Journal of Healthcare Engineering

Data Mining in Medicine

Lead Guest Editor: Lei Zhang
Guest Editors: Liangyin Chen and Ke Yan

Data Mining in Medicine

Journal of Healthcare Engineering

Data Mining in Medicine

Lead Guest Editor: Lei Zhang
Guest Editors: Liangyin Chen and Ke Yan

Associate Editors

Academic Editors

Contents

Hindawi
Journal of Healthcare Engineering
Volume 2021, Article ID 6158961, 8 pages
https://doi.org/10.1155/2021/6158961

Research Article

Cost Control of Treatment for Cerebrovascular Patients Using a Machine Learning Model in Western China

Siyu Zeng , Li Luo, Yuanchen Fang , and Xiaozhou He

Business School, Sichuan University, No. 24 South Section 1, Yihuan Road, Chengdu, China

Correspondence should be addressed to Yuanchen Fang; y.fang@scu.edu.cn

Received 7 May 2021; Accepted 8 November 2021; Published 22 November 2021

Academic Editor: Antonio Gloria

Background. Cerebrovascular disease has been the leading cause of death in China since 2017, and the control of medical expenses for these diseases is an urgent issue. Diagnosis-related groups (DRG) are increasingly being used to decrease the costs of healthcare worldwide. However, the classification variables and rules used vary from region to region. Of these variables, the question of whether the length of stay (LOS) should be used as a grouping variable is controversial. *Aim*. To identify the factors influencing inpatient medical expenditure in cerebrovascular disease patients. The performance of two sets of classification rules, and the effects of the extent of control of unreasonable medical treatment, were compared, to investigate whether the classification variables should include LOS. *Methods*. Data from 45,575 inpatients from a Healthcare Security Administration of a city in western China were used. Kruskal–Wallis H tests were used for single-factor analysis, and multiple linear stepwise regression was used to determine the main factors. A chi-squared automatic interaction detector (CHAID) algorithm was built as a decision tree model for grouping related data. The intensity of oversupply of service was controlled step by step from 10% to 100%, and the performance was calculated for each group. *Results*. The average hospitalization cost was 1,284 US dollars, and the total was 51.17 million US dollars. Of this, 43.42 million were paid by the government, and 7.75 million were paid by individuals. Factors including gender, age, type of insurance, level of hospital, LOS, surgery, therapeutic outcomes, main concomitant disease, and hypertension significantly influenced inpatient expenditure ($P < 0.05$). Incorporating LOS, the patients were divided into seven DRG groups, while without LOS, the patients were divided into eight DRG groups. More clinical variables were needed to achieve good results without LOS. Of the two rule sets, smaller coefficient of variation (CV) and a lower upper limit for patient costs were found in the group including LOS. Using this type of economic control, 3.35 million US dollars could be saved in one year.

1. Introduction

Cerebrovascular disease and its complications are the leading cause of disability and death worldwide. Of all the diseases of the nervous system, cerebrovascular diseases have the greatest impact on disability and produce the highest economic burden [1–3]. Since 2017, this disease has become the leading cause of death in China [4]. The number of people suffering from cardiovascular and cerebrovascular diseases in China was 330 million in 2019, and these diseases are the leading cause of death among urban and rural residents [5]. In 2017, the total cost of treating cerebrovascular diseases in China reached 83.83 billion US dollars, ranking first among all diseases and accounting for 17% of the total medical cost of treating diseases, equivalent to 0.66% of GDP [6]. One city alone spent 51.17 million US dollars a year on these diseases in this study. In the face of so much economic pressure, the government must take effective action to reduce the economic burden of cerebrovascular diseases.

Diagnosis-related groups (DRG) are one of the most advanced medical payment management methods, aiming to reduce inefficiency and contain costs [7]. Based on factors such as a patient's demographic information, diagnosis, and disease severity, DRG-based payment systems group patients with similar clinical attributes requiring similar care, providing the necessary framework to aggregate patients into case types or products, which entail the use of similar resources [8]. DRG adopt a standard pricing framework for a single disease group [9] and provide equity in payments across healthcare providers for services of the same kind. Most studies have found DRG to have positive effects on controlling medical expenses and reducing the economic burden among patients [10]. Studies into cerebrovascular diseases have found that DRG can effectively reduce unreasonable costs incurred in the treatment of cerebrovascular diseases [11, 12]. However, the rules of the grouping vary between countries and regions; for example, length of stay (LOS) is widely used as a statistical classification index in research into DRG management in Poland, Britain, and other developed countries [10]. Japan uses LOS as a secondary parameter [9]. However, Finland and Sweden do not consider LOS [13].

China Healthcare Security Diagnosis-Related Groups (CHS-DRG) are the unified grouping standard used by the national pilot city [14]. Due to the unbalanced development of China's economy, the Chinese government requires cities to develop localized grouping rules based on their actual conditions, so there are variations of DRG payment policy design and grouping rules across China [15]. Beijing Diagnosis-Related Groups (BJ-DRG) are the earliest localization group in China; Beijing built Chinese Diagnosis-Related Groups (CN-DRG) following the model of the All-Patient Diagnosis-Related Groups (AP-DRG) in the USA, and Shanghai built a Shanghai-DRG and National standards for paying fees according to DRG (C-DRG) based on the Australia Refined DRG (AR-DRG). However, these grouping methods are all based on the data collected from the first-tier developed cities in China, and there is no research into the underdeveloped cities in the west of the country. It is inappropriate for cities in the west to use the same rules, due to the unbalanced economic and technological development in China [16]. None of those grouping rules take into account the LOS, unlike most countries in Asia, which incorporate LOS [17].

In this study, we collected data from an underdeveloped city in western China. Machine learning was used to group patients with similar costs, and two sets of rules were built, one incorporating LOS and the other without LOS. We compared the performance of the grouping rules based on the coefficient of variation (CV) to assess the heterogeneity within a group, as has been done in previous studies [8]. We identified the outliers in each group and considered them to represent unreasonable costs. Finally, we tried to control these costs to different extents. This study fills the gap in previous studies, which have only focused on developed cities and which use CV as the standard measure of the results of grouping. In our study, underdeveloped cities and control performance were considered.

The rest of this paper is organized as follows. In Section 2, we introduce our materials and methods. In Section 3, we present our results, including general information and inpatient medical expenditure, single and multiple factor analysis of the factors influencing inpatient medical expenditure, the results of two sets of rules for DRG grouping, medical expenses in different DRG, and payment method adjustment results. In Section 4, we discuss the results. Section 5 concludes this study and provides a description of directions for future research.

2. Materials and Methods

2.1. Patient Data. The data used in this research were collected from the Healthcare Security Administration of a city in western China during 2018. The data included medical records and cost information related to 93,185 inpatients with cerebrovascular diseases (ICD-10:60-69) as the principal diagnosis, all of which under the major diagnostic categories (MDC) of diseases and dysfunction of the nervous system (MDCB). Original information on these patients included 58 variables, such as gender, age, LOS, cost of hospitalization, payment of medical insurance, and type of insurance.

2.2. Data Cleaning. In the first step of data cleaning, we selected data from only the comprehensive grade tertiary and secondary hospitals. The patients from township hospitals, community hospitals, and school hospitals were removed. As a second step, we eliminated outliers in costs [8] and patients younger than 18 years of age. Finally, patients who were not hospitalized in our study city but were reimbursed by the city's Medical Insurance Bureau were excluded. Valid data from a total of 45,575 patients were obtained after screening.

2.3. Statistical Analysis and Data Grouping. The proportions of the training set and the test set were 80% and 20%, respectively. Firstly, the training set is grouped, and the effect of grouping is detected with the data of the test set. Finally, all the data are put into grouping rules and analyzed.

Kruskal–Wallis tests were used for single factor analysis to determine the factors influencing hospitalization expenses. Values of $P < 0.05$ were considered to be statistically significant [18]. Stepwise multiple generalized linear regression was used for variance analysis [19]. The medical costs for different subgroups were calculated, and the statistically significant variables with the greatest impacts on medical costs were selected for grouping analysis.

The Chi-Squared Automatic Interaction Detection (CHAID) algorithm was used to establish the combination of DRG [10, 20]. In the selection of grouping variables, we considered both the inclusion and exclusion of LOS, CV, and the percentage of outliers. We considered a CV value of less than 1 to indicate no heterogeneity within a group, as has been done in previous studies [8]. We regarded outliers to represent unreasonable medical treatment and calculated the variation in unreasonable medical costs among different participants under different degrees of control. We used inpatient hospitalization expenditure as the dependent variable, and the variables selected by the generalized linear stepwise model were set as the independent variables. LOS was shown to have a significant positive influence on medical expenditure. In order to further investigate the grouping performance of LOS, we built two decision tree models. The first model used the LOS as a classification variable, and the second model omitted the LOS. We have conducted more than ten random trials using data sampling samples, and the results of each trial are consistent, which indicates that the performance of the algorithm is stable. All analyses were carried out using R.studio 4.0.2 software [21] with the CHAID package [22].

3. Results

In the following section, we summarize general information about the patients' medical costs in Section 3.1, and single factor and multiple analysis are shown in Sections 3.2 and 3.3, respectively. The results of grouping using the two sets of rules based on machine learning are shown in Section 3.4. Finally, the performance of the algorithm using different levels of implementation control is presented in Section 3.5.

3.1. General Information and Inpatient Medical Expenditure. As shown in Table 1, women, individuals over 60 years old, and urban residents accounted for the majority of patients, while men, the elderly, and rural residents had relatively high expenses. Of the patients, 50.18% spent less than nine days in hospital, and 82.26% recovered after hospitalization. Of the patients with complete data, 19,488 (42.76%) were male and 26,087 (57.24%) were female; 1,995 (4.37%) were under the age of 45, while 9,117 (20%) were aged between 45 and 60, and 34,463 patients (75.64%) were older than 65. With respect to residence, 30,243 (66.36%) patients were urban workers, and 15,332 (33.64%) were rural residents. Among them, 24,482 (53.74%) were from a secondary grade hospital, and 21,087 (46.26%) were from a tertiary grade hospital. We also carried out statistical analysis on the effect of LOS, with surgery or without surgery, discharge status, and comorbidities complications (CCs) and whether there was grade III hypertension, on the distribution of patients' medical expenditure in different subgroups. The average expenditure of these patients was 1,284 US dollars. Among the subgroups, males, individuals aged over 65, rural residents, patients from tertiary grade hospitals, LOS more than 13 d, surgery, death, and CCs with insufficiency of blood supply to the cerebral arteries were more expensive.

3.2. Single Factor Analysis of the Factors Influencing Inpatient Medical Expenditure. In this study, 58 variables were examined using single-factor analysis (Table 1). Ten factors—gender, age, type of insurance, surgery, LOS, status on discharge, CCs, and a hypertension level of three—were shown to be associated with statistically significant differences in hospital expenditure, using Kruskal–Wallis tests ($P < 0.01$). Expenditure on men, individuals older than 60, rural residents, patients with longer LOS, patients undergoing surgery, death, and patients with CCs was the highest.

3.3. Multiple Factor Analysis of the Factors Influencing Inpatient Medical Expenditure. Generalized linear stepwise models were used for multiple regression analysis. Gender, LOS, level of hospital, surgery, status on discharge, type of insurance, comorbidities complications, and age had significant impacts on medical expenditure (Table 2). The R-squared value of the model was 0.521, and the kappa value was 12.08, indicating that the model performed well, and there was no multicollinearity between variables. All of these variables could be regarded as reasonable data for DRG grouping.

3.4. Two Rules for DRG Grouping and Medical Expenses in Different DRG. There were seven subgroups in model one and eight groups in model two. The hospital level was the main factor, and the second rule, without LOS, required more disease-related information, such as details of CCs. The group without LOS was more stringent. For example, grade A tertiary and grade B tertiary were in the same group under the rule incorporating LOS, while they were in different groups without LOS. The number of individuals in each group and details of expenses are shown in Tables 3 and 4. Most of the CVs of the first grouping method were less than 0.5, indicating that the homogeneity within the group was good, and the grouping effect was better in the grouping rules incorporating LOS. The weight calculation formula was (the average cost of the group)/(all the average costs). The higher the weight, the more resources consumed by the patients in the group. We set P75 + 1.5 IQR as the cost limit of each group, and the excess amount indicates the number of each group's medical expenses that were outside the cost limit.

We also analyzed the outliers of each group. Using the first grouping rules, the outliers were older than the normal patients, while using the second grouping rules, the outliers had a significantly longer LOS than the average.

3.5. Prediction of Medical Expenses Based on an Increasing Control Ratio of Unreasonable Treatment. In 2018, a total of 51.17 million US dollars medical expenses were related to 45,575 inpatients with cerebrovascular diseases as the principal diagnosis. The average cost was 1,248 US dollars. Among them, 43.42 million were paid by the Healthcare Security Administration, and 7.75 million were paid by patients themselves. All of this expenditure was based on the Fee for Service (FFS) payment system. We took the mean cost of each group as the payment standard for the DRG group and calculated the average cost to the Healthcare Security Administration, hospital, and patient.

