Journal of Information Hiding and Multimedia Signal Processing ©2024 ISSN 2073-4212
Ubiquitous International Volume **15**, Number **1**, March **2024**

XGAN: A Medical Insurance fraud Detector based on GAN with XGBoost

Yi Mao, Yun Li, Bei Xu, Jingyu Han,

School of Computer Science and Technology
Nanjing University of Posts and Telecommunications (NJUPT)
No.9, Wenyuan Road,Yadong new District, Nanjing, 210023, China
maoyi@nuk.edu.tw;liyun@njupt.edu.cn; xubei@njupt.edu.cn; jingyuhan@njupt.edu.cn

Received December 29, 2023; revised February 17, 2024; accepted March 5, 2024

Abstract. *Efficient and accurate identification of medical insurance fraud is of great significance to ensure the standardized use of medical insurance funds and the improvement and development of national medical system. Apart from the economic loss caused by such fraud, real patients in need of medical care suffer because they cannot access services that are being lost due to a lack of funds.This paper mainly studies the establishment of medical insurance fraud model based on machine learning. Aiming at the unbalanced distribution of fraudulent samples and normal samples of insured personnel, the traditional processing methods of over-sampling and under-sampling and the new samples generated by GAN (Generative Adverse Networks) neural network are used to solve the problem of sample imbalance, which effectively reduces the impact of the imbalance of original data on the effect of binary classification detection. After processing the data samples of insured persons, the establishment of medical insurance fraud monitoring model is completed by using logistic regression model and XGBoost model respectively, and finally the evaluation of the monitoring model is realized on the original data set. From the final prediction results of the model, over-sampling and GAN processing methods can better retain the sample characteristics than under sampling, while XGBoost model performs better in classification and detection problems than logistic regression model.*

Keywords: Medical Insurance; Fraud detection; Characteristic Engineering; Deep Learning; Classification detection.

1. Introduction.

With rapid economic development, health is not only an important part of the people's pursuit of a better life but also sets higher standards for national governance. However, the fraud of Medicaid has become a major obstacle to the healthy development of medical management mechanism. Medical insurance fraud swindled a lot of money, which not only affected the rational application of medical insurance funds, but also greatly weakened the credibility of the government's medical insurance system. In 2021, the National Medical Insurance Bureau released the 2020 National Medical Security Development Statistical Communique, which pointed out that in 2020, 1.36 billion people in China have basically participated in medical insurance, with a participation rate of up to 95%, and the trend of universal medical insurance has taken shape [1]. The U.S. federal government estimates that about 700 billion is lost due to fraud, waste and abuse in the U.S. healthcare system [7]. Even though the state has compiled statistics on cases of fraud involving basic medical insurance, the data presented is merely the tip of the iceberg of the actual figures, indicating that the real amount of medical insurance fraud far exceeds the official statistics. The defrauding of medical insurance funds is characterized by a diversity of actors, complex methods, and covert actions, which makes the fraud detection process more and challenging.

The prevalence of health insurance fraud has led to the emergence of anti-fraud research. According to the investigation, the cases investigated and handled by the supervision of China's medical insurance fund include most of the personnel involved in the medical insurance process, including the staff of designated medical institutions, the staff of designated retail pharmacies, the insured personnel and the relevant personnel of medical security institutions. The fraud and fraud of the medical insurance fund not only caused losses to the fund itself, but also damaged the efficiency and effectiveness of the use of the fund, and seriously damaged the trust of patients in medical insurance institutions and the government by patients who urgently needed "life-saving money". In order to make rational use of medical insurance funds, it is necessary to strengthen the prevention of medical insurance fraud risks.

The "China Health Insurance Development Research Report" points out that insurance fraud globally accounts for 15% of the total insurance payouts, with an estimated fraud rate between 20%-30%. The "2019 China Insurance Industry Intelligent Risk Control White Paper" reveals that, according to calculations by the International Association of Insurance Supervisors (IAIS), global losses due to insurance fraud amount to $80 billion annually. Insurance fraud primarily includes intentionally fabricating insurance subjects to defraud insurance money [17]; concocting non-existent insurance incidents, inventing false causes of accidents, or exaggerating the extent of losses to claim insurance money; and deliberately causing insurance incidents to claim insurance money [9].

At this stage, the scale of insurance fraud is growing day by day, and the methods of crime are becoming more and more digital and hidden. However, nowadays, most of the medical insurance professionals in the medical insurance department randomly select the files accumulated for a long time [27], but this manual sampling method is time-consuming, laborious, and inefficient, and it is easy to make judgments based on experience alone. Most importantly, this approach does not accurately identify anomalous relevant data in the health insurance information system. Deep learning and machine learning technology can effectively reveal the inherent relationship between medical insurance fraud and enrollment records, improve the accuracy of identification, and greatly reduce inefficient labor expenditure. Based on machine learning and deep learning, this paper builds a medical insurance fraud detection model and contributes to the digitalization and modernization of medical insurance detection.

This paper is organized as follows: Section 2 summarizes the research background and significance of this topic, the research status at home and abroad, and then introduces the main research content of this paper, finally, expounds the overall structure of the paper. Section 3 and Section 4introduce the process of model building and the evaluation of results. Data is generated by building GAN network to further deal with the problem of data imbalance. The model evaluation mechanism of confusion matrix is introduced, and a variety of model evaluation indexes are introduced. In model training, combined with the data obtained after data preprocessing, the model is modeled by logistic regression and XGBoost models, and the model is trained. The principle of classification model is briefly described, and the relevant skills in model training are introduced. Finally, the advantages of GAN-XGBoost model in fraud detection are highlighted. In Section 5summarizes the research process and the relevant results, puts forward the shortcomings in the research, and prospects the development of the research.

2. **Related Work.** Medical insurance fraud goes hand in hand with the medical insurance industry, and medical insurance is an important part of social insurance, so it is

closely related to social stability, so scholars at home and abroad have paid great attention to the detection of medical insurance anomalies. Research on medical insurance fraud mainly focuses on fraud risk analysis and fraud identification. This section mainly describes different research studies revolving in this area. In addition, we will pay special attention to the research related to medical insurance fraud detection in the context of class imbalance issues.

In the field of fraud risk analysis, in 2003, Sparrow proposed that there are 7 levels of health insurance fraud. The FBI has categorized some common methods of health insurance fraud: virtual services, duplicate billing, over-reporting, over-checking, kickbacks, and more. Most of the people involved in the health insurance process are likely to be involved in fraud. Mitka [18] and Home [2] et al. argued in 2011 that health-care providers were most likely to be involved in fraud among the various actors involved. Therefore, it is particularly necessary to study the related work in this field, whose main methods could be classification-related, such as Deep learning, ensemble and feature ranking, and etc al.

2.1. **Data Mining Approaches.**

Data mining can discover deep and valuable information from massive data, find hidden patterns, associations, trends, and anomalies from unknown structures or structured datasets, uncover relevant knowledge embedded within the data, thereby revealing the truth of matters and aiding in decision-making [19,23,25]. In 1999, Biafore proposed to use data mining to find connections between large amounts of insurance data to support the identification of fraud [24]. In 2001, Sokol et a.l built a fraud identification model that can identify different characteristics of medical services through data mining technology, and achieved good results.

