-means-LOF, and deep clustering, respectively. Similarly, on the KDD-cup99 dataset, it outperforms LOF, K-means-LOF, and deep clustering by 0.065, 0.01, and 0.034, respectively, with the best detection rate. Additionally, ILOF-DPMM demonstrates a superior anomaly detection performance. For the real health insurance dataset, the outliers are determined based on the distribution of unlabelled data. Both disease types rank in the top 30% of data for all four methods, accounting for 1,827 and 1,301 cases, or 0.156 and 0.195 of the total data, respectively. In the multiple iterations of ILOF-DPMM, the top 20% of data comprises 941 and 590 data points, or 0.080 and 0.088, indicating a high level of anomalies. These results indicate that the ILOF-DPMM algorithm is capable of effectively processing medical data with discrete and continuous features, achieving not only high detection rates but also more pronounced partitioning effects and model stability. In future research, to enhance the timeliness and accuracy of detection, the algorithm can be further optimised to reduce complexity and integrated with relevant diagnosis and treatment processes.

References

- Boukela, L., Zhang, G., Yacoub, M., Bouzeffrane, S., Ahmadi, S. B.B. and Jelodar, H. (2021) 'A modified LOF-based approach for outlier characterization in IoT', *Annals of Telecommunications*, Vol. 76, pp.145-153, DOI: 10.1007/s12243-020-00780-5.
- Ding, D. and Karabatsos, G. (2021) 'Dirichlet process mixture models with shrinkage prior', *Stat.*, Vol. 10, No. 1, DOI: 10.1002/sta4.371.
- Elliott, M.R. and Xia, X. (2021) 'Weighted Dirichlet process mixture models to accommodate complex sample designs for linear and quantile regression', *Journal of Official Statistics*, Vol. 37, No. 1, pp.71-95, DOI: 10.2478/jos-2021-0004.
- Guo, L., Ge, P., Zhao, Y., Wang, D. and Huang, L. (2023) 'LightMOT: a lightweight convolution neural network for real-time multi-object tracking', *International Journal of Bio-Inspired Computation*, Vol. 22, No. 3, pp.152-161, DOI: 10.1504/IJBIC.2023.135467.
- Huang, Z., Li, X., Du, H., Zou, W., Zhou, G., Mao, F., ... and Hu, M. (2023) 'An algorithm of forest age estimation based on the forest disturbance and recovery detection', *IEEE Transactions on Geoscience and Remote Sensing*, Vol. 61, DOI: 10.1109/TGRS.2023.3322163.
- Kalidindi, A. and Arrama, M.B. (2023) 'Botnet attack detection in IoT using hybrid optimisation enabled deep stacked autoencoder network', *International Journal of Bio-Inspired Computation*, Vol. 22, No. 2, pp.77-88, DOI: 10.1504/IJBIC.2023.134981.