The current FFS method encourages an oversupply of service in order to increase revenue [9]. We consider expenditure less than the cost limit in each group to be a normal supply and the instances in which the outliers exceed the upper limit as an oversupply of services. We increased the control intensity step by step from 10% to 100% for this oversupply service, to simulate performance under the payment system of DRG. The control effect of the two grouping rules is shown in Table 5. If we took full control, the rules with LOS could save 598,570 US dollars, and 3.35 million US dollars could be saved based on the grouping rules without LOS.

Table 1: Factor assignment and result of single factor analysis of the factors influencing inpatient medical expenditure ($n = 45{,}575$).

Variables	Assignment of influencing factors	Simple size	Expenditure $ (M + IQR)	*t*/*F*	*P* value
Gender	Male = 1	19,488 (42.76%)	1111.82 ± 741.66	247.81	<0.000
	Female = 2	26,087 (57.24%)	1026.12 ± 688.18		
Age	Age ≤ 45 = 1	1,995 (4.37%)	808.56 ± 535.69	788.16	<0.000
	Age between 45 and 60 = 2	9,117 (20.00%)	954.13 ± 640.03		
	Age ≥ 60 = 3	34,463 (75.62%)	1108.76 ± 745.46		
Level of hospital	Grade B secondary hospital = 1	7,609 (16.70%)	784.12 ± 431.26	23,064	<0.000
	Grade A secondary hospital = 2	16,879 (37.04%)	932.23 ± 671.61		
	Grade B tertiary hospital = 3	9,208 (20.20%)	1140.05 ± 825.05		
	Grade A tertiary hospital = 4	11,879 (26.06%)	1597.27 ± 1160.64		
LOS	≤9 d = 1	22,870 (50.18%)	779.17 ± 537.85	16,838	<0.000
	9~13 d = 2	9,544 (20.94%)	1180.41 ± 884.23		
	≥13 d = 3	13,161 (28.88%)	1708.35 ± 1252.09		
Surgery	Yes = 1	2,820	1315.17 ± 765.10	330.79	<0.000
	No = 2	36,987	1013.85 ± 684.21		
	Others = 3	5,768	1212.19 ± 711.22		
Discharge status	Recovery = 1	37,492 (82.26%)	1108.33 ± 750.21	1,413	<0.000
	Transfers = 2	707 (1.55%)	933.52 ± 608.18		
	Death = 3	262 (0.57%)	1525.31 ± 760.29		
	Others = 4	7,954 (15.48%)	842.92 ± 570.62		
	Midway check-out = 5	60 (0.13%)	1112.47 ± 605.51		
Comorbidities complications	The cerebral arteries lack blood supply = 1	20,936 (45.94%)	1031.86 ± 884.13	6285.7	<0.000
	Lacunar infarction = 2	10,111 (22.19%)	1081.60 ± 740.06		
	Cerebral infarction = 3	5,246 (11.51%)	1477.85 ± 1031.87		
	Chronic cerebral ischemia = 4	1,989 (4.36%)	1089.20 ± 763.47		
	Others = 5	7,293 (16%)	1560.89 ± 974.51		
Hypertension level three	Yes = 1	9,303 (20.41)	1162.96 ± 778.03	84.216	
	No = 0	36,272 (79.59%)	1116.37 ± 699.49		

Table 2: Multiple linear stepwise regression results of factors influencing hospitalization expenditure in cerebrovascular disease patients.

Variables	Regression coefficient	Standard deviation	*T*-statistic	*P* value
Gender (take the male as a reference)				
Female	−229.84	44.53	−5.16	<0.000
LOS (take less than nine days as a reference)				
9 days~13 days	2360.58	57.49	40.12	<0.000
More than 13 days	6463.63	59.93	119.86	<0.000
Level of hospital (take grade B secondary hospital as a reference)				
Grade A secondary hospital	1369.27	66.80	20.50	<0.000
Grade B tertiary hospital	3542.44	75.10	47.17	<0.000
Grade A tertiary hospital	6038.54	76.85	78.57	<0.000
Surgery (take having surgery as a reference)				
No surgery	−1480.79	90.11	−16.43	<0.000
Status on discharge (take recovery as a reference)				
Transfers	−669.11	265.52	−2.52	0.012
Death	3035.03	265.08	11.45	<0.000
Others	16.50	64.99	0.25	0.80
Midway check-out	−665.90	819.97	−0.81	0.416
Type of insurance (take urban as a reference)				
Rural	−260.14	46.82	−5.56	<0.000
Main comorbidities complications (take cerebral arteries lack blood supply as a reference)				
Lacunar infarction	−3814.00	67.45	−56.55	<0.000
Cerebral infarction	−3440.60	74.52	−46.17	<0.000
Chronic cerebral ischemia	−1622.16	86.57	−18.74	<0.000
Others	−3925.15	120.92	−32.46	<0.000
Age (take less than 45 as a reference)				
45~60	385.83	115.15	3.35	0.001
More than 60	512.51	108.66	4.72	<0.000

Table 3: Results of first grouping rule (with LOS) and hospitalization expense for cerebrovascular disease patients (US dollars).

Group no.	Grouping rules	*N* (%)	Mean in US dollars	Median ± IQR	CV	Weight	Cost limit	Excess amount *N* (%)
DRG 1	Grade secondary hospital, LOS < 13 d	16,787 (42%)	760	710 ± 501	0.47	0.59	11,020	29 (1.74%)
DRG 2	Grade tertiary hospital, LOS ≤ 9 d	8,344 (21%)	1,125	1004 ± 761	0.60	0.88	15,755	282 (3.38%)
DRG 3	Grade tertiary hospital, LOS 9~13 d	3,741 (9%)	1,620	1503 ± 1195	0.42	1.26	23,700	62 (1.66%)
DRG 4	Grade secondary hospital, LOS ≥ 13 d	5,046 (13%)	1,409	1336 ± 1043	0.36	1.09	21,044	1 (0.02%)
DRG 5	Grade tertiary hospital, CCs with the cerebral arteries lack blood supply, lacunar infarction, chronic cerebral ischemia	2,618 (7%)	2,007	1834 ± 2007	0.42	1.56	29,132	52 (1.99%)
DRG 6	Grade tertiary hospital, CCs with cerebral infarction, principal diagnosis ICD I60, I63, I65, I66, I67, I69	2,007 (5%)	2,957	2602 ± 1923	0.46	2.30	42,263	42 (2.09%)
DRG 7	Grade tertiary hospital, CCs with cerebral infarction, principal diagnosis ICD I61, I62, I64, I68	1,246 (3%)	3,368	3403 ± 2337	0.44	2.86	46,271	21 (1.68%)

Table 4: Results of second grouping rule (without LOS) and hospitalization expense for cerebrovascular disease patients (US dollars).

Group no.	Grouping rules	*N* (%)	Mean in US dollars	Median ± IQR	CV	Weight	Cost limit	Excess amount *N* (%)
DRG 1	Grade secondary hospital, CCs with the cerebral arteries lack blood supply, lacunar infarction, chronic cerebral ischemia	17,490 (44%)	838	773 ± 535	0.48	0.65	8,640	2,015 (11.52%)
DRG 2	Grade secondary hospital, CCs with cerebral infarction	4,343 (11%)	1,096	1001 ± 765	0.49	0.85	12,357	289 (6.65%)
DRG 3	Grade B tertiary hospital, CCs with the cerebral arteries lack blood supply, lacunar infarction, chronic cerebral ischemia	5,031 (13%)	1,203	1106 ± 735	0.51	0.94	11,852	687 (13.65%)
DRG 4	Grade A tertiary hospital, CCs with the cerebral arteries lack blood supply, lacunar infarction, chronic cerebral ischemia	5,660 (14%)	1,548	1405 ± 1063	0.47	1.21	17,142	380 (6.67%)
DRG 5	Grade B tertiary hospital, CCs with cerebral arteries supply blood, lacunar infarction, chronic ischemic cerebral, main diagnosis ICD I60, I63, I65, I66, I67, I69	2,132 (5%)	1,789	1483 ± 1048	0.63	1.39	16,980	325 (15.52%)
DRG 6	Grade A tertiary hospital, CCs with cerebral infarction, main diagnosis ICD I60, I63, I65, I66, I67, I69	2,710 (7%)	2,360	1953 ± 1187	0.67	1.84	19,147	316 (11.66%)
DRG 7	Grade B tertiary hospital, CCs with cerebral infarction, main diagnosis ICD I61, I62, I64, I68	1,172 (3%)	2,363	1983 ± 1426	0.56	1.84	22,997	430 (36.69%)
DRG 8	Grade A tertiary hospital, CCs with cerebral infarction, main diagnosis ICD I61, I62, I64, I68	1,251 (3%)	3,151	2884 ± 1746	0.57	2.45	28,160	329 (26.29%)

4. Discussion

Cerebrovascular disease has been the leading cause of death in China since 2017. The total cost of treating cerebrovascular diseases in China reached 540.6 billion yuan, ranking first among all kinds of diseases and accounting for 17% of the total cost of treating diseases, equivalent to 0.66% of GDP. Our study showed that the total cost of cerebrovascular disease was 51.17 million US dollars in an underdeveloped city in western China during 2018. The average medical expenditure on cerebrovascular disease patients was 1,284 US dollars, of which the government paid 43.42 million US dollars and patients paid 7.75 million US dollars, accounting for 84.8% and 15.2%, respectively. The government therefore paid an average of 1,087 US dollars for each patient, and each patient paid 196 US dollars for themselves. The expenditure in developed cities was even higher. Control of the medical expenses caused by cerebrovascular disease is an urgent problem for the Chinese government.

The city we chose uses a Fee for Service system, which may provide an incentive to oversupply services. We used local data to classify the patients into different groups with similar medical costs. Two models with different rules were

Table 5: The payment situation under different control intensity (US dollars).

			FFS	DRG control intensity									
				10%	20%	30%	40%	50%	60%	70%	80%	90%	100%
Rules with LOS	Society	Total	51.17M	51.03M	50.97M	50.91M	50.86M	50.80M	50.73M	50.67M	50.61M	50.55M	50.40M
		Average	1,284	1,282	1,281	1,279	1,278	1,276	1,275	1,273	1,272	1,270	1,263
	Hospital	Total	0	−0.06M	−0.12M	−0.18M	−0.24M	−0.30M	−0.33M	−0.42M	−0.48M	−0.54M	−0.60M
		Average	0	−1.55	−2.95	−4.50	−6.05	−7.44	−9.00	−10.54	−11.94	−13.49	−15.04
Rules without LOS	Society	Total	51.17M	50.71M	50.40M	50.09M	49.78M	49.31M	49.00M	48.69M	48.38M	48.07M	47.76M
		Average	1,284	1,275	1,267	1,258	1,250	1,242	1,233	1,225	1,216	1,208	1,199
	Hospital	Total	0	−0.33M	−0.67M	−1.01M	−1.34M	−1.68M	−2.01M	−2.35M	−2.68M	−3.01M	−3.35M
		Average	0	−8.37	−16.75	−25.27	−33.65	−42.18	−50.55	−58.93	−67.30	−75.83	−84.20

M: millions.

built, based on whether the LOS was included as a classification variable. We used the CV to measure the quality of the grouping and analyzed the characteristics of the outliers in each group. We then increased the intensity of control of the oversupply of services step by step, from 10% to 100%, to simulate the performance based on the two grouping rules. The model incorporating LOS had a smaller CV than the model without LOS. If our standard model was built without LOS, it could reduce the occurrence of medical oversupply, saving 3.35 million US dollars in one year. These figures apply to only one city; if the whole country controlled costs in this way, the economic pressures on healthcare could quickly be alleviated.

Although it is generally recognized that LOS is the main factor influencing medical expenses [23], the inclusion of LOS as a classification variable of DRG is inconsistent. It is generally believed that considering LOS as a classification variable may lead to upcoding [11]. Most European countries, including England, Estonia, and Finland, do not consider LOS as a classification variable. The official Chinese CHD-DRG, modelled on the American MS-DRG, does not include LOS [14], and the Shanghai-DRG, based on the Australia AR-DRG, also does not consider LOS. However, some studies indicate that omitting LOS may increase the frequency of readmission and moves between hospitals, with services provided in alternative ways [17]. Omitting LOS also leads to poorer care for patients who should have a longer stay. The grouping rules of some countries, such as France, Ireland, and Poland, consider LOS to be an important factor [13].

Tables 3 and 4 show the results of grouping. The grouping rule with LOS has a smaller CV, indicating that the cost difference within grouping rule one was smaller, and the grouping was more reasonable. We used the P75 + 1.5 IQR as the upper limit to test for outliers in each group. The proportion of outliers was higher in the group without LOS. This observation implies that the use of LOS can lead to accurate grouping. Both grouping rules demonstrate that the hospital level is very important. In grouping rules without LOS, hospital levels and comorbidity are more finely divided. It is therefore counterproductive to consider only one hospital level.