The growing popularity of data mining also benefits from the widespread use of electronic billing and classification systems. CMS has utilized data-mining techniques to establish a digital, evidence-based reimbursement system for inpatient treatments. [6]. The Texas Medicaid Fraud and Abuse Detection System accumulated Millions of electronic medical records, test data, and monitoring have been collected to identify aberrant behaviors in medical treatments, thereby detecting potential fraud. Supervised methods like Neural Networks [11,21], Association Rules [20,29,30] , Genetic Algorithms [15], Fuzzy Bayesian classifier [4], Logical Regression methods [11,12], Bayesian Networks [26,28] KNN classifiers [15], and Classification Trees [11] have been used by researchers to identify fraudulent activities within the health-care system. Whether it's data analytics to track costs and payments made, upgraded health information technologies, or benchmarks for evidence and progress, data mining is the driving force behind value-based health-care transformation.

However, an important challenge in using machine learning algorithms to predict health insurance fraud is data imbalance. In many machine learning methods, data imbalance can lead to bias towards the majority class, resulting in a loss of fairness in predictions.

2.2. **Deep Learning Approaches.**

Deep learning was initially proposed by Geoff Hinton in 2006 [10]. It solves the problem of local optimal solution by replacing sigmoid response function with ReLU, maxout and other functions to solve the gradient disappearance problem. The Australian government uses BP neural networks to manage health insurance and identify health fraud [8]. In China, medical insurance fraud detection research started late, but Chinese scholars have given it a very high degree of attention. The main forms of medical insurance fraud in China include: impersonation, cause fraud, exaggerated losses, bill fraud, false medical treatment, etc. [22,32]. Yi and other scholars proposed the conception of using graph convolutional neural networks to solve the shortage of fraud samples [31]. Guo applied the decision tree algorithm to hospitalization information, dividing the characteristics of diseases to monitor the medical insurance fund [13]. In 2017, Chen proposed the parallelization algorithm of DBSCAN and random forest in medical insurance fraud detection based on Hadoop platform , which was implemented using Map-Reduce technology [14].

Due to the imbalance of the medical insurance fraud training data-set, in order to improve the accuracy of the algorithm on such data-sets, Hancock et al. [16] used Catboost to detect fraud in the unbalanced medical insurance claims data. Experiments show that the Catboost algorithm has better classification performance on such data-sets than previous algorithms. Generative Adversarial Networks (GANs), are currently one of the most important research hot-spots in the field of artificial intelligence. Their outstanding generative capabilities are not only suitable for generating various types of images and natural language data but also inspire and promote the development of various semi-supervised and unsupervised learning tasks.

Generative Adversarial Networks (GANs) is considered the most important advancement in the area of deep learning, enabling algorithms to generate new data. The goal of GAN is to synthesize new data with the same distribution as the training dataset. Therefore, the original form of GAN is considered to fall under the category of unsupervised learning for machine learning tasks because there is no need to label the data. As scholars study machine learning more deeply, it is found that extensions to primitive GANs can be located in semi-supervised and supervised tasks. The original authors and numerous other researchers have proposed numerous improvements measures and applied them in practice across multiple branches of engineering and science [3]. In the field of computer vision, GANs are applied to various tasks, including achieving image to image conversion (the technique of converting input images into specific output images), improving image resolution (generating high-definition versions from low definition images), repairing images (learning methods to repair lost or damaged parts of images), and widely used in other scenarios). For example, recent advances in GAN research have led to the ability to generate new models of high-resolution facial images. How to use the generative capabilities in GAN to improve the accuracy in medical insurance fraud detection to address the imbalance in the training data-set is still need to be further discussed.

3. **Methond.** The Proposition of GANs is to define two models, one generative and one discriminant. The flow diagram of the GAN network is shown in the figure, and the generative model is used to learn the incoming picture data to generate a picture that looks like a real sample, that is, to create a pseudo-image data. Then the generated pseudo-picture is transmitted to the discriminant model, and the function of the discriminant model is to determine whether the given picture comes from the real fake or the picture generated by the generation model, that is, to judge the authenticity of the given picture. Both models were initially not trained, the discriminant model and the generative model were trained together for adversarial training, the generative model generated a fake picture to the discriminant model for judgment, and this process was repeated, and the two models slowly enhanced their training ability to reach steady state. The ideal state is that the discriminant model determines that the probability of an image coming from a real data set and from the generative model is 50%, that is, the discriminant model cannot distinguish the authenticity of the picture, and the GAN model will achieve the ideal equilibrium state.

As shown in Figure 1, during the training process of the GAN network model, the generated network G generates as many real images as possible to fool the discriminant network D, and the goal of the discriminant network D is to differentiate the fake image

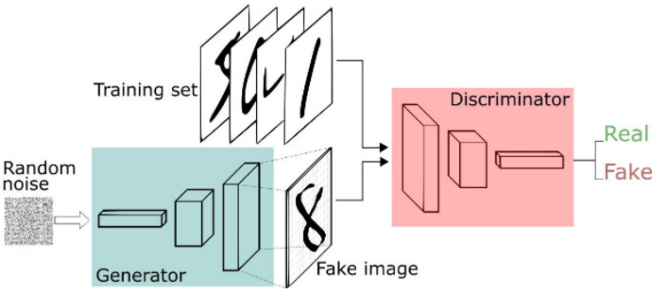


Figure 1. GAN network flowchart.

generated by G from the real image. In like manner, the interaction between the generative network G and the discriminative network D can be seen as a dynamic "game process", in which both parties continuously adjust their strategies to obtain maximum benefits. The final equilibrium point reached during this process is called the Nash equilibrium point.

3.1. **Construction of GAN model loss function.** GAN is different from other models in that GAN needs to run two optimization algorithms at the same time during training, so it is necessary to define an optimizer for the generative network and the discriminant network, one to minimize the loss of the discriminant network, and the other to minimize the loss of the generated network. If the samples come from a real data set, as shown in the formula, the samples are all true, so it is hoped that the discriminant network can also judge the samples to be true.

$$\begin{aligned} D(x) = 1, x \to Pdata(x) \\ E_{x \to Pdata(x)} log[D(X)] \end{aligned} \tag{1}$$

Similarly, as described in the following formula, if the sample comes from a generative network forgery, it is hoped that the discriminating network can identify it as fake.

$$\begin{aligned} D(x) = 0, x \to P_g(x) \\ E_{z \to P_z(z)} log[D(G(z))] \Longrightarrow min \\ or E_{z \to P_z(z)} log[1 - D(G(z))] \Longrightarrow max \end{aligned} \tag{2}$$

where x is the real sample data, which is its probability distribution. z refers to noise, such as Gaussian noise, and $p(z)$ refers to the probability density of noise. When x satisfies the probability distribution of the true sample, $x \sim p_d ata(x)$,the discriminant network outputs 1. The discriminant network wants itself to output as many correct situations as possible, that is, the mean is as large as possible. $G(z)$ refers to the sample generated from the noise z by generating the network, which is discriminated by the discriminant network$D$. $D(G(z))$ is the log-likelihood of it and is generally desired to be as small as possible, i.e. the logarithmic likelihood equivalent to $1 - D(G(z))$ is also should be as large as possible. Concurrently, because the data is divided into real data and generated data, it is necessary to add the two probabilities together. So as the formula shows, when generating a network given G, $V(G, D)$ should be as large as possible.

$$\begin{aligned} V(G, D) = E_{x \to Pdata(x)} \\ log[D(X)] + E_{z \to P_z(z)} log[1 - D(G(z))] \end{aligned} \tag{3}$$