- Kemp, J., Barker, C., Good, N. and Bain, M. (2022) 'Sequential pattern detection for identifying courses of treatment and anomalous claim behaviour in medical insurance', in *2022 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, December, pp.3039–3046, DOI: 10.1109/BIBM55620.2022.9995541.
- Kou, Y., Lin, G., Qian, Y. and Liao, S. (2023) 'A novel multi-label feature selection method with association rules and rough set', *Information Sciences*, Vol. 624, pp.299–323, DOI: 10.1016/j.ins.2022.12.070.
- Li, W., Ye, P., Yu, K., Min, X. and Xie, W. (2023) 'An abnormal surgical record recognition model with keywords combination patterns based on TextRank for medical insurance fraud detection', *Multimedia Tools and Applications*, Vol. 82, pp.30949–30963, DOI: 10.1007/S11042-023-14529-4.
- Liu, J., Wang, Y. and Yu, J. (2023) 'A study on the path of governance in health insurance fraud considering moral hazard', *Frontiers in Public Health*, Vol. 11, DOI: 10.3389%2Ffpubh.2023.1199912.
- Meng, L., van den Elzen, S., Pezzotti, N. and Vilanova, A. (2023) 'Class-constrained t-SNE: combining data features and class probabilities', *IEEE Transactions on Visualization and Computer Graphics*, DOI: 10.1109/TVCG.2023.3326600.
- Pan, S. and Zhang, W. (2017) 'Fraudulent medical behavior detection based on hybrid approach', *Journal of East China Normal University (Natural Science)*, Vol. 2017, No. 5, pp.125–137, DOI: 10.3969/j.issn.1000-5641.2017.05.012.
- Premkumar, M., Ashokkumar, S.R., Jeevanantham, V., Mohanbabu, G. and Anupallavi, S. (2023) 'Scalable and energy efficient cluster based anomaly detection against denial of service attacks in wireless sensor networks', *Wireless Personal Communications*, Vol. 129, No. 4, pp.2669–2691, DOI: 10.1007/s11277-023-10252-3.
- Rao, D. and Shi, H. (2023) 'Study on air traffic flow recognition and anomaly detection based on deep clustering', *Computer Science*, Vol. 50, No. 3, pp.121–128.
- Shu, Z., Song, A., Wu, G. and Pedrycz, W. (2023) 'Variable reduction strategy integrated variable neighborhood search and NSGA-II hybrid algorithm for emergency material scheduling', *Complex System Modeling and Simulation*, Vol. 3, No. 2, pp. 83–101, DOI: 10.23919/CSMS.2023.0006.
- Tang J. (2018) 'Combating medicare fraud and safeguarding funds', *Chinese Medical Insurance*, Vol. 5, No. 5, pp.5–7.
- Tayeh, T., Aburakhia, S., Myers, R. and Shami, A. (2020) 'Distance-based anomaly detection for industrial surfaces using triplet networks', in *2020 11th IEEE Annual Information Technology, Electronics and Mobile Communication Conference (IEMCON)*, November, pp.372–377, DOI: 10.1109/IEMCON51383.2020.9284921.
- Vakulenko, M.O. (2023) 'Semantic comparison of texts by the metric approach', *Digital Scholarship in the Humanities*, Vol. 38, No. 2, pp.766–771, DOI: 10.1093/lc/fqac059.
- Villegas-Ortega, J., Bellido-Boza, L. and Mauricio, D. (2021) 'Fourteen years of manifestations and factors of health insurance fraud, 2006–2020: a scoping review', *Health and Justice*, Vol. 9, pp.1–23, DOI: 10.1186/s40352-021-00149-3.
- Wang, K., An, J., Yu, Z., Yin, X. and Ma, C. (2021) 'Kernel local outlier factor-based fuzzy support vector machine for imbalanced classification', *Concurrency and Computation: Practice and Experience*, Vol. 33, No. 13, DOI: 10.1002/cpe.6235.
- Wu, H., Ye, F., Gao, Y., Cong, Y. and Hao, A. (2022) 'Real-time laparoscopic cholecystectomy simulation using a particle-based physical system', *Complex System Modeling and Simulation*, Vol. 2, No. 2, pp.186–196, DOI: 10.23919/CSMS.2022.0009.
- Xiao, Q., Wu, L., Wu, X. and Rättsch, M. (2023) 'Simulating temporally and spatially correlated wind speed time series by spectral representation method', *Complex System Modeling and Simulation*, Vol. 3, No. 2, pp.157–168, DOI: 10.23919/CSMS.2023.0007.
- Xiao, S., Bai, T., Cui, X., Wu, B., Meng, X. and Wang, B. (2023) 'A graph-based contrastive learning framework for medicare insurance fraud detection', *Frontiers of Computer Science*, Vol. 17, No. 2, DOI: 10.1007/S11704-022-1734-0.
- Xin, J., Qu, Y., Zhang, F. and Negenborn, R. (2022) 'Distributed model predictive contouring control for real-time multi-robot motion planning', *Complex System Modeling and Simulation*, Vol. 2, No. 4, pp.273–287, DOI: 10.23919/CSMS.2022.0017.
- Xu, Z. (2022) 'Medical insurance anomaly detection based on deep semi-nonnegative matrix factorization', *Technology Innovation and Application*, No. 25, pp.90–93, DOI: 10.19981/j.CN23-1581/G3.2022.25.021.
- Yabroff, K.R., Zhao, J., Han, X. and Zheng, Z. (2019) 'Prevalence and correlates of medical financial hardship in the USA', *Journal of General Internal Medicine*, Vol. 34, pp.1494–1502, DOI: 10.1007/s11606-019-05002-w.
- Yu, Z., Pei, C., Zhang, S., Wen, X., Li, J., Xie, G. and Pei, D. (2023) 'AutoKAD: empowering KPI anomaly detection with label-free deployment', in *2023 IEEE 34th International Symposium on Software Reliability Engineering (ISSRE)*, pp.13–23, DOI: 10.1109/ISSRE59848.2023.00063.
- Zhao, S., Wang, Q., Zheng, J. and Wang, X. (2023) 'Gaussian process based modeling and control of affine systems with control saturation constraints', *Complex System Modeling and Simulation*, Vol. 3, No. 3, pp.252–260, DOI: 10.23919/CSMS.2023.0009.