We analyzed the outliers (Tables 3 and 4) and found that in the LOS group, the age of the outliers was significantly higher than the average value of the group, while in the group without LOS, the LOS was significantly higher than the average. A study using MS-DRG hospital data from Malta also found that most of the outliers were older and higher costs were associated with higher LOS [8]. Further analysis of these results could help identify the reasons for the high costs.

In Asia, only the Republic of Korea considers the type of hospital as a factor for DRG-based payment [9]. In this study, we found that the level of the hospital crucially influenced inpatient medical expenditure. Although there have been studies looking at the impact of hospital levels on costs [19], research into DRG has tended to focus only on tertiary hospitals. Our research therefore complements previous studies that only grouped hospitals at one level [13].

The major diagnosis was directly related to the differences in the cost of hospitalization. Comorbid patients often require special treatment and care, and different comorbidities may affect the cost of additional care, making comorbid diseases an important grouping variable. Medical costs are higher for the elderly, who require special treatments [13], but age did not show up in our grouping variables. In China, many DRG subgroups, such as the pneumonia subgroup, have age as the primary factor [19], possibly because the high cost of this group is mainly concentrated in the elderly and children. However, the age distribution of cerebrovascular disease is mainly concentrated in the elderly. In most of the European countries, like England and Estonia, age is not a factor used in grouping [13]. This observation is consistent with our findings. Most grouping rules have found surgery to be an important variable, and our single analysis also showed that surgery has a significant impact on costs. But surgery was not a variable identified in our results. This situation may have something to do with the choice of disease species. A cluster study in Beijing, China, also confirmed that in stroke, one of the cerebrovascular diseases, surgery is rare [24].

Table 5 shows the performance if the oversupply of services is controlled under the payment system of DRG. The intensity of control was increased step by step from 10% to 100%, and the results of application of the two rule sets were compared. More money could be saved without the LOS. Experience in Europe indicates that use of LOS leads to upcoding, and the medical cost was high when considering the LOS. These results imply that without LOS the cost could be controlled better, but with LOS the patients could be classified better. More incentives and oversight are needed if DRG is to be introduced. For one city, 21 million RMB could be saved by applying the results of our research, an outcome which is highly desirable for the government.

There were some limitations in this study. Due to the lack of standards for the data reported by the hospitals, there were 5,768 cases lacking information on whether surgery was performed, so these data were excluded from the grouping. Since there is no uniform surgical code between each hospital, we could not use the surgical code as our research object. Due to the large amount of data, we only considered data from one year. In the future, data from more years could be included, or the data from another year could be used for the CV of the test group.

5. Conclusions

We used real data from less developed regions for grouping for the DRG, filling the gap in previous studies, which took developed regions as research objects. To the best of our knowledge, this is the first time that secondary grade hospitals have been considered in a Chinese DRG study. We compared two grouping methods and discussed the results of the grouping. DRG payments were fixed, and this study adjusted the payment ratio of medical insurance, patients, and hospitals to achieve a satisfactory result for all three parties. To speed the development of DRG and rationalize the costs of cerebrovascular disease, the structure of hospital information and the standardization of data entry are essential. More research in this area is urgently needed.

Data Availability

All the data were taken from the Medical Insurance Laboratory of ChengDu Healthcare Security Administration.

Ethical Approval

The study does not involve human subjects and adheres to all current laws of China.

Conflicts of Interest

The authors report no conflicts of interest concerning the materials or methods used in this study or the findings presented in this paper.

Acknowledgments

This project was sponsored by the National Natural Science Foundation of China (nos. 72001153 and 72102159) and the Fundamental Research Funds for the Central Universities (Grant no. SXYPY202135).

References

[1] C. J. Murray, T. Vos, R. Lozano et al., "Disability-adjusted life years (DALYs) for 291 diseases and injuries in 21 regions, 1990-2010: a systematic analysis for the Global Burden of Disease Study 2010," *Lancet*, vol. 380, pp. 2197–2223, 2012.

[2] R. Lozano, M. Naghavi, K. Foreman et al., "Global and regional mortality from 235 causes of death for 20 age groups in 1990 and 2010: a systematic analysis for the global burden of disease study 2010," *Lancet*, vol. 380, pp. 2095–2128, 2012.

[3] S. Zhang, W.-B. He, and N.-H. Chen, "Causes of death among persons who survive an acute ischemic stroke," *Current Neurology and Neuroscience Reports*, vol. 14, no. 8, p. 467, 2014.

[4] M. Zhou, H. Wang, X. Zeng et al., "Mortality, morbidity, and risk factors in China and its provinces, 1990-2017: a systematic analysis for the global burden of disease study 2017," *The Lancet*, vol. 394, no. 10204, pp. 1145–1158, 2019.

[5] China cardiovascular health and disease report, "Summary of China cardiovascular health and disease report 2019," *Chinese Journal of Circulation*, vol. 35, pp. 833–854, 2020.

[6] Y. Zhang, P. Chai, T. Zhai, and Q. Wang, "Accounting and analysis of the treatment cost of cardiovascular and cerebrovascular diseases in China in 2017," *Chinese Journal of Recycling*, vol. 35, no. 9, pp. 859–865, 2020.

[7] R. Busse, A. Geissler, A. Aaviksoo et al., "Diagnosis related groups in Europe: moving towards transparency, efficiency, and quality in hospitals?" *BMJ*, vol. 346, p. f3197, 2013.

[8] C. Camilleri, M. Jofre-Bonet, and V. Serra-Sastre, "The suitability of a DRG casemix system in the Maltese hospital setting," *Health Policy*, vol. 122, no. 11, pp. 1183–1189, 2018.

[9] P. L. Annear, S. Kwon, L. Lorenzoni et al., "Pathways to DRG-based hospital payment systems in Japan, Korea, and Thailand," *Health Policy*, vol. 122, 2018.

[10] S. W. Wu, Q. Pan, T. Chen et al., "Research of medical expenditure among inpatients with unstable angina pectoris in a single center," *Chinese Journal of Medical Sciences*, vol. 130, no. 13, pp. 1529–1533, 2017.

[11] V. Bystrov, A. Staszewska-Bystrova, D. Rutkowski, and T. Hermanowski, "Effects of DRG-based hospital payment in Poland on treatment of patients with stroke," *Health Policy*, vol. 119, no. 8, pp. 1119–1125, 2015.

[12] Y. Yan, Z. Xie, Y. Luo, H. Liu, A. Liu, and S. Zhu, "Grouping study of DRGs in stroke patients in Beijing," *Chinese Health Statistics*, vol. 25, no. 4, pp. 347–350, 2008.

[13] M. Peltola and W. Quentin, "Diagnosis-related groups for stroke in Europe: patient classification and hospital reimbursement in 11 countries," *Cerebrovascular Diseases*, vol. 35, no. 2, pp. 113–123, 2013.

[14] http://www.nhsa.gov.cn/art/2019/10/24/art_37_1878.html.

[15] K. Zou, H. Y. Li, D. Zhou, and D. Liao, "The effects of diagnosis-related groups payment on hospital healthcare in China: a systematic review," *BMC Health Services Research*, vol. 20, 2020.

[16] D. Li, "The unbalance of economic and social development in China: analysis and correction," *Journal of Sichuan University (Engineering Science Edition)*, vol. 2004, no. 1, pp. 27–32, 2004.

[17] A. Geissler, D. Scheller-Kreinsen, W. Quentin, and E. Group, "Do diagnosis-related groups appropriately explain variations in costs and length of stay of hip replacement? a comparative assessment of drg systems across 10 European countries," *Health Economics*, vol. 21, no. S2, pp. 103–115, 2012.

[18] T. Grubinger, C. Kobel, and K.-P. Pfeiffer, "Regression tree construction by bootstrap: model search for DRG-systems applied to Austrian health-data," *BMC Medical Informatics and Decision Making*, vol. 10, no. 1, p. 9, 2010.

[19] W. Xue-zhi, "DRGs case combination study on hospitalization expenses of pneumonia patients," *China Health Statistics*, vol. 37, no. 2, pp. 235–238, 2020.

[20] C. Martin, K. Odell, J. C. Cappelleri, T. Bancroft, R. Halpern, and A. Sadosky, "Impact of a novel cost-saving pharmacy

program on pregabalin use and health care costs," *Journal of Managed Care & Specialty Pharmacy*, vol. 22, no. 2, pp. 132–144, 2016.

[21] R Core Team, *R: A Language and Environment for Statistical Computing*, R Foundation for Statistical Computing, Vienna, Austria, 2013, URL http://www.R-project.org/.

[22] G. V. Kass, "An exploratory technique for investigating large quantities of categorical data," *Applied Statistics*, vol. 29, no. 2, pp. 119–127, 1980.

[23] J. M. Kahn, G. D. Rubenfeld, J. Rohrbach, and B. D. Fuchs, "Cost savings attributable to reductions in intensive care unit length of stay for mechanically ventilated patients," *Medical Care*, vol. 46, no. 12, pp. 1226–1233, 2008.

[24] Y. Yan, Z. Xie, and Y. Luo, "Structural, elastic and electronic properties of ReO2," *Chinese Journal of Health Statistics*, vol. 25, no. 4, pp. 347–350, 2008.

Hindawi
Journal of Healthcare Engineering
Volume 2021, Article ID 5578089, 10 pages
https://doi.org/10.1155/2021/5578089

Research Article

Identification of the Vas Deferens in Laparoscopic Inguinal Hernia Repair Surgery Using the Convolutional Neural Network

Peng Cui,[1] **Song Zhao**,[2] **and Wenxi Chen**[1]

[1] *Biomedical Information Engineering Lab, The University of Aizu, Aizuwakamatsu, Japan*
[2] *Department of General Surgery, People's Hospital of Hannan District, Hannan District, Wuhan, China*

Correspondence should be addressed to Wenxi Chen; wenxi@u-aizu.ac.jp

Received 28 January 2021; Revised 25 August 2021; Accepted 7 September 2021; Published 22 September 2021

Academic Editor: Jialin Peng

Inguinal hernia repair is one of the most frequently conducted surgical procedures worldwide. Laparoscopic inguinal hernia repair is considered to be technically challenging. Artificial intelligence technology has made significant progress in medical imaging, but its application in laparoscopic surgery has not been widely carried out. Our aim is to detect vas deferens images in laparoscopic inguinal hernial repair using the convolutional neural network (CNN) and help surgeons to identify the vas deferens in time. We collected surgery videos from 35 patients with inguinal hernia who underwent laparoscopic hernia repair. We classified and labeled the images of the vas deferens and used the CNN to learn the image features. Totally, 2,600 images (26 patients) were labeled for training and validating the neural network and 1,200 images (6 patients) and 6 short video clips (3 patients) for testing. We adjusted the model parameters and tested the performance of the model under different confidence levels and IoU and used the chi-square to analyze the statistical difference in the video test dataset. We evaluated the model performance by calculating the true positive rate (TPR), true negative rate (TNR), accuracy (ACC), positive predictive value (PPV), and *F*1-score at different confidence levels of 0.1 to 0.9. In confidence level 0.4, the results were TPR 90.61%, TNR 98.67%, PPV 98.57%, ACC 94.61%, and *F*1 94.42%, respectively. The average precision (AP) was 92.38% at IoU 0.3. In the video test dataset, the average values of TPR and TNR were 90.11% and 95.76%, respectively, and there was no significant difference among the patients. The results suggest that the CNN can quickly and accurately identify and label vas deferens images in laparoscopic inguinal hernia repair.

1. Introduction

Inguinal hernias are found in 3%–8% of the general population [1]. Inguinal hernia repair (IHR) is one of the most commonly used surgical methods in general surgery. More than 20 million inguinal or femoral hernias are repaired every year, and 75% of abdominal wall hernias are classified as inguinal hernia [2, 3].

With the development of medical technology and medical equipment, minimally invasive surgery has been performed widely. Transabdominal preperitoneal (TAPP) [4] and total extraperitoneal (TEP) [5] surgeries are the most commonly used minimally invasive methods for inguinal hernia.

However, laparoscopic IHR is considered to be technically challenging. Laparoscopic minimally invasive surgery method can reduce the extent of skin incisions, nerve damage, and hematoma; lower postoperative pain and risk of infection of the surgical site; and lead to quicker recovery [6, 7], but it also has several disadvantages: for example, the surgeon initially needs a longer surgery time before plateauing on his/her learning curve; the surgery has a higher risk of complications; it needs much more knowledge of pelvic anatomy; and it needs a high level of surgical skill [8, 9], which can lead to more mistakes and harm to patients during the learning process. In clinical practice, young surgeons need a long learning curve to carry out laparoscopic repair surgery well, especially TEP technology. With the increase of surgical experience and familiarity with local anatomy, complications and recurrence rates will gradually decrease [10].