The optimal solution for discriminant network D is:

$$D_G^* = \underset{D}{argmax}\, V(G, D) \tag{4}$$

For the generative network G, when the discriminant network D is given, the optimal solution of the generative network G is shown in :

$$G_D^* = \underset{G}{argmin}\, V(G, D) \tag{5}$$

3.2. **Optimization of the confrontation process.** The optimization process can be redefined as a minimum maximum game problem involving multiple participants, with the core objective of finding a strategy combination in which no participant can achieve better results by individually changing their strategy. This strategy combination is known as the Nash equilibrium point, that is, until the discriminant model is unable to distinguish between the fake samples generated by the generative model and the real samples.

$$G_D^* = \underset{G}{argmin}\, V(G, D) \tag{6}$$

3.3. **Insurance fraud detection model based on GAN.** At present, there are many evolutionary models of GAN, such as CGAN to solve the problem of GAN being too free, propose GAN with conditional constraints, and introduce conditional variables in modeling to add conditions to the model to guide the generation of data; WGAN Optimizes generative adversarial networks from the perspective of loss functions, so that it also has good performance results in the fully connected layer, and the loss function of the discriminant network and the generation network is not logged; WCGAN, like CGAN, adds conditional variables to WGAN. When applied to image processing, the generator of Generative Adversarial Networks (GANs) is responsible for creating two-dimensional images containing three color channels, with each pixel performing the same, while the discriminator evaluates these images. In order to explore the spatial structure characteristics of image data, convolutional transformation is usually used between network layers. In this convolutional layer, each neuron only processes a small portion of the input and output (such as neighboring pixels in an image) to capture spatial relationships. However, the variables in the Medicare dataset do not contain this spatial structure, so it is necessary to adjust the convolutional network to include fully connected layers. In a fully connected layer, each neuron is connected to all inputs and outputs of that layer, allowing the network to reveal the interrelationships between features.

The principle of GANs is that there are two models, one generative model and one discriminant model. The discriminant model is employed to determine whether a given data is real data (data obtained from the dataset), and the task of generating the model is to create data that is as identical as possible to the real data. These two models work together against training, generate new data to generate new data to deceive the discriminant model, and then discriminant model to judge whether these data are real or fake. In the process of training these two models, the performance of the two models continues to improve, and finally reaches steady state. In order to be able to enhance the text type, all the convolutional layers in the traditional GAN network are replaced with fully connected layers, and the traditional CGAN, WGAN, WCGAN network structure is also replaced, taking GAN as an example, as shown in Figure 2, the created model structure includes the generative network and the discriminant network part.

Figures 3, 4and 5 are the generative network structure, discriminant network structure and core adversarial training structure of GAN network, respectively. After the generation network model and the discriminant network model are trained, they are combined, the upper part is the generation network, the lower part is the discriminant network, the

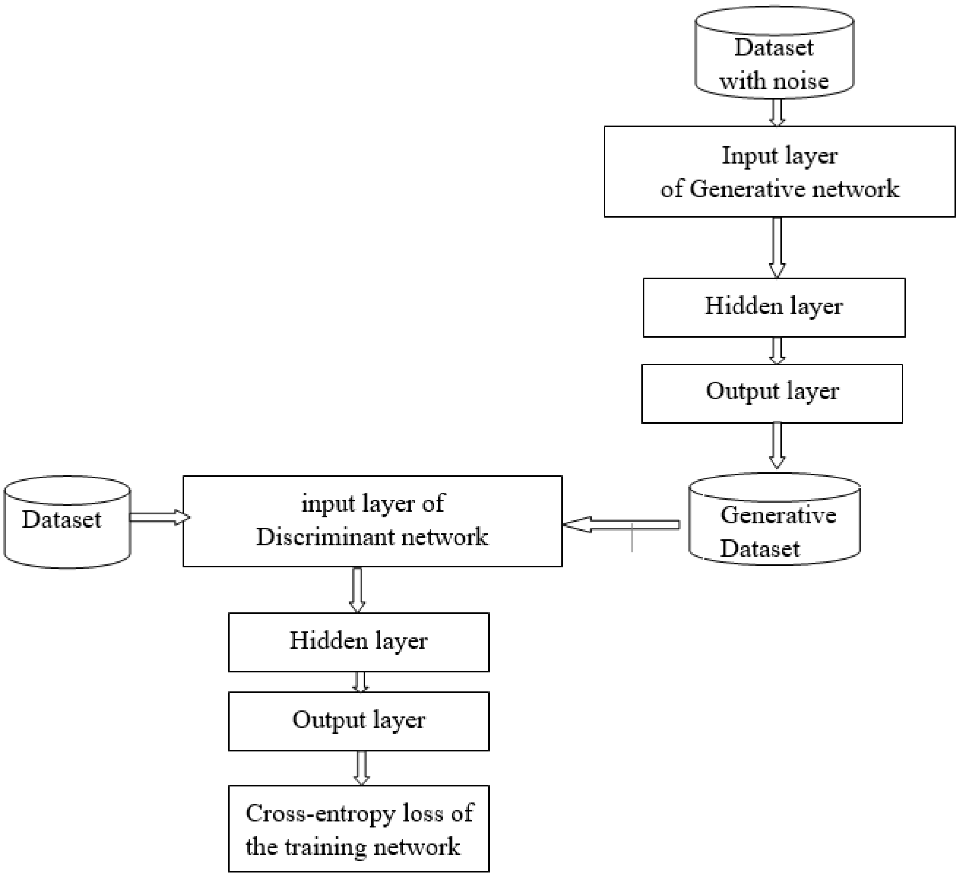
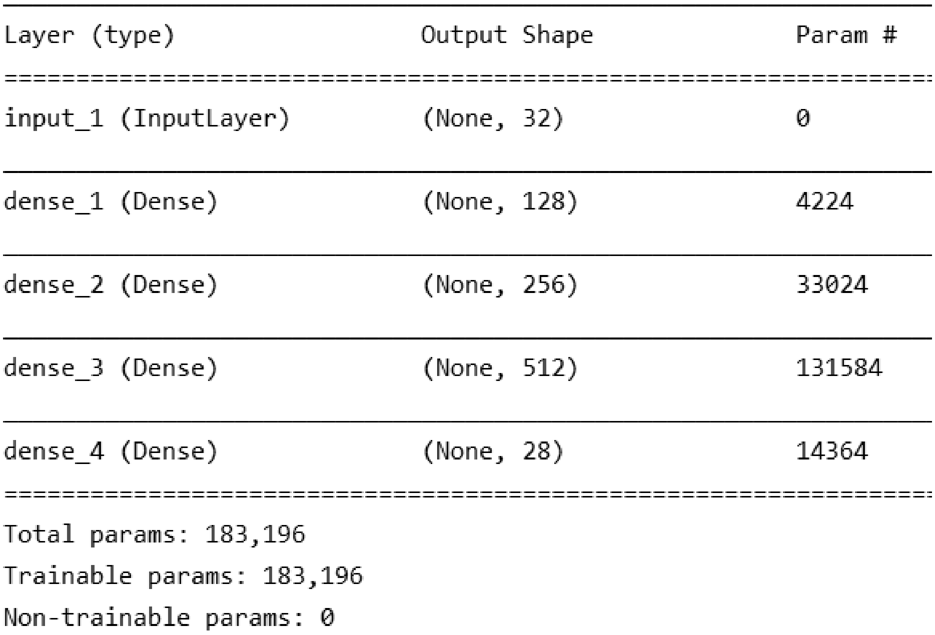
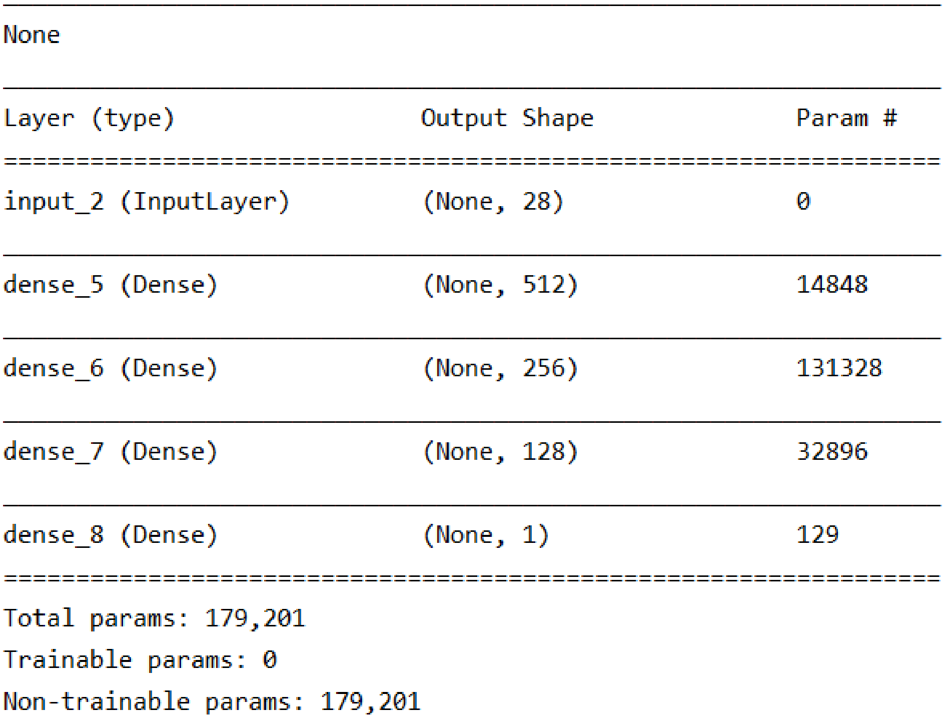

Dataset with noise
Input layer of Generative network
Hidden layer
Output layer
Generative Dataset
Dataset
input layer of Discriminant network
Hidden layer
Output layer
Cross-entropy loss of the training network

Figure 2. Flowchart of the GAN network in Insurance fraud detection.



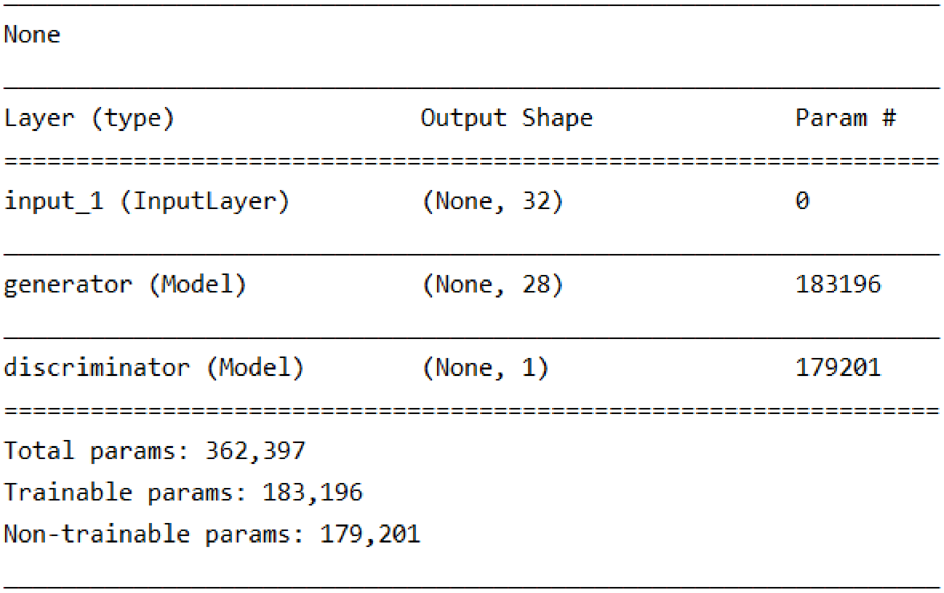
```
Layer (type)                 Output Shape              Param #
=================================================================
input_1 (InputLayer)         (None, 32)                0

dense_1 (Dense)              (None, 128)               4224

dense_2 (Dense)              (None, 256)               33024

dense_3 (Dense)              (None, 512)               131584

dense_4 (Dense)              (None, 28)                14364
=================================================================
Total params: 183,196
Trainable params: 183,196
Non-trainable params: 0
```

Figure 3. Generative network structure of GAN network.



```
_________________________________________________________________
None
_________________________________________________________________
Layer (type)                 Output Shape              Param #
=================================================================
input_2 (InputLayer)         (None, 28)                0
_________________________________________________________________
dense_5 (Dense)              (None, 512)               14848
_________________________________________________________________
dense_6 (Dense)              (None, 256)               131328
_________________________________________________________________
dense_7 (Dense)              (None, 128)               32896
_________________________________________________________________
dense_8 (Dense)              (None, 1)                 129
=================================================================
Total params: 179,201
Trainable params: 0
Non-trainable params: 179,201
```

Figure 4. Discriminant network structure of GAN network.



```
_________________________________________________________________
None
_________________________________________________________________
Layer (type)                 Output Shape              Param #
=================================================================
input_1 (InputLayer)         (None, 32)                0
_________________________________________________________________
generator (Model)            (None, 28)                183196
_________________________________________________________________
discriminator (Model)        (None, 1)                 179201
=================================================================
Total params: 362,397
Trainable params: 183,196
Non-trainable params: 179,201
_________________________________________________________________
```

Figure 5. Adversarial training structure of GAN network.

generated network generates medical fraud sample data, and sends it to the discriminant network to judge the authenticity of the fraud sample, and during the entire adversarial training process, only the generation network is adjusted, and the parameters of the discriminant network do not need to be changed.

The proposal of GAN has triggered a research boom in the academic industry. Primitive GANs have problems such as difficulty in interpretation and training. Therefore, many scholars have also studied derivative models of GAN, including CGAN (Conditional Generative Adversarial Network), WGAN (Generative Adversarial Network with gradual penalty), WCGAN (combination of the first two), SSGAN (Semi-supervised Generative Adversarial Network) and so on. CGAN is a simple extension of GAN, adding the conditional variable y in the input data of G and D, y can be a category label attribute or data supplement, to a certain extent, it solves the problem that the GAN generation result is difficult to control, and has certain restrictions on the label category of the generated sample. That is, the input of the CGAN generator G has an extra part of the condition y. At this point, the objective function during CGAN optimization is shown in Equation 7:

$$\begin{aligned} &\underset{G}{min}\,\underset{D}{max}\,V(G,D) = \underset{G}{min}\,\underset{D}{max}\,E_{x\rightarrow Pdata(x)} \\ &log[D(x|y)] + E_{Z\rightarrow P_Z(Z)}log[1-D(G(z|y))] \end{aligned} \qquad (7)$$

$$(8)$$

Compared with traditional GAN, WGAN introduces the concept of Wasserstein distance, and changes the output of the discriminator and the update process of the loss function. The changes of WGAN have improved the problems of unstable GAN network training and gradient explosion to a certain extent, making convergence more stable. The discriminator in SSGAN not only needs to judge the source of the sample, but also classify the sample, which makes the discriminator more discriminant.