Common complications of the operation include bleeding; bladder lesions; intestinal obstructions; intestinal perforations; injury to the iliac vein, femoral nerve, and vas deferens; and even death [8].

With the development of artificial intelligence (AI) technology, CNNs have become an effective method for medical image analysis, disease prediction and diagnosis, and lesion detection and have been widely used [11–13]. CNNs are not a solution to replace doctors, but will help doctors optimize their routine tasks, thus having a potentially positive impact on medical practice [14]. We have also applied deep learning technology to nerve and dura mater recognition under spinal endoscopy and achieved satisfactory results [15].

There are many neural network models for object detection [16–19], but each neural network has its own advantages. Some can detect objects quickly, but the precision is not optimal, while others can detect an object with higher precision, but the speed of detection is quite slow. In this research, we want to use a CNN model that can detect objects fast enough to be used at 30 frames per second and have good precision. YOLO is a one-stage convolutional neural network for object detection [20, 21], whose rate of detection can reach 65 fps with an average precision of up to 43.5% on the COCO dataset [21]. The innovations of this article are as follows: (1) combined with computer CNN technology and clinical data, a new method for identifying vas deferens images in laparoscopic inguinal hernia repair using the CNN model is proposed, and the object detection ability of the CNN model (YOLOv4) on the medical dataset is also tested; (2) different annotation methods are used to train the CNN model and to examine the performance of the model in the process of training and testing; (3) discussed with clinical experts and selected the appropriate IoU value to evaluate the performance of the model for reference by clinical surgeons.

2. Materials and Methods

2.1. Data Collection. This research was approved by the Ethics Committee at Hannan Hospital, Hannan District, Wuhan City, Hubei Province, China.

In this study, 35 adult male patients with inguinal hernia disease admitted to the hospital for laparoscopic surgery from April 2018 to December 2019 were selected. The laparoscopic image device used was KARL STORZ Endoscopy (22202020-110), America. All patients underwent laparoscopic hernia repair and signed a patient consent form. We collected information such as gender, age, disease name, and interoperation videos. All endoscopic surgeries were performed by senior endoscopic experts at Hannan Hospital. Details of the dataset are shown in Table 1, and the patient age distribution is shown in Figure 1.

2.1.1. Inclusion Criteria. We selected those subjects satisfying all of the following three criteria: (i) adult male; (ii) the patient was diagnosed with inguinal hernia and underwent hernia repair for the first time; and (iii) the patient agreed to allow the use of video recordings for scientific research.

Table 1: Dataset and patient information.

Dataset	Number of patients	Age (years)	Number of images
Training	26	63.15 ± 7.64	2,600
Image test	6	64 ± 8.12	1,200
Video test	3	55 ± 6.56	5,433

All patients are male.

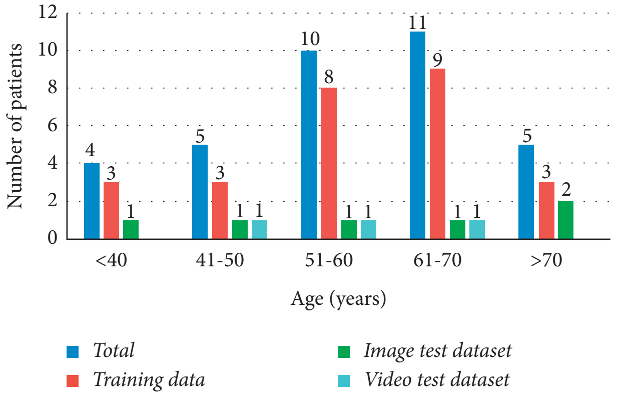


Figure 1: Patient age distribution.

2.1.2. Exclusion Criteria. We excluded subjects matching any of the following three criteria: (i) female patient; (ii) patients with irreducible inguinal hernia; and (iii) the laparoscopic surgery was converted into an open surgery for any reason.

2.2. Data Processing. The patients were randomly divided into three groups, 26 patients in the training dataset, 6 patients in the image test dataset, and 3 patients in the video test dataset. In the training dataset and image test dataset, we used MATLAB (9.6.0.1174912 (R2019a) Update 5, academic use) to decompose the surgical videos into images according to different datasets and then saved them. A laparoscopic expert selected these images manually; then, the other laparoscopic expert verified them and finally deleted the disputed images and reached an agreement. In the video test dataset, we selected two short video clips from each patient's full surgical video, each of which is 30 seconds. All the videos used in this study were 30 frames per second. One of them is a clip with the vas deferens image, and the other is a clip without the vas deferens image. Then, two laparoscopic experts verified these video clips and reached an agreement.

In order to balance the training data, we selected 100 images containing the vas deferens for each patient in the training dataset. In the image test dataset, we chose 200 images for each patient (100 images that included the vas deferens and 100 images without the vas deferens). Thus, a total of 3800 pictures (2600 images in the training dataset and 1200 images in the image test dataset) of the vas deferens and 180 seconds (90 seconds with the vas deferens and 90 seconds without the vas deferens) of video clips were chosen to form an experimental database. There was no overlap in patients and images between the training dataset, image test dataset, and video test dataset.

All the training data were labeled using software LabelImg (v1.8.1) and then validated by two laparoscopic experts; none of the researchers had any objection to the labeling results. The research flowchart is shown in Figure 2. The original images with the vas deferens, labeled images, and original images without the vas deferens are depicted in Figure 3.

2.3. Training Parameters and Computer Configuration. In order to achieve higher accuracy with a faster processing speed, we used a one-stage neural network, YOLOv4 (based on the Darknet framework), to train and test the above datasets.

For training the model, we randomly divided the training dataset (2,600 images) into training data and internal validation data according to the ratio of 9 : 1. The details of neural network training parameters are as follows: input size = 416 ∗ 416, batch = 64, subdivisions = 32, initial learning rate = 0.001, momentum = 0.95, and max-batches = 10000.

Because our study is only a binary classification task, vas deferens tissue will only appear in one area of an image in laparoscopic IHR. Therefore, in the target detection stage, we adjusted the parameters in the process of nonmaximum suppression (NMS) and set NMS-IoU to 0.1 to reduce the number of detection boxes.

The computer was an Intel i9 9900k CPU @3.6 GHz × 16, RAM 32 GB, with a CUDA-enabled Nvidia Titan 312 GB graphics processing unit (Nvidia), based on hardware of the NVIDIA GeForce RTX 2070 SUPER GPU. The whole training time was about 12 hours.

3. Results

We examined the test dataset using the best training weight in the training process, and then two laparoscopic experts verified whether the images in the test dataset were correctly labeled by the model. The two laparoscopic experts verified the labels in all the images and further discussed any labeled images that were controversial. Finally, they reached an agreement on all the labeling results.

We defined those images with the vas deferens that were correctly labeled as true positive (TP); the images without the vas deferens but wrongly identified and labeled incorrectly as false positive (FP); the images containing the vas deferens but not identified and not labeled as false negative (FN); and the images without the vas deferens and without any label as true negative (TN).

In the image test dataset, we used different confidence levels from 0.1 to 0.9 to evaluate the performance of the model. The model was used to label the images in the image test dataset, and two laparoscopic experts examined these results. For 600 positive symbols (images include the vas deferens), a total of 607 detection boxes were labeled on these images. The detailed test results at different confidence levels are shown in Table 2. Example images of TP, FP, and FN are shown in Figure 4.

$$\begin{aligned}
\mathrm{TPR} &= \frac{\mathrm{TP}}{(\mathrm{TP}+\mathrm{FN})},\\
\mathrm{TNR} &= \frac{\mathrm{TN}}{(\mathrm{TN}+\mathrm{FP})},\\
\mathrm{PPV} &= \frac{\mathrm{TP}}{(\mathrm{TP}+\mathrm{FP})},\\
\mathrm{ACC} &= TP + \frac{\mathrm{TN}}{(\mathrm{TP}+\mathrm{TN}+\mathrm{FP}+\mathrm{FN})},\\
F1 &= \frac{2 * \mathrm{PPV} * \mathrm{TPR}}{(\mathrm{PPV}+\mathrm{TPR})}.
\end{aligned} \tag{1}$$

According to the test results in Table 2, we used the following indicators to evaluate the performance of the CNN model: true positive rate (TPR), true negative rate (TNR), accuracy (ACC), positive predictive value (PPV), intersection over union (IoU), average precision (AP), and $F1$-score. We also draw the receiver operating characteristic (ROC) curve and calculate the AUC value as indicators to evaluate the performance of the model. The formulas used to calculate these values are as follows, and the results are shown in Table 3. The ROC curve is shown in Figure 5. The performance of the model for different IoUs is shown in Table 4.

According to the ROC curve, we calculate that the optimal confidence threshold of the model is around confidence level 0.4, and the AUC value is 0.97, so in the process of testing the video test dataset, we use the confidence level of 0.4 to test the real-time detection function of the model. After saving the video detection results, we decompose the video clips into images for verification. A total of 5433 images were decomposed from the video clips (2719 with the vas deferens and 2714 without the vas deferens). Two laparoscopic experts verified these images one by one and confirmed the results. The evaluation indicators and specific results are shown in Table 5.

In the video test dataset, these results were analyzed using IBM SPSS Statistics version 23.0. We used the chi-square test to analyze the statistical differences between patients, with $p > 0.05$, meaning no significant difference between the tested objects. The p value of TPR was 0.768, the p value of TNR was 0.608, and all p values were greater than 0.05; the average TPR, TNR, PPV, and ACC were 90.11%, 95.76%, 95.52%, and 92.93%, respectively. The results show that the model can effectively identify the vas deferens images in laparoscopic inguinal hernia repair, and the sensitivity and specificity of the model to different patients in the video test dataset are not statistically different.

We also tested the real-time detection ability of the model. The results show that when the input size is 416 ∗ 416, the detection speed of the model can reach 45.32 frames per second, and the detection speed is greater than 30 frames per second, which meets the real-time detection requirements of laparoscopic inguinal hernia repair.

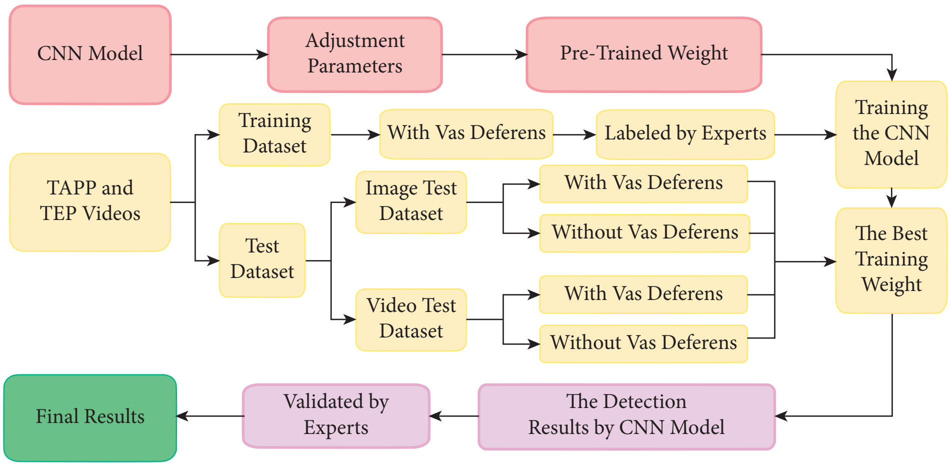


Figure 2: Research flowchart.