4. **Experiment.**

4.1. **Data set.**

For the sake of ensuring the consistency and authenticity of the data, the data-set used in this paper is the dataset of the Tianchi Precision Social Security Competition, which is derived from the real medical insurance data and has great reference value for the establishment of medical insurance models. The data given by the competition questions mainly include personal codes, the amount of various drugs generated, the amount of subsidies, the number of medical visits, and the time of medical treatment. The dataset contains two data tables, the data table $df_train_file_final$ records information about the insured person, and the other df_id_traintable records the insured person's code and the label of participation in fraud (0-normal, 1-fraud), which are joined by the primary key enrolle code. The relationship between the two data is shown in Figure 6:

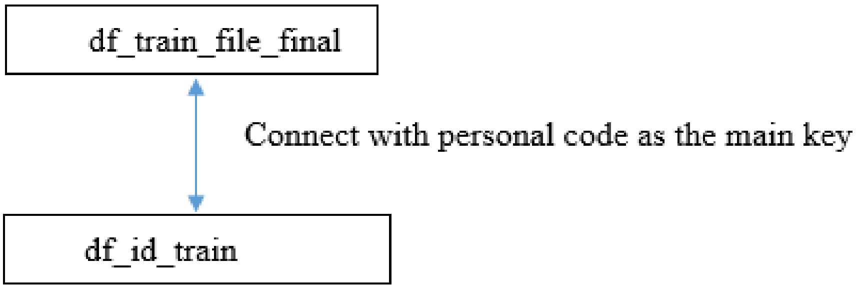


FIGURE 6. The relationship between the two data.

Data preprocessing is a crucial step in the machine learning process. The preparation of data significantly affects the effectiveness of classification models when identifying medical insurance fraud activities. We found that the dataset contains three types of variables: time-type, numeric, and categorical.

The time-type variables are the transaction time and the time of the doctor's visit, and the categorical variables are divided into dichotomous and multi-categorical variables. The fraud label of the insured person is a dichotomous variable, with two attributes, 0 (no) and 1 (yes), and the hospital category and discharge diagnosis disease are multi-categorical variables with multiple different attributes. Categorical variables are discrete variables and have an impact on the model if they are directly converted to ordinal variables for processing. For the handling of discrete variables, One-hot encoding is a relatively common way. One-hot coding, also known as one-bit effective coding, is used to encode each different state through multi-bit registers, each of which corresponds one-to-one to its unique register state, with only one effective at any time. This coding method enables the classifier to effectively process categorical data and expand the features to a certain extent. In this paper, One-hot coding is used to encode some categorical features and obtain the standard deviation.

4.2. **Machine learning model evaluation.** Generally, the following metrics are used for model evaluation of the trained model:

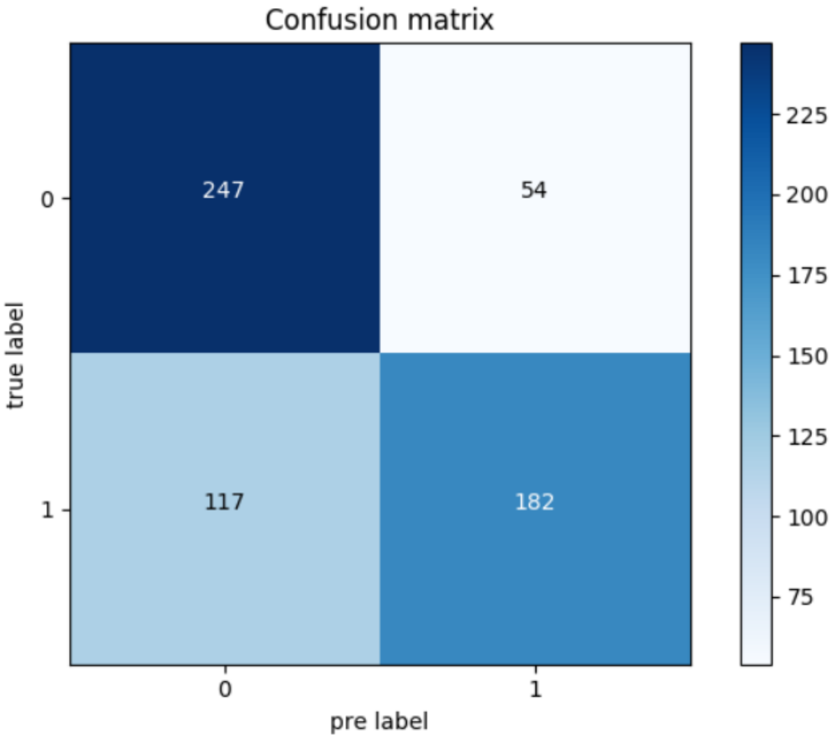


FIGURE 7. Confusion matrix generated by health care data.

4.2.1. *Confusion matrix.* The confusion matrix is composed of a coordinate system, with the x-axis representing the predicted values and the y-axis representing the true values. we could calculate the metric value measured by the current model by comparing the outcome difference. Figure 9 plots a confusion matrix for down-sampling of medicare fraud data:

True Positive (TP): A positive sample predicted as positive by the model.

False Positive (FP): A negative sample predicted as positive by the model.

False Negative (FN): A positive sample predicted as negative by the model.

True Negative (TN): Negative samples predicted as negative by the model.

True Positive Rate (TPR): TPR=TP/(TP+FN), that is, the number of positive samples predicted as positive / the actual number of positive samples.

False Positive Rate (FPR): FPR=FP/(FP+TN), that is, the number of negative samples predicted as positive / the actual number of negative samples.

False Negative Rate (FNR): FNR=FN/(TP+FN), that is, the number of positive samples predicted as negative / the actual number of positive samples.

True Negative Rate (TNR): TNR=TN/(TN+FP), that is, the number of negative samples predicted as negative / the actual number of negative samples/2.

4.2.2. *Accuracy.* Accuracy is the most commonly used indicator for evaluating the performance of classification models, which is evaluated as: $Accuracy = (TP + TN)/(TP + FN + FP + TN)$.

That means the number of correctly predicted positive and negative examples divided by the total.

4.2.3. *Recall.* Recall is a more scientific way to detect models. Recall = TP/(TP+FN).

4.2.4. *Precision.* Precision is only for positive samples that are correctly predicted, not for all samples that are correctly predicted. It is manifested by predicting how much of what is positive inside is really positive. It can be describled as the accuracy rate. Precision = TP/(TP+FP) is the number of correctly predicted positive cases / the total number of predicted positive cases.

4.2.5. *F1 score.* The F1 score is the harmonic value of precision and recall, which is more inclined towards the smaller of two numbers, so the F1 value is the largest when the precision and recall are proximate. Many recommendation systems are evaluated using F1 values.

$$2/F1 = 1/Precision + 1/Recall \tag{9}$$

Table 1. prediction results comparison between downsampling and oversampling strategy.

Method	Accuracy	Recall	F1
Downsampling	0.957	0.74	0.98
oversampling	0.956	0.69	0.98

The generator network is set up with an input layer with 32 neurons and 3 hidden layers, and the number of neurons is 128,256,512, each layer uses ReLU as an activation function. Since the original data (dimensionally reduced) has 28 dimensions, the output layer is set to 28 neurons, representing a vector that mimics the original data. The input layer of the discriminator network has 28 neurons, which is consistent with the total number of data features. The three hidden layers contain 512,256,128 neurons each, and the output layer uses sigmiod as an activation function to output the classification probability. After defining the generator model and discriminator model respectively, a fusion model is defined. The model combines the input-output functions of the two models, the input layer receives the input vector of the generator, and the output layer outputs the output vector of the discriminator.

The Figure 11 a real comparison chart of the network-generated data and real fraud data generated by the network with the increase in the number of model training cycles, and the number of steps is 0, 100, 200, 300, 400, 500, respectively, it is obvious that the model gradually generates data that is more and more similar to the original data with

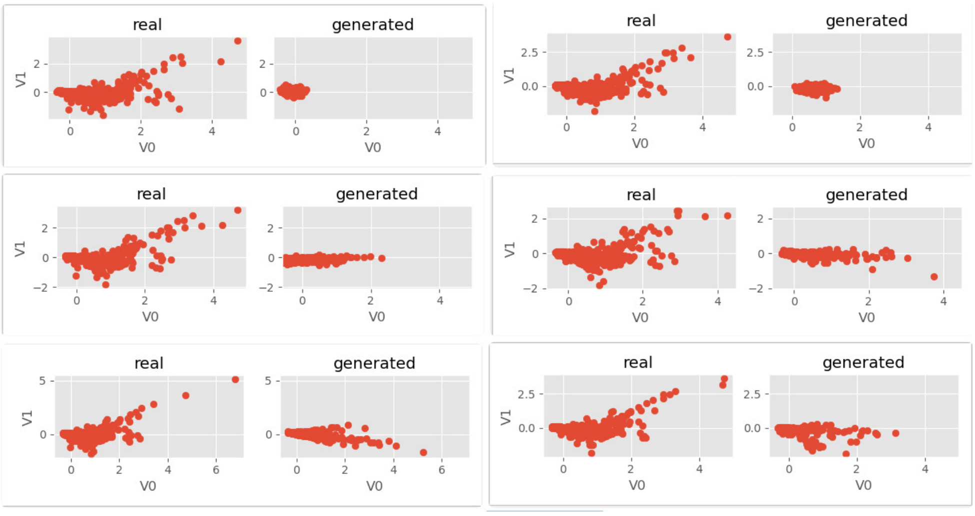


Figure 8. Training data generation process (GAN, CGAN, WGAN, WCGAN, respectively).

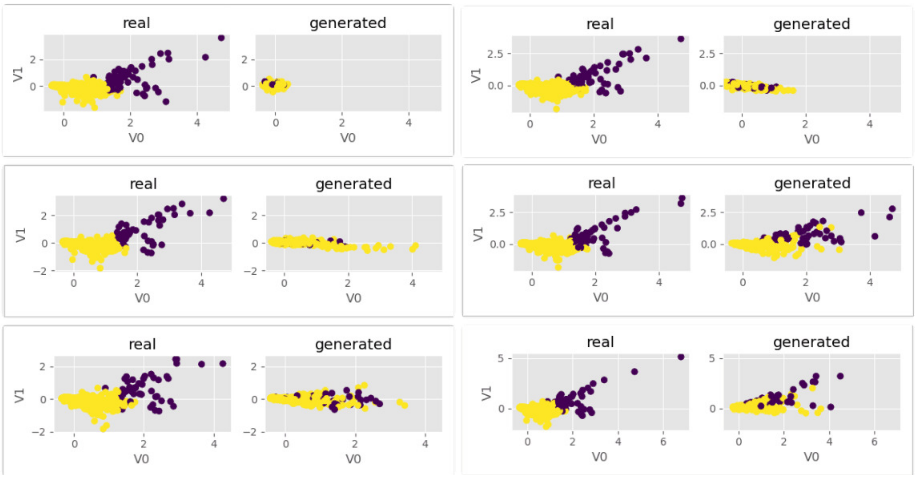


Figure 9. A real comparison chart of the network-generated data and real fraud data generated by the network.

the increase of the number of steps during the training process. From the perspective of data distribution, the WCGAN model works better. Figure 5.7 is the probability that the network in WCGAN for the generation of data generated by the network is false, and it can be seen that the model is not good for the generation of medical insurance fraud data, and it remains floating as the number of steps slowly drops to around 80%. The reason for this phenomenon may be that the processing of fraud dataset features in feature engineering is not good enough, and GAN networks are usually used for the generation of image data, and are not suitable for processing text data, while medical insurance fraud data itself is discrete, for the traditional GAN model, the generation network outputs different results, and the discrimination network gives the discriminant results, but can not well transmit the gradient update information to the generation network, so the discrimination of the discriminant network output is meaningless. Although GAN has been improved to replace the convolutional layer with a fully connected layer, it can be seen from the results that the generated data samples do not work well.

We visualize the training effect of GAN network to find out the steps with better training effect. Set the total number of iterations of GAN to 500, and view the training effect once every 100 steps, a total of 6 training results, and the results of GAN, CGAN, WGAN, WGAN training data generation are shown in Figure 10, Figure 11, Figure 12,and Figure 13 respectively.

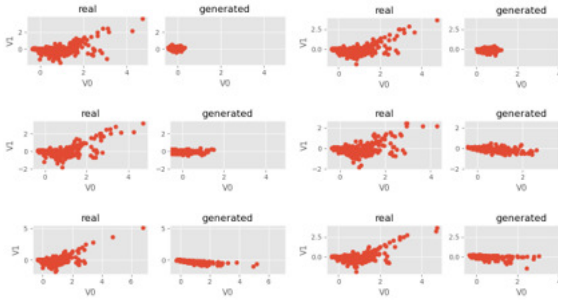


Figure 10. Training results of GAN.

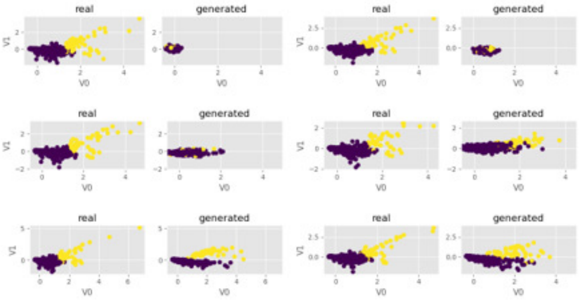


Figure 11. Training results of CGAN.

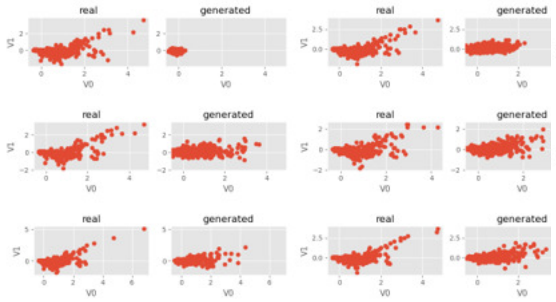


Figure 12. Training results of WGAN.

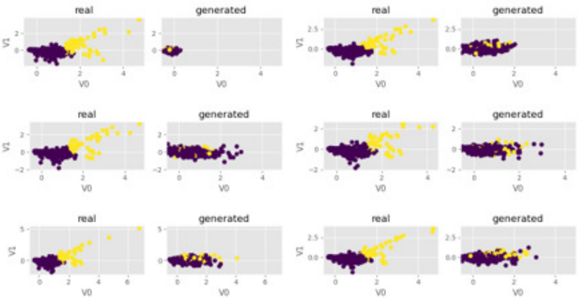


Figure 13. Training results of WCGAN.