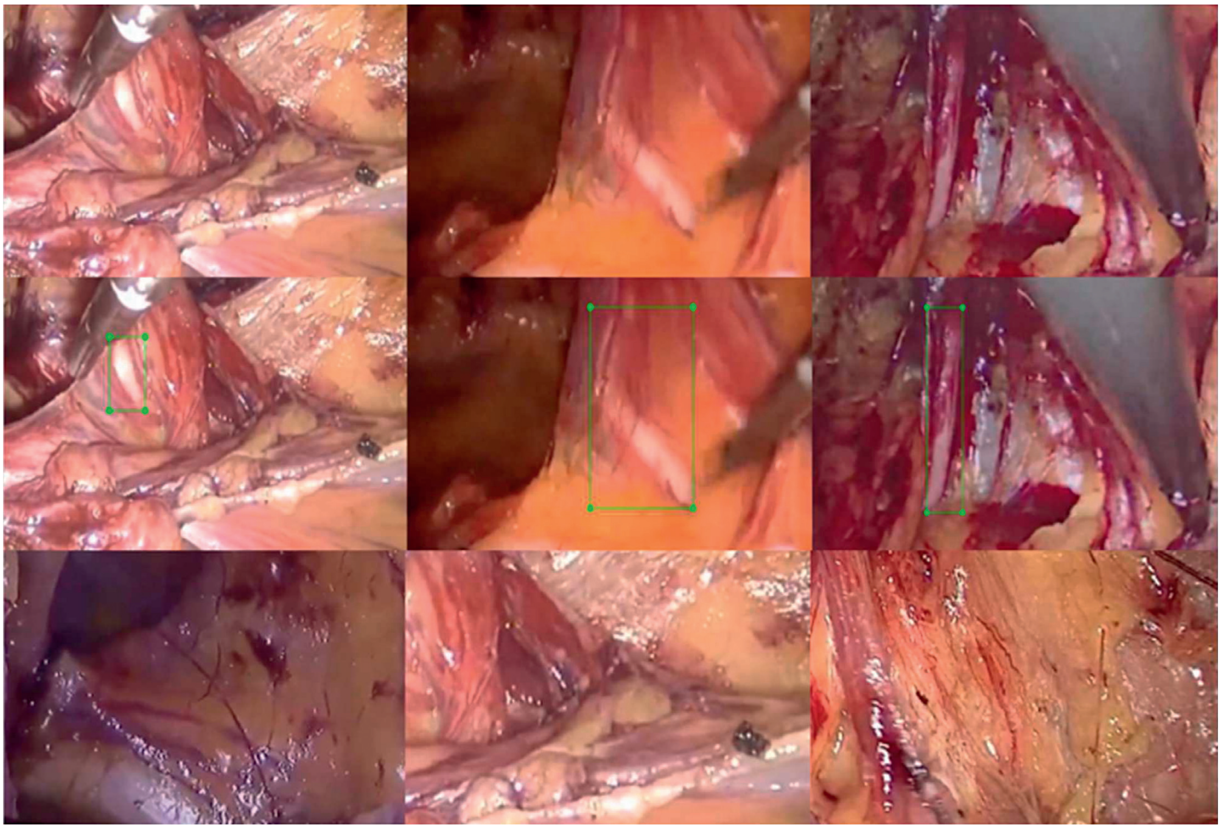


Figure 3: Original images. The first row shows original images with the vas deferens; the second row shows labels included by the experts; the third row shows original images without the vas deferens.

4. Discussion

Using CNNs for medical image detection is not new, but using this technology to detect the vas deferens under laparoscopic IHR surgery is novel; we use the CNN model to train and test the vas deferens images and obtained good results.

4.1. Performance of the CNN-Based Identification. In order to select the appropriate parameters and indicators to evaluate the performance of the model, we set different confidence levels to calculate the evaluation indicator. Laparoscopic experts verified the labeled images one by one and concurred on the final results.

Table 2: The test result at different confidence levels.

Confidence level	TP	FP	TN	FN
≥0.1	577	56	544	30
≥0.2	565	32	568	42
≥0.3	555	17	583	52
≥0.4	550	8	592	57
≥0.5	545	4	596	62
≥0.6	531	2	598	76
≥0.7	516	1	599	91
≥0.8	494	0	600	113
≥0.9	461	0	600	146

These results are based on all the detection boxes.

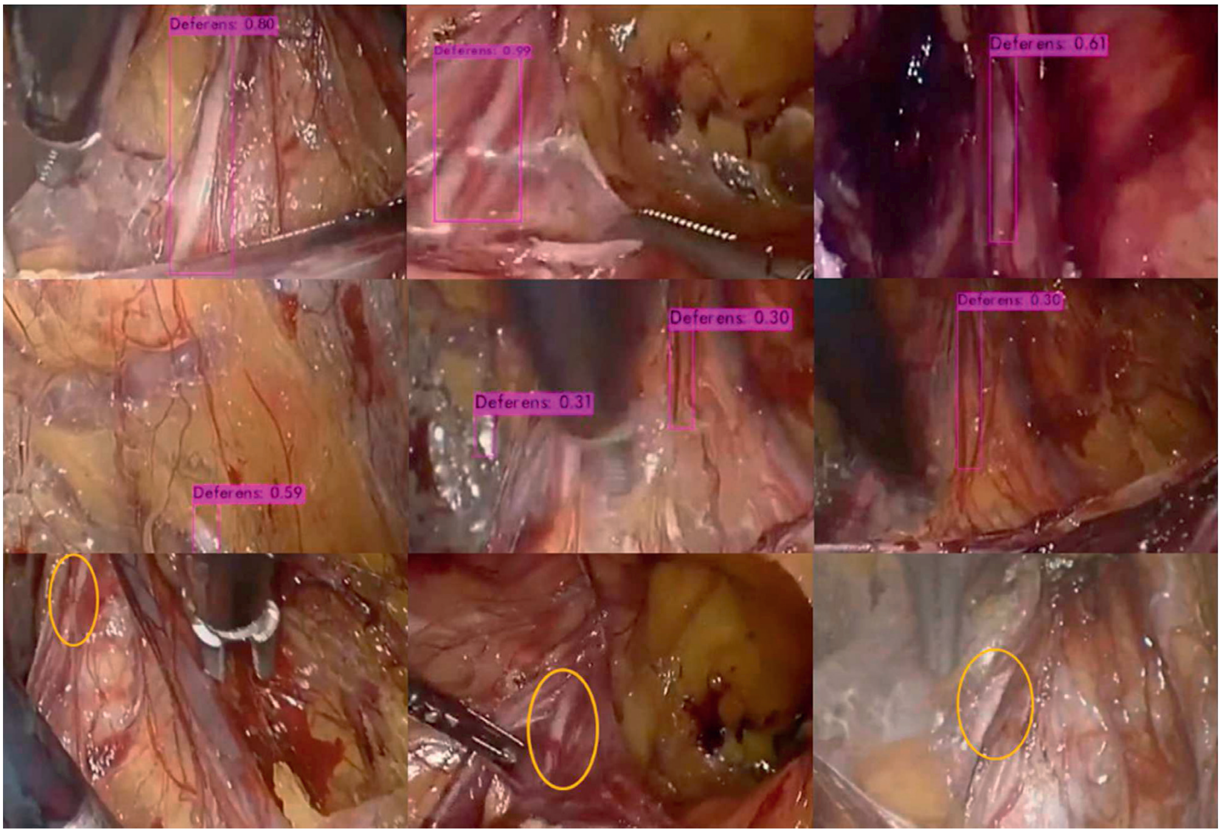


Figure 4: CNN-labeled vas deferens area and corresponding confidence level. The first row shows the true positive (TP); the vas deferens is labeled with a purple rectangle, and the confidence level is shown above the rectangle. The second line shows the false positive (FP). The third line shows the false negative (FN). The model does not give any labels on the image. The researchers circled the area of the vas deferens in yellow.

According to Table 3, it is clear that, with increasing confidence levels, TPR gradually decreased from 95.06% to 75.95%, while TNR and PPV gradually increased from 90.67% and 91.15%, respectively, to 100%. In order to observe the performance of the model more comprehensively, we also calculated the ACC and *F*1-score, both of which first increased and then decreased with increasing confidence levels.

The *F*1-score is often used as a comprehensive index to judge the performance of a CNN model; it is a combination of TPR and PPV. When the confidence level is 0.2, 0.3, and 0.4, the *F*1-score of the CNN model is higher than 94%. The ROC curve can show the influence of different thresholds on the generalization performance of the model, which is helpful to select the best threshold [22, 23]. We analyzed the ROC curve, compared the *F*1-score, and discussed with laparoscopy experts. Finally, we thought that when the confidence level was 0.4, the comprehensive performance of the model was more suitable for the dataset.

4.2. Medical Image Labeling Method. Using the correct labeling method is also an important aspect of the CNN model's target detection to achieve good results. Due to the

Table 3: The evaluation indicators at different confidence levels.

Confidence level	TPR (%)	TNR (%)	PPV (%)	ACC (%)	*F*1 (%)
≥0.1	95.06	90.67	91.15	92.87	93.06
≥0.2	93.08	94.67	94.48	93.87	93.85
≥0.3	91.43	97.17	97.03	94.28	94.16
≥0.4	90.61	98.67	98.57	94.61	94.42
≥0.5	89.79	99.27	99.27	94.53	94.29
≥0.6	85.01	99.62	99.62	93.54	93.16
≥0.7	85.01	99.81	99.81	92.38	91.82
≥0.8	81.38	100	100	90.64	89.73
≥0.9	75.95	100	100	87.9	86.33

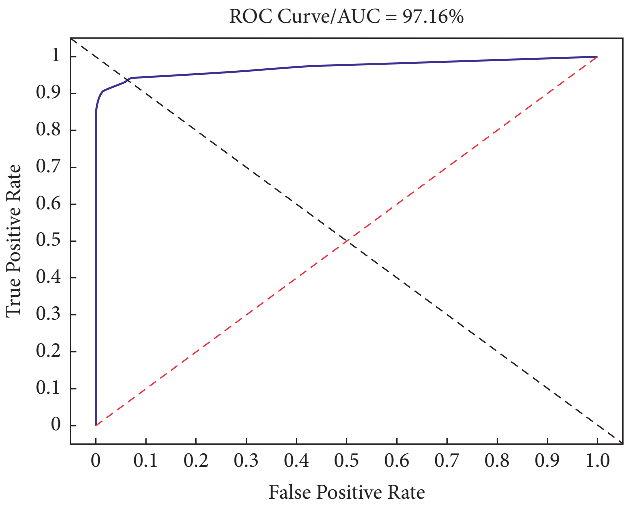


Figure 5: The ROC curve and AUC value.

Table 4: The evaluation indicators under different IoUs.

IoU	TPR (%)	PPV (%)	AP (%)	*F*1 (%)
≥0.2	88.73	99.82	95.64	93.95
≥0.3	87.75	98.71	92.38	92.91
≥0.4	85.95	96.69	89.31	91.00
≥0.5	80.72	90.81	80.88	85.47
≥0.6	72.22	81.25	68.73	76.47
≥0.7	52.29	58.82	36.97	55.36

These evaluation indicators are calculated based on 600 positive data items in the test dataset.

Table 5: The test results in the video test dataset.

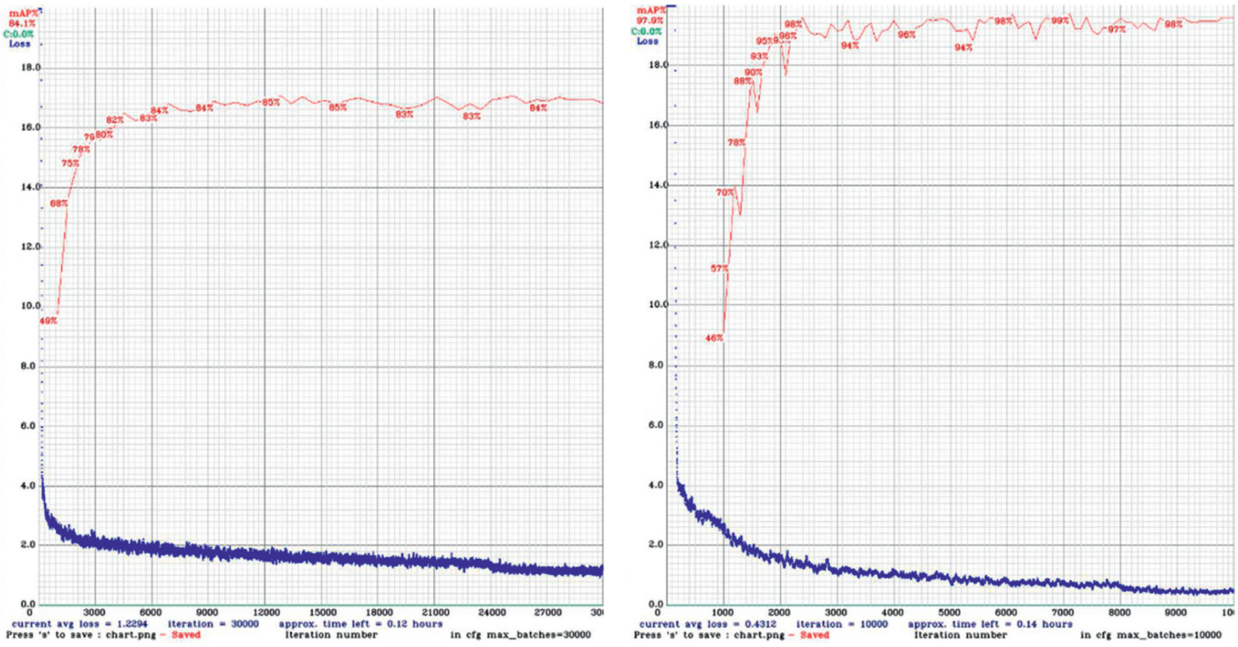
Patients	TPR (%)	TNR (%)	PPV (%)	ACC (%)
1	90.23 (822/911)	96.14 (872/907)	95.92	93.18
2	89.50 (810/905)	95.23 (859/902)	94.96	92.36
3	90.58 (818/903)	95.91 (868/905)	95.67	93.25
Average	90.11 (2450/2719)	95.76 (2599/2714)	95.52	92.93

These evaluation indicators are calculated based on confidence level 0.4.

complex environment of the endoscopic surgery, target detection in the endoscopic surgery is also a new challenge. There is currently no clear method for labeling the target tissue in the endoscopic surgery.

At the beginning of this study, when we labeled the vas deferens on the surgical images, because the features of the target do not take on a regular shape, in order to lessen the proportion of nonvas deferens tissue in the label box and reduce the influence of nonvas deferens tissue on the target tissue, we only labeled the area with obvious vas deferens features on the image. However, the results show that the model is unable to obtain all information pertaining to the vas deferens in these images, which leads to unsatisfactory training and testing results.

We adjusted the labeling method and expanded the scope of the label box so that the vas deferens tissue in the image could be included in the label box as much as possible. Although the proportion of the nonvas deferens in the label box increases, when we use the same training image and validation image and the same training parameters to train and test the model again, the training process and results show that the training effect of the new labeling method is



(a)

(b)

Figure 6: Processes of training and validation using different labeling methods. (a) The graph showing the training process of the partial labeling method. (b) The graph showing the training process of the new labeling method. The red line shows the internal validation precision, and the blue line shows the loss curve.

better than our previous labeling method. The example after adjusting the labeling method is shown in line 2 of Figure 3, and the data details of the training process are shown in Figure 6.

Through the observation of the model training process, it is not difficult to find that different annotation methods will significantly affect the training effect of the model. During surgery, although the target tissue may be partially occluded by other nontarget tissues, our suggestion is to label the target tissue as completely as possible.

4.3. Indicator IoU and AP. IoU is the ratio of the overlap area and the union area of the label box and the test box. In the field of medical image recognition, the higher the IoU is, the higher the positioning accuracy is and the more it meets the needs of clinicians.

We tested the performance of the model under different IoU thresholds (0.7 to 0.2). The results showed that AP increased from 36.97% to 95.64%, and PPV increased from 58.82% to 99.82%.

We discussed with laparoscopic experts whether these labeled areas are enough to remind surgeons to identify the vas deferens. Experts believe that, in laparoscopic surgery, the role of computer-aided surgery is to remind surgeons to readily discover target tissue and to pay more attention to this target area. When surgeons know the general location of the vas deferens, they will be more alert and cautious, so as not to damage the vas deferens and other target tissues during surgery.

By comparing the images in the test results, experts believed that when the IoU was greater than 0.3, the labeling result was acceptable. However, when the IoU dropped to 0.2, some labels were inaccurate, the proportion of target tissue in the label box was too low to correctly represent the vas deferens, and the center of the label box was not located over the vas deferens. These results also tell us that IoU is also an important indicator to evaluate the performance of the model, and higher IoU will better assist doctors to observe and discover target organizations.

We found that when the IoU threshold was greater than 0.2, the TPR was lower than 90%, which indicated that the label box given by the model was not accurate enough, and the learning of the vas deferens was not enough. Although the model could recognize obvious features of the vas deferens, if the vas deferens' surface was partially obscured, the model could not accurately recognize and label the vas deferens. The details are shown in Figure 7.

4.4. Limitations. We used the CNN model to identify the vas deferens in laparoscopic inguinal hernia repair for the first time. Although we achieved satisfactory results, our research also has some limitations. First, we should further expand the number of patients and the absolute number of vas deferens images as the training dataset so that the CNN model can fully learn the characteristics of the vas deferens. Second, more indicators and test data should be set to evaluate the performance of the model, and we need to compare the performance with surgeons in different levels, so as to make the evaluation of the model more objective. Third, we only use the data of one hospital to train and verify the model, and the multicenter data will more effectively

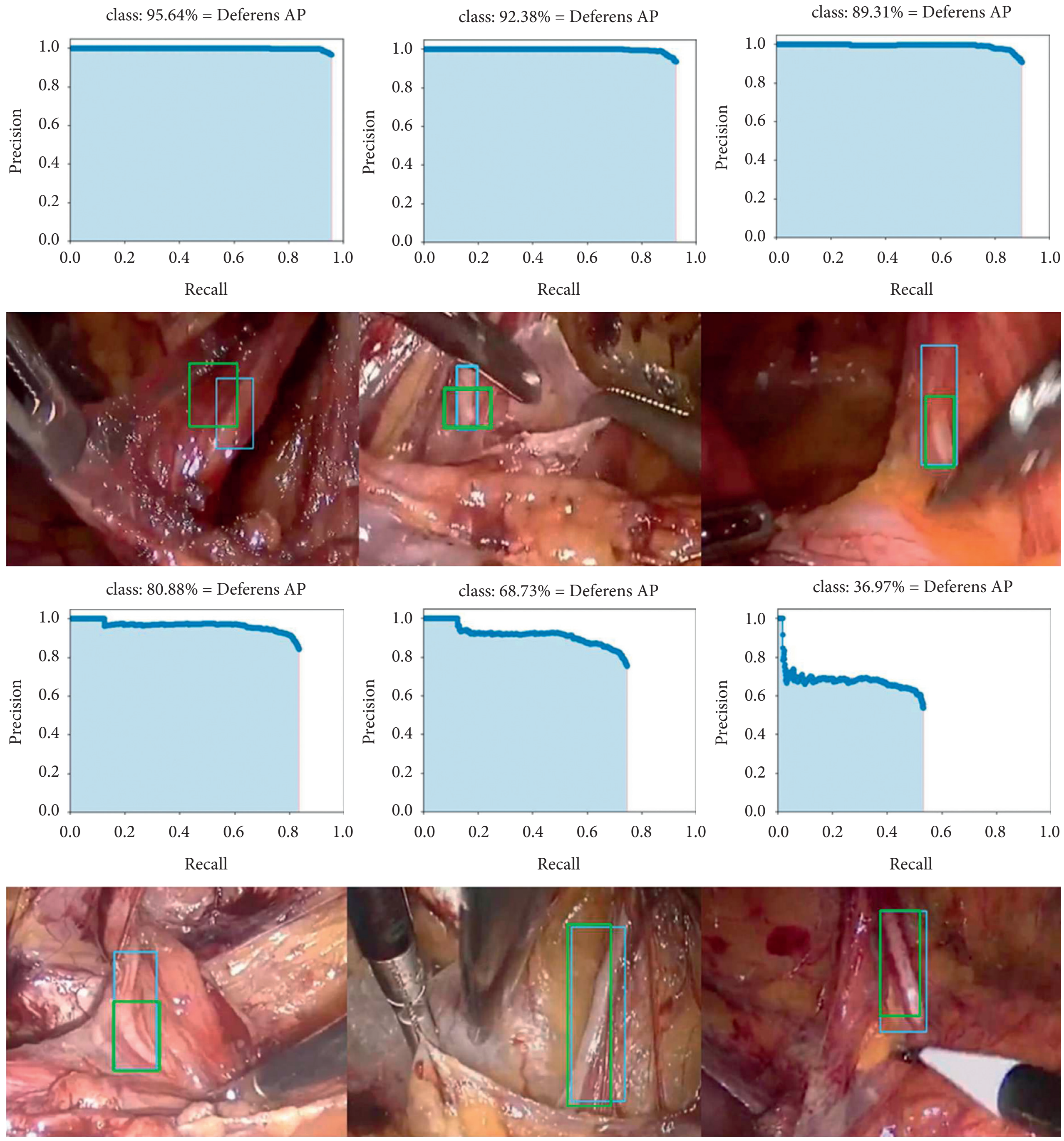


Figure 7: The AP of the model under different IoU values and the corresponding image examples. In the first row and the third row, from left to right, AP curves at IoU 0.2 to 0.4 and 0.5 to 0.7 are shown. In the second row and the fourth row, the corresponding images at different IoU values are shown. The blue rectangle boxes were labeled by laparoscopic experts in advance, and the green rectangle boxes were labeled by the model.

prove the detection and generalization ability of the model. Fourth, this model has a large number of parameters. Although the detection performance is satisfactory, it requires high computer configuration and time-consuming training process. We should further optimize the model structure, reduce the model parameters, and improve the model detection ability.

In future studies, we plan to collect more data from more hospitals, compare the neural network with the indicators of identifying important tissues in laparoscopic inguinal hernia repair by surgeons with different levels of experience, and further test whether this technology can help young general surgeons optimize the learning curve and reduce the incidence of vas deferens injury complications.

5. Conclusion

As an effective target detection method, computer deep learning technology has been widely used in medical image recognition [24–27]. In this study, we used YOLO (v4) to identify vas deferens images under laparoscopic inguinal hernia repair. We used different confidence levels from 0.1 to 0.9 to calculate various evaluation indicators in the image test dataset; picked the best confidence level for the video test dataset; adjusted the IoU thresholds from 0.2 to 0.7 to understand the positioning accuracy and AP of the model; and discussed with laparoscopic experts to select appropriate parameters to evaluate the performance of the model. In the image test dataset, the values of TPR, TNR, PPV, ACC, and $F1$ were 90.61%, 98.67%, 98.57%, 94.61%, and 94.42% (confidence level 0.4), respectively. In the video test dataset, the values of TPR, TNR, PPV, and ACC were 90.11%, 95.76%, 95.52%, and 92.93%, respectively. In IoU 0.3, the average precision (AP) was 92.38%.

We confirmed that a CNN can identify and label vas deferens images efficiently in laparoscopic inguinal hernia repair. This will help laparoscopic surgeons, especially young ones, to better carry out clinical work, optimize the learning curve of the laparoscopic surgery, improve surgical efficiency, and reduce surgical complications.

Data Availability

The image and video data used to support the findings of this study are restricted by the Ethics Committee at Hannan Hospital in order to protect patient privacy.

Consent

All participants gave their informed consent. All approaches were performed in accordance with the regulations and relevant guidelines.

Conflicts of Interest

The authors declare that they have no conflicts of interest.

Authors' Contributions

PC, SZ, and WXC conceptualized and designed the study and critically revised the article. SZ and PC contributed to the acquisition of the data and analyzed and interpreted the data. PC drafted the article. WXC and PC reviewed the submitted version of the manuscript and performed statistical analysis. WXC approved the final version of the manuscript on behalf of all authors, provided administrative/technical/material support, and supervised the study.

Acknowledgments

This study was supported in part by the Competitive Research Fund (2020-P-24) of the University of Aizu.

References

[1] A. N. Kingsnorth, M. R. Gray, and D. M. Nott, "Prospective randomized trial comparing the shouldice technique and plication darn for inguinal hernia," *British Journal of Surgery*, vol. 79, no. 10, pp. 1068–1070, 1992.

[2] P. Schumpelick, K. H. Treutner, and G. Arlt, "Inguinal hernia repair in adults," *The Lancet*, vol. 344, no. 8919, pp. 375–379, 1994.

[3] A. Kingsnorth and K. LeBlanc, "Hernias: inguinal and incisional," *The Lancet*, vol. 362, no. 9395, pp. 1561–1571, 2003.

[4] S. A. Kapiris, W. A. Brough, C. M. S. Royston, C. O'Boyle, and P. C. Sedman, "Laparoscopic transabdominal preperitoneal (tapp) hernia repair," *Surgical Endoscopy*, vol. 15, no. 9, pp. 972–975, 2001.

[5] C. Tamme, H. Scheidbach, C. Hampe, C. Schneider, and F. Köckerling, "Totally extraperitoneal endoscopic inguinal hernia repair (TEP)," *Surgical Endoscopy*, vol. 17, no. 2, pp. 190–195, 2003.

[6] R. Bittner and J. Schwarz, "Inguinal hernia repair: current surgical techniques," *Langenbeck's Archives of Surgery*, vol. 397, no. 2, pp. 271–282, 2012.

[7] F. Köckerling, D. Jacob, W. Wiegank et al., "Endoscopic repair of primary versus recurrent male unilateral inguinal hernias: are there differences in the outcome?" *Surgical Endoscopy*, vol. 30, no. 3, pp. 1146–1155, 2016.

[8] A. Meyer, P. Blanc, J. G. Balique et al., "Laparoscopic totally extraperitoneal inguinal hernia repair. twenty-seven serious complications after 4565 consecutive operations," *Revista do Colégio Brasileiro de Cirurgiões*, vol. 40, no. 1, pp. 32–36, 2013.

[9] A. Meyer, P. Blanc, R. Kassir, and J. Atger, "Laparoscopic hernia: umbilical-pubis length versus technical difficulty," *Journal of the Society of Laparoendoscopic Surgeons: Journal of the Society of Laparoendoscopic Surgeons*, vol. 18, no. 3, 2014.

[10] F. Y. Suguita, F. F. Essu, L. T. Oliveira et al., "Learning curve takes 65 repetitions of totally extraperitoneal laparoscopy on inguinal hernias for reduction of operating time and complications," *Surgical Endoscopy*, vol. 31, no. 10, pp. 3939–3945, 2017.

[11] G. Litjens, T. Kooi, B. E. Bejnordi et al., "A survey on deep learning in medical image analysis," *Medical Image Analysis*, vol. 42, pp. 60–88, 2017.