The number of iteration steps of training is 0, 100, 200, 300, 400, and 500, respectively, and it can be found that with the increase of the number of steps, the dispersion of generated data gradually approaches the distribution of real data. When training the CGAN and WCGAN models, the conditional variable y input by the generator and discriminator is the categorical attribute of the label itself, so in the training process, it can be found that the real data and the generated data contain the distribution structure of fraud samples and normal samples. From the training effect graph, it can be seen that CGAN and WGAN perform better in imitating the distribution of the original dataset. However, because GAN and WGAN networks do not have label attributes, they are too free in generating sample categories, and even if the effect is better, they cannot fix the fraudulent label samples we need. Therefore, we choose the CGAN and WCGAN networks with the tag attribute. It can be seen from the figure that the CGAN network is relatively close to the real data distribution, so the CGAN model is selected to generate fraudulent data.

After training, load the model, take some raw data and import it for generation. We compare the characteristics between the original data and the generated data. It can be found that the distribution of the generated data and the original data in each feature is pretty close, which shows the high quality of the generated data. XGBoost (eXtreme Gradient Boosting) is an ensemble learning framework, a gradient enhancement algorithm enhancement library, proposed by Chen et al. in 2015 [5]. It is a strong classifier, and the design idea is to combine several weak classifiers (usually tree structures) with low classification accuracy to produce a classifier with high accuracy. XGBoost is an improved version of the GBDT (Gradient Boosting Decision Tree) algorithm. The idea of GBDT is to add a new weak classifier every time, correct the bias of other classifiers in front, optimize it in each round of iteration, and finally the accuracy of the classifier after continuous training will be stronger than that of a single weak classifier, follow the gradient descent idea when generating each tree, and refer to the information of all trees before generating a new tree to minimize the loss function. To achieve high accuracy, thousands of iterative calculations are required to generate a large number of trees, which takes a long time. GBDT only seeks the first-order derivation, and XGBoost performs the second-order Taylor expansion in the cost function processing, which makes the optimal solution of the model more efficient, and adds regular terms to the cost function to limit the total number of leaf nodes and weight size, so that the model expression is more concise and avoids the model falling into an over-fitting state.

XGBoost uses multiple CPU cores for parallel training, training speed is fast, and supports user-defined objective functions and evaluation functions, only needs to meet the second-order drivability of functions . For each round of iterations, XGBoost has built-in cross-validation rules, which can easily obtain the optimal number of iterations. The XGBoost algorithm also has excellent classification effect on the basis of fast running speed, and supports the function of custom hyper-parameters and loss functions.

Here, the data generated by the GAN network and the data in part of the original dataset are used as the training set. I selected half of the fraud samples in the original data-set (i.e., Class category 1) and sent them to the generation network to simulate the fraud data, and then adjusted some parameter information of model training to improve the performance of the model. The GridSearchCV method is used here. The graph below shows the change in the XGBoost loss curve as the number of iterations increases. It can be seen that after nearly 300 iterations, the curve tends to level off. Therefore, set the hyper-parameter $n_estimators$ to 300.

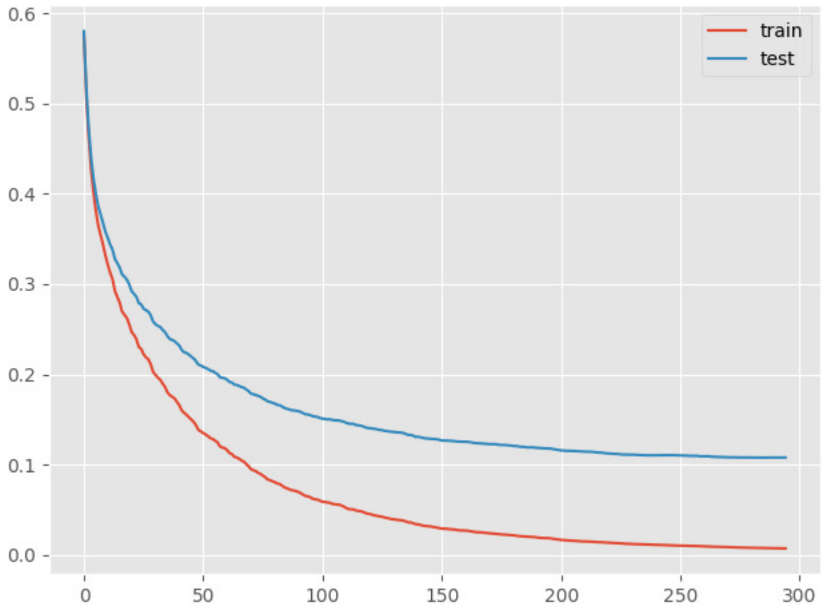


FIGURE 14. Loss curve of XGBoost.

After adjusting the parameters, the XGBoost model basically avoids the phenomenon of model over-fitting, and performs well on data such as classification accuracy and summoning rate. After the model is built, the influence of each feature on the model is as follows.

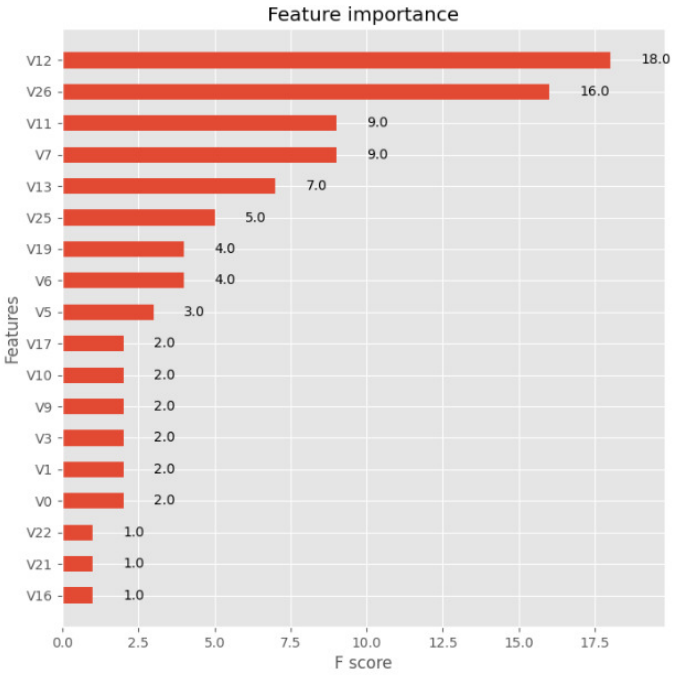


Figure 15. Importance of the features.

Table 2shows the prediction results on the test set of the original medical insurance data-set after under-sampling, over-sampling and GAN network processing above, and the model obtained by logistic regression model and XGBoost model training.