[12] J. Ker, L. Wang, J. Rao, and T. Lim, "Deep learning applications in medical image analysis," *Ieee Access*, vol. 6, pp. 9375–9389, 2017.

[13] M. I. Razzak, S. Naz, and A. Zaib, "Deep learning for medical image processing: overview, challenges and the future," in *Classification in BioApps*, pp. 323–350, Springer, Berlin, Germany, 2018, Lecture Notes in Computational Vision and Biomechanics.

[14] A. Fourcade and R. H. Khonsari, "Deep learning in medical image analysis: a third eye for doctors," *Journal of stomatology, oral and maxillofacial surgery*, vol. 120, no. 4, pp. 279–288, 2019.

[15] P. Cui, Z. Guo, J. Xu et al., "Tissue recognition in spinal endoscopic surgery using deep learning," in *Proceedings of the 2019 IEEE 10th International Conference on Awareness Science and Technology (iCAST)*, pp. 1–5, IEEE, Morioka, Japan, October 2019.

[16] R. Girshick, J. Donahue, T. Darrell, and J. Malik, "Rich feature hierarchies for accurate object detection and semantic segmentation," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 580–587, Columbus, OH, USA, June 2014.

[17] K. Simonyan and A. Zisserman, “Very deep convolutional networks for large-scale image recognition,” arXiv preprint arXiv:1409.1556, 2014.

[18] R. Girshick, “Fast r-cnn,” in *Proceedings of the IEEE international conference on computer vision*, pp. 1440–1448, Santiago, Chile, December 2015.

[19] S. Ren, K. He, R. Girshick, and J. Sun, “Faster r-cnn: towards real-time object detection with region proposal networks,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 39, no. 6, pp. 1137–1149, 2016.

[20] R. Joseph and F. Ali, “Yolov3: an incremental improvement,” arXiv preprint arXiv:1804.02767, 2018.

[21] A. Bochkovskiy, C. Wang, and H. Mark Liao, “Yolov4: optimal speed and accuracy of object detection,” arXiv preprint arXiv:2004.10934, 2020.

[22] A. P. Bradley, “The use of the area under the roc curve in the evaluation of machine learning algorithms,” *Pattern Recognition*, vol. 30, no. 7, pp. 1145–1159, 1997.

[23] R. Kumar and A. Indrayan, “Receiver operating characteristic (roc) curve for medical researchers,” *Indian Pediatrics*, vol. 48, no. 4, pp. 277–287, 2011.

[24] A. Teramoto, H. Fujita, O. Yamamuro, and T. Tamaki, “Automated detection of pulmonary nodules in PET/CT images: ensemble false-positive reduction using a convolutional neural network technique,” *Medical Physics*, vol. 43, no. 6Part1, pp. 2821–2827, 2016.

[25] P. Lakhani and B. Sundaram, “Deep learning at chest radiography: automated classification of pulmonary tuberculosis by using convolutional neural networks,” *Radiology*, vol. 284, no. 2, pp. 574–582, 2017.

[26] H. Luo, G. Xu, C. Li et al., “Real-time artificial intelligence for detection of upper gastrointestinal cancer by endoscopy: a multicentre, case-control, diagnostic study,” *The Lancet Oncology*, vol. 20, no. 12, pp. 1645–1654, 2019.

[27] A. A. Mohamed, W. A. Berg, H. Peng, Y. Luo, R. C. Jankowitz, and S. Wu, “A deep learning method for classifying mammographic breast density categories,” *Medical Physics*, vol. 45, no. 1, pp. 314–321, 2018.

Hindawi
Journal of Healthcare Engineering
Volume 2021, Article ID 6033860, 14 pages
https://doi.org/10.1155/2021/6033860

Research Article

Evaluating Drug Risk Using GAN and SMOTE Based on CFDA's Spontaneous Reporting Data

Jianxiang Wei,[1,2] Guanzhong Feng,[3] Zhiqiang Lu,[3] Pu Han,[1] Yunxia Zhu,[3] and Weidong Huang[1,2]

[1] *School of Management, Nanjing University of Posts and Telecommunications, Nanjing 210003, China*
[2] *Key Research Base of Philosophy and Social Sciences in Jiangsu-Information Industry Integration Innovation and Emergency Management Research Center, Nanjing 210003, China*
[3] *School of Internet of Things, Nanjing University of Posts and Telecommunications, Nanjing 210003, China*

Correspondence should be addressed to Jianxiang Wei; jxwei@njupt.edu.cn

Received 31 May 2021; Accepted 19 August 2021; Published 29 August 2021

Academic Editor: Ayush Dogra

Adverse drug reactions (ADRs) pose health threats to humans. Therefore, the risk re-evaluation of post-marketing drugs has become an important part of the pharmacovigilance work of various countries. In China, drugs are mainly divided into three categories, from high-risk to low-risk drugs, namely, prescription drugs (Rx), over-the-counter drugs A (OTC-A), and over-the-counter drugs B (OTC-B). Until now, there has been a lack of automated evaluation methods for the three status switch of drugs. Based on China Food and Drug Administration's (CFDA) spontaneous reporting database (CSRD), we proposed a classification model to predict risk level of drugs by using feature enhancement based on Generative Adversarial Networks (GAN) and Synthetic Minority Over-Sampling Technique (SMOTE). A total of 985,960 spontaneous reports from 2011 to 2018 were selected from CSRD in Jiangsu Province as experimental data. After data preprocessing, a class-imbalance data set was obtained, which contained 887 Rx (accounting for 84.72%), 113 OTC-A (10.79%), and 47 OTC-B (4.49%). Taking drugs as the samples, ADRs as the features, and signal detection results obtained by proportional reporting ratio (PRR) method as the feature values, we constructed the original data matrix, where the last column represents the category label of each drug. Our proposed model expands the ADR data from both the sample space and the feature space. In terms of feature space, we use feature selection (FS) to screen ADR symptoms with higher importance scores. Then, we use GAN to generate artificial data, which are added to the feature space to achieve feature enhancement. In terms of sample space, we use SMOTE technology to expand the minority samples to balance three categories of drugs and minimize the classification deviation caused by the gap in the sample size. Finally, we use random forest (RF) algorithm to classify the feature-enhanced and balanced data set. The experimental results show that the accuracy of the proposed classification model reaches 98%. Our proposed model can well evaluate drug risk levels and provide automated methods for status switch of post-marketing drugs.

1. Introduction

Drug risk has always been a worldwide concern, and its most intuitive manifestation is adverse drug reactions (ADRs). The severity of adverse reactions of different drugs varies greatly. In some cases, it can even be fatal, which poses a great threat to people's health [1]. ADRs refer to harmful reactions of qualified drugs that have nothing to do with the purpose of medication under normal usage and dosage. Edwards and Aronson proposed a clearer definition of ADRs: "An appreciably harmful or unpleasant reaction, resulting from an intervention related to the use of a medicinal product, which predicts hazard from future administration and warrants prevention or specific treatment, or alteration of the dosage regimen, or withdrawal of the product [2]."

In order to reduce the harm caused by ADRs, the classification system of prescription (Rx) drugs and over-the-counter (OTC) drugs has become an internationally common model. According to the regulations of the US

Food and Drug Administration (FDA), drugs are classified into Rx drugs and OTC drugs based on indicators such as toxicity and dependence [3]. Compared with Rx drugs, OTC drugs have less adverse reactions, and they can be purchased without a doctor's prescription to treat mild diseases. For OTC drugs, the China Food and Drug Administration (CFDA) further divides OTC drugs into two categories, namely, OTC-A drugs and OTC-B drugs, of which OTC-B is safer [4]. Therefore, drugs in China are divided into three categories, and the order of risk levels is Rx > OTC-A > OTC-B. At present, the drug regulatory authorities of many countries implement a re-evaluation system for post-marketing drugs, and switch Rx drugs and OTC drugs based on the frequency and severity of ADRs [5, 6]. As Brass argues, removal of the requirement for prescriptions saves both the health care professional and the patient time, but assessment of the ability of patients to use drugs in this manner is a critical component of the regulatory review [7]. This method mainly relies on the judgement of medical experts and lacks an automated risk identification technology. We hope to build a multi-classifier to determine whether a drug belongs to one of the above three categories by evaluating ADRs, in order to provide an objective and automatic method for the status switch of drugs. Furthermore, the accurate classification of drugs will provide more convenience for patients' medication, while reducing the risk of ADRs as much as possible.

The ADR reports used in this experiment all originate from CFDA's spontaneous reporting system (SRS). Spontaneous reporting means that medical workers voluntarily report suspicious ADRs discovered in the clinic to drug manufacturers, adverse reaction monitoring agencies, drug regulatory departments, etc. [8]. SRS is suitable for wide deployment in various regions and can collect large amounts of ADR data [9]. Nowadays, most members of WHO Uppsala Monitoring Centre (UMC) have adopted this system [10]. However, the information in many reports is too rough, which may affect the causality of adverse reactions, leading to over or under attribution [11]. At the same time, incomplete or missing reports make it impossible to calculate the incidence of ADR accurately. Therefore, it is necessary to standardize the original data and use the method of signal detection to extract effective information. At present, the commonly used signal detection method for ADRs is disproportionality analysis (DPA) [12, 13]. The proportional reporting ratio (PRR) used in this paper is one of the DPA methods. Based on the PRR method, we can build a data matrix with drugs as samples, ADR symptoms as features, and signal detection results as feature values. The last column of the matrix represents the category label of each drug.

Since the overall data contains many types of ADRs, and only part of the adverse reactions is caused by one drug, this data matrix is high-dimensional and sparse. A large number of features can increase interference noise and may obscure some important ADR data. In order to improve the classification accuracy, it is necessary to perform feature selection on the data set. The principle of feature selection is to use detection methods to evaluate all features from the data set, and retain features that are efficient and reliable for data classification [14]. The experiment uses machine-learning methods to extract features with high importance scores.

Considering that the high-dimensional feature space contains a lot of information about ADR symptoms, we cannot simply keep important features and delete those that are not helpful for classification, because this method may cause some serious ADR features to disappear, making it difficult to accurately evaluate the potential risks of drugs. On the basis of retaining the existing ADR features, we hope to expand the feature space with more effective data that are helpful for the classification. Goodfellow et al. proposed the concept of generative adversarial networks (GANs) in 2014 [15]. As a popular theory in deep learning in recent years, GAN has achieved outstanding performance in data generation. Therefore, according to the feature selection data, we use GAN to generate similar artificial data, which are added to the feature space to achieve feature enhancement.

In terms of sample space, the number of Rx drugs and OTC drugs in our ADR data set is extremely imbalanced. Traditional classification techniques perform poorly on this type of data because they tend to favor the majority class. The synthetic minority over-sampling technique (SMOTE) algorithm proposed by Chawla et al. is one of the most representative external methods to balance data sets through resampling [16]. We use SMOTE to balance the data set by adding samples based on k-nearest neighbors in the minority class. The samples of Rx, OTC-A, and OTC-B drugs will reach a balanced state after SMOTE resampling, which lays a data foundation for using conventional random forest (RF) classification algorithm.

The purpose of this paper was to build a high-accuracy drug risk level classification model, which can be deployed in the Chinese spontaneous reporting system. When a drug manufacturer applies to the drug regulatory authority for drug category switch, CFDA organizes medical experts to conduct drug risk assessment. In this process, the proposed model can automatically identify the drug category according to the ADR monitoring data after the drug is put on the market, which can provide auxiliary decision support for experts.

2. Related Work

With the development of computer science, the use of machine learning to solve ADR problems is common and widely used, which makes great contribution to the control of medication risks. In 2011, Pouliot et al. used more than 480,000 molecular activity data in the PubChem database to establish a logistic regression model to predict the level of ADRs that may be caused by target drugs [17]. The results show that 75% of the adverse reaction signals mined by this model could be verified by relevant medical literature or drug instructions. In the same year, Santiago et al. proposed a new ADR detection method, which compared and screened the drugs involved in the adverse reaction signals, and then obtained the final mining results [18]. In this way, the sensitivity of mining adverse reactions related to rhabdomyolysis reached 70%, and the positive detection rate reached 45%. In 2013, Chen et al. realized the identification of high-risk proteins and the discovery of potential adverse reaction mechanisms through the analysis of the high-risk protein network of ADRs [19]. This study analyzed the data of drug target proteins, protein pathways, and proteins related to adverse reactions. The results found a total of 41 ADR protein subnetworks, and found that certain biological enzymes and transport proteins are the key factors causing adverse reactions.

In the field of risk mining of ADRs, researchers have conducted a lot of research on various types of spontaneous report databases and have achieved sufficient results. In 2014, Roberto et al. used signal detection methods to conduct data mining, trying to detect the serious cardiovascular adverse reaction signals of triptans drugs [20]. The results show that triptans drugs are related to a variety of adverse reactions such as ischemic cerebrovascular complications. In 2015, Mai et al. conducted data mining in the spontaneous report database and found that the use of statin drugs may increase the risk of rectal cancer or pancreatic cancer [21]. In 2018, Scholl et al. proposed a prediction-model-based approach to improve the efficiency of full database screening [22]. The AUC value and the ratio of potential signals of this method have been greatly improved compared with traditional signal detection methods. In 2019, to resolve entity-level ADR classification tasks, Alimova and Tutubalina investigated deep neural network models in the natural language processing (NLP) field based on various ADR corpus [23].

In recent years, researchers have analyzed ADR from multiple perspectives such as patient age and drug interaction, and have proposed many new risk detection methods. In 2020, Martocchia et al. evaluated the incidence of adverse events and drug-drug interactions exposed to polypharmacy and proposed that the application of certain software programs could significantly reduce the incidence of adverse events at every level of healthcare [24]. In 2021, Giangreco and Tatonetti pointed out that detection of ADR is challenging due to dynamic biological processes during ontogeny, which alter pharmacokinetics and pharmacodynamics [25]. The population modeling technique they proposed exhibited normally distributed and robust ADR risk estimation at all development stages of children. In the same year, Mehta et al. reviewed the risk assessment methods of prescription drug and developed more than two dozen prescription drug-based risk indices, which differ significantly in design, performance, and application [26].

Regarding China's spontaneous report data, scholars have integrated and evaluated ADR information, and have begun to measure drug risks in an intelligent way. In 2015, Ge et al. used the NLP method to extract knowledge of adverse reactions in a large number of Chinese clinical narrative texts [27]. Based on the results of knowledge extraction, they established a knowledge base corresponding to drugs and adverse reactions, and set up a website to provide online query and to download ADR information. In 2020, we compared four drug-risk prediction models using machine-learning methods as classifiers, and determined the best risk prediction framework [28], with a classification accuracy rate of 95%.

In order to further improve the classification accuracy so that the risk prediction model can be applied in practice, this paper is based on the previous research, and realizes the feature enhancement of high-dimensional ADR feature space through the combination of GAN and feature selection. Furthermore, by comparing with our previous models, we propose a better predictive model for evaluating drug risks.

3. Materials

ADR reports used in this study were obtained from the CFDA in Jiangsu Province. The data set covered a total of 985,960 ADR reports in Jiangsu Province from 2011 to 2018, including report ID, report address, patient age, gender, drug name, and ADRs symptom. Due to invalid and duplicate reports, we deleted data with no reference value and standardized the names of drugs and ADR symptoms. Then 1,047 drug names and 751 ADR symptoms were prepared. In more detail, for each drug, the ADR mentioned in one report would increase the total of corresponding ADR symptoms by one. The result of the final statistics is a table with the drug names corresponding to the frequency of all types of ADRs. Data set could be described as the following.

Sample space:

$$X = \{x_1, x_2, \ldots, x_m\}. \tag{1}$$

Here, $m = 1047$. The drugs and ADRs make up the sample space together. Drugs are the samples and ADRs are the features.

Feature vector:

$$x_i = \left(x_{i1}, x_{i2}, \ldots, x_{i\,d}\right) \in X. \tag{2}$$

Here, $d = 751$. Every sample is composed of d features, and x_i is one of the feature vectors in sample space, where $x_{i\,d}$ represents the frequency under the matching drug-ADR pairs.

According to the China Medical Information Platform, we manually labeled all drug samples, with values 0, 1, and 2 representing Rx, OTC-A, and OTC-B, respectively.

The statistical results in Table 1 show that Rx drugs account for a high proportion, while the two categories of OTC drugs are the opposite, which means that the classes in the data set are imbalanced.

4. Methods

4.1. Model Framework. Figure 1 shows the flowchart of the proposed model. The model is mainly divided into four stages: signal detection stage, feature enhancement stage, minority expansion stage, and RF classification stage.

(1) Signal detection stage: the first step of the proposed model is to use signal detection on the preprocessed spontaneous report data, and calculate the PRR value of the drug-ADR pairs to obtain the ADR imbalanced data set.

(2) Feature enhancement stage: based on the ADR imbalanced data set, the model selects the top 200 features of classification importance, and uses GAN

Table 1: Quantity information of drugs in data set.

Drug category	Label	Sample size	Percentage
Rx	0	887	84.72
OTC-A	1	113	10.79
OTC-B	2	47	4.49

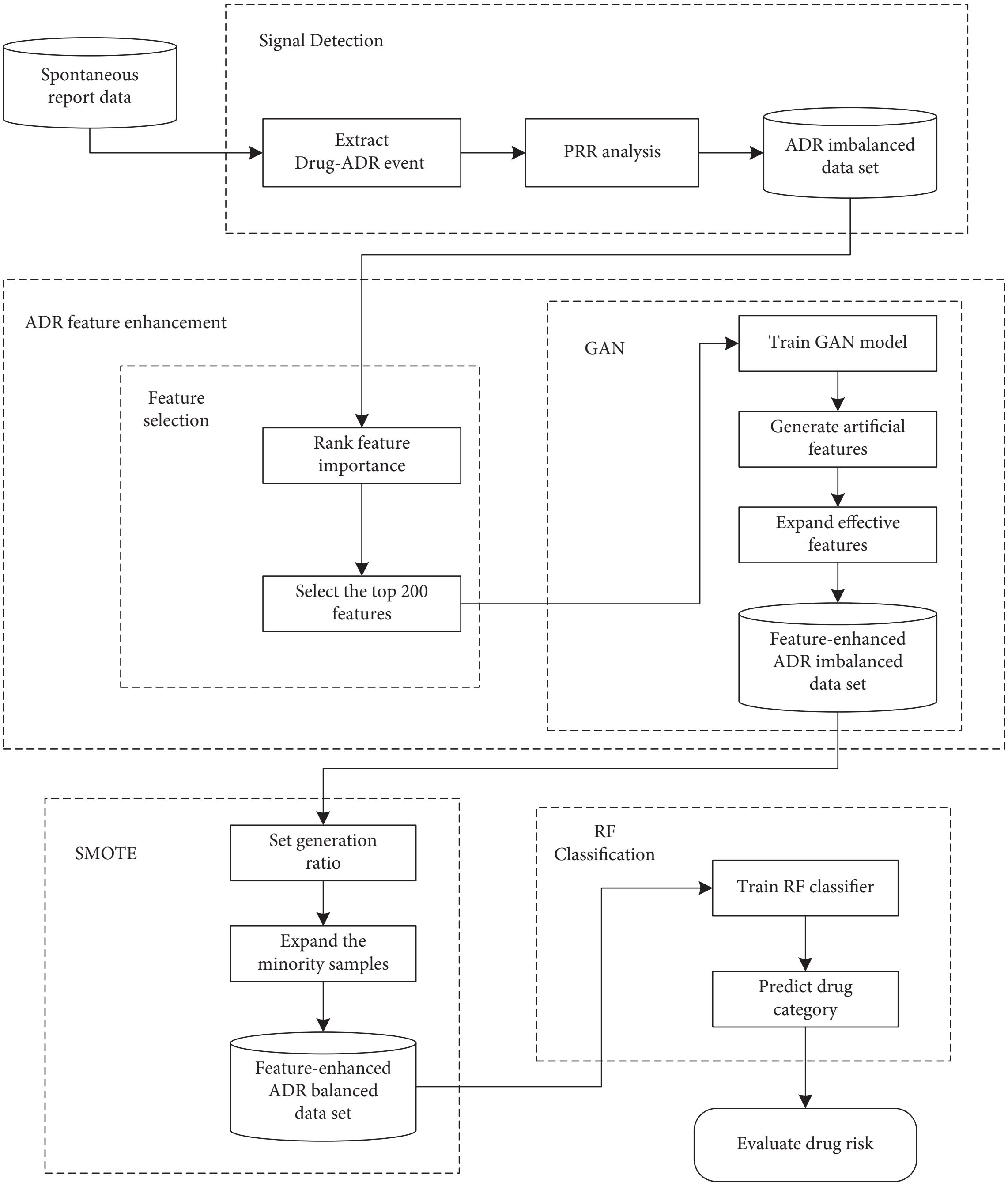


Figure 1: Flowchart of the proposed model.

to generate artificial features that meet the real data distribution. The generated features are added to the imbalanced ADR data set, so that the feature space contains more effective features, which can improve the classification accuracy. So far, the feature-enhanced ADR imbalance data set has been obtained.

(3) Minority expansion stage: SMOTE is used to expand the minority samples (OTC-A and OTC-B) to equalize the number of the three categories of drugs, which helps to obtain a feature-enhanced ADR balanced data set.

(4) RF classification stage: the RF algorithm is used to classify the feature-enhanced ADR balanced data set. Finally, we analyze the results of the proposed model based on multiple indicators, and further evaluate the risks of postmarketing drugs.

4.2. Signal Detection. The DPA method is currently the most used ADR signal detection technology [29]. DPA is used to measure the disproportion or imbalance of the sample distribution in the database. If the number of occurrences associated with drug and adverse event is greater than the expected number or the number of other combinations, it is considered that there is a potential connection between the drug and the adverse event, which may be a positive ADR signal. Calculation of DPA is based on the principles using the two-by-two contingency table.

Proportional reporting ratio (PRR) is one of the DPA methods, which was proposed in 2001 by Evans of the British Medical Regulatory Authority [30], and it is a key method for ADR signal detection in the world. The calculation of PRR is similar to the relative risk in epidemiological studies, which is used to quantify the strength of the drug-ADR association. According to Table 2, the formula to compute PRR value is

$$\mathrm{PRR} = \frac{(A/(A+B))}{(C/(C+D))}. \tag{3}$$

Formula (3) indicates that if the PRR value of a drug-ADR pair is larger, the relative risk is higher, so the risk of the adverse reaction corresponding to the drug is greater. In our study, after calculating the PRR value of all drug-ADR pairs based on statistical data, the data matrix is established with drugs as the sample, ADR as the feature, and PRR results as the matrix value. The last column of the data matrix is the category label of each drug, where Rx is “0,” OTC-A is “1,” and OTC-B is “2.” Due to the quantitative difference among the three categories of drugs, we got the ADR imbalance data set.

4.3. Feature Enhancement. Since the overall data contain many types of ADRs, and only part of the adverse reactions is caused by one drug, this data matrix obtained by PRR is high-dimensional and sparse. In order to improve the model’s classification accuracy of this data set, we expand the effective ADR data in the feature space to achieve feature enhancement.4.3.1. Algorithm ID4 is Used for Feature Selection (FS)

In the process of decision tree attribute splitting, the Gini index is used to calculate the contribution of a single feature for the correct classification. During tree growth, the purity measure of split at node k is:

Table 2: DPA two-by-two contingency table.

	Target ADRs	Other ADRs	Total
Target drugs	A	B	$A+B$
Other drugs	C	D	$C+D$
Total	$A+C$	$B+D$	$A+B+C+D$

$$\mathrm{Gini}(p_k) = \sum_{k=1}^{n} p_k(1-p_k) = 1 - \sum_{k=1}^{n} p_k^2. \tag{4}$$

In formula (4), p_k represents the probability that the sample is correctly classified at node k. The sample is divided into different branches to produce their branch sets T^v, and the purity measure is as follows:

$$\mathrm{Gini}_{\mathrm{index}(T,k)} = \sum_{v=1}^{V} \frac{|T^v|}{|T|}\mathrm{Gini}(T^v). \tag{5}$$

T represents the current divided set, and the Gini index reflects the probability that any two branch sets are inconsistent. A smaller Gini index in formula (5) indicates that the branch set is purer, which also means that the classification accuracy will be higher. Therefore, node k strives to meet the minimum purity:

$$k^* = \arg\min_k \mathrm{Gini}_{\mathrm{index}(T,k)}. \tag{6}$$

The feature f_i is the classification basis of node k, and left and right branches can be obtained. They are measured according to the Gini changes of the branches:

$$\mathrm{Gini}_{(f_i,k)} = \mathrm{Gini}(P_k) - \mathrm{Gini}(P_1) - \mathrm{Gini}(P_r). \tag{7}$$

$\mathrm{Gini}(P_1)$, $\mathrm{Gini}(P_r)$ represent the Gini index of the left and right branches, respectively. After calculating $\mathrm{Gini}_{(f_i,k)}$ in formula (7), the importance of the feature f_i in the j-th tree is:

$$\mathrm{Im}_j^{\mathrm{Gini}} = \sum_{m\in M} \mathrm{Gini}_{(f_i,k)}. \tag{8}$$

The importance of feature f_i on a single tree is calculated by formula (8). Furthermore, in the total number of m trees, the feature f_i appears when part of the tree nodes split. Then, the overall importance of measuring feature f_i is:

$$\mathrm{Im}_{f_i} = \sum_{j=1}^{m} \mathrm{Im}_j^{\mathrm{Gini}}. \tag{9}$$

In order to select features that are more effective for classification, the features