Table 2. prediction results comparison on the test set of the original medical insurance data-set

Method	Precision	Recall	F1
Downsampling-LR	0.956	0.730	0.828
Downsampling-XGBoost	0.811	0.869	0.839
SMOTE-LR	0.958	0.783	0.862
SMOTE-XGBoost	0.912	0.903	0.908
BorderlineSMOTE-LR	0.956	0.805	0.874
BorderlineSMOTE-XGBoost	0.926	0.869	0.896
GAN-LR	0.938	0.728	0.819
GAN-XGBoost	0.944	0.964	0.954

It can be seen from Table 2 that when under-sampling is used as the sampling method, the classification performance on logistic regression and XGBoost model is not very good due to the large loss of features, while the processing effect of the two over-sampling is relatively good, and the BorderlineSMOTE and SMOTE methods have no particularly obvious difference in the test set in this paper. Compared with the logistic regression model and the application of XGBoost in other sampling methods, GAN-XGBoost performs best in classification accuracy, recall and f1 score, and has more ideal fraud detection ability and practical application prospects.

5. **Conclusion.** In this paper, the medical insurance fraud detection model is established by using of machine learning and deep learning, and after specific learning of the relevant algorithms, a model processed by various methods is established, and GAN-XGBoost, a detection model with certain effectiveness, is obtained, and certain results are achieved. On the basis of the collected relevant data sets, the data is preprocessed, including data cleaning, missing value processing, dimensionlessness, feature coding, etc. The imbalanced samples were focused on processing, two sampling methods were adopted, and considering the limitations of the SMOTE method, the BorderlineSMOTE method was introduced. On the basis of the two sampling methods, a GAN network is built to further process unbalanced samples in view of the problems caused by the two sampling methods. After processing the data-set, logistic regression and XGBoost classification model were used to train, and good training parameters were obtained by grid search and cross-validation. Through comparative experiments, the model was evaluated by using various evaluation indicators such as accuracy, recall, and F1 score, and it was concluded that the medical insurance detection model of GAN-XGBoost had obvious advantages in evaluation indicators.

Acknowledgment. This research is sponsored by National Natural Science Foundation of China under Grant No.62002174. We also deeply appreciate the valuable feedback and suggestions provided by the reviewers, which have helped us improve the quality of the article.

REFERENCES

[1] People's daily. there are 1.36 billion people enrolled in basic medical insurance in china.

[2] H. A. Improving efforts to combat health care fraud. *U.s.department of Health and Human Services*, 2011.

[3] Y.-B. C. and U. OA. Review of generative adversarial networks and its application in cybersecurity. *Artif Intell Rev*, 53:1721–1736, 2020.

[4] C. L. Chan and C. H. Lan. A data mining technique combining fuzzy sets theory and bayesian classifier-an application of auditing the health insurance. In *International conference on artificial intelligence*, 2001.

[5] T. Chen and C. Guestrin. Xgboost: A scalable tree boosting system. In *KDD*, pages 785–794, 2016.

[6] R. D, R. G, and C. G. Data mining and the implementation of a prospective payment system for inpatient rehabilitation. *Health Serv Outcomes Res Method*, 3(3-4):247–266, 2002.

[7] T. D. Predicting healthcare fraud in medicaid: A multidimensional data model and analysis techniques for fraud detection. *ScienceDirect*, pages 1252–1264, 2003.

[8] T. D, M. R. M, S. P, and et al. Predicting healthcare fraud in medicaid:a multidimensional data model and analysis techniques for fraud detection. *Procedia Technology*, 9:1252–1264, 2013.

[9] H. Denenberg. The denenberg report: the insurance commissioners, other government agencies, and the insurance companies focus on insurance fraud committed by policyholders, but nothing is done about the multi-billion dollar racket of insurance fraud committed by insurance companies, 2005.

[10] H. G. E, O. S, and T. Y. A fast learning algorithm for deep belief nets. *Neural Computation*, 18:1527–1554, 2006.

[11] L. F, T. Y, and C. J. A process-mining framework for the detection of healthcare fraud and abuse. *Health Care Manage*, pages 353–358, 2008.

[12] T. F, R. A, M. S, and K. A. Fraud in the health systems of chile: a detection model. *Am J Public Health*, pages 56–61, 2008.

[13] M. Guo. *Application of data mining technology in disease diagnosis related grouping.* PhD thesis, Central south university, 2009.

[14] M. Guo. *Research and application of medical insurance fraud detection based on Hadoop platform.* PhD thesis, University of Electronic Science and Technology of China,UESTC, 2017.

[15] H. He, S. Hawkins, and X. Yao. Application of genetic algorithm and k-nearest neighbour method in real world medical fraud detection problem. *Journal of Advanced Computational Intelligence and Intelligent Informatics*, 4(2):130–137, 2020.

[16] H. J. and K. T.M. Medicare fraud detection using catboost. pages 97–103, 2020.

[17] L. J, H. K, J. J, and S. J. A survey on statistical methods for health care fraud detection. *Health Care Manage Sci*, 20(10):275–287, 2008.

[18] M. M. Predicting healthcare fraud in medicaid: A multidimensional data model and analysis techniques for fraud detection. *ScienceDirect*, pages 1252–1264, 2003.

[19] K. ML, F. LB, S. MK, R. B, U. PH, and B. N. T. LPC. Data mining in healthcare: Applying strategic intelligence techniques to depict 25 years of research development. *Int J Environ Res Public Health*, 18(6):3099–3114, 2021.

[20] V. MS, N. JP, and R. MJ. Applying data mining techniques to a health insurance information system. In *Proceedings of the 22nd VLDB conference*, 1996.

[21] O. PA, F. CJ, and R. GA. A medical claim fraud/abuse detection system based on data mining: a case study in chile. In *Proceedings of international conference on data mining*, 2006.

[22] M. F. Rabbi, M. N. Sultan, M. Hasan, and M. Z. Islam. Tribal dress identification using convolutional neural network. *Journal of Information Hiding and Multimedia Signal Processing*, 14(3):72–80, 2023.

[23] M. M. I. Raju, S. Sarker, and M. M. Islam. Chronic kidney disease prediction using ensemble machine learning. *Journal of Information Hiding and Multimedia Signal Processing*, 14(1):1–9, 2023.

[24] B. S. Predictive solutions bring more power to decision makers. *Health Management Technology*, 20(10):12–14, 1999.

[25] M. A. S. Sardar, H. Saha, M. N. Sultan, and M. F. Rabbi. Intrusion detection in electric vehicles using machine learning with model explainability. *Journal of Information Hiding and Multimedia Signal Processing*, 14(3):81–89, 2023.

[26] O. T, M. N, B. L, L. C, and S. C. Using ethnography to design a mass detection tool (mdt) for the early discovery of insurance fraud. In *Proceedings of the ACM CHI conference*, 2003.

[27] Z. Tingting. *Research on medical insurance anomaly detection based on deep learning.* PhD thesis, University of Electronic Science and Technology of China, 2019.

[28] Y. WS. Process analyzer and its application on medical care. In *Proceedings of 23rd International conference on information systems (ICIS02)*, 2002.

[29] Y. WS and H. SY. Detecting hospital fraud and claim abuse through diabetic outpatient services. *Expert Syst Appl*, pages 31(56–58), 2006.

[30] S. Y, J. D, M. D, and Sutinen. A mining medical specialist billing patterns for health service management. In *Proceeding 7th Australasian data mining conference (AusDM 2008)*, 2008.

[31] D. Yi, G. Deng, C. Dong, M. Zhu, and et al. Medical insurance fraud detection algorithm based on graph convolutional neural network. *Computer Application*, 40(5):1272–1277, 2020.

[32] Q. Zhao. *Analysis of fraud forms and countermeasures of medical insurance funds.* PhD thesis, Chinese Academy of Social Sciences,CASS, 2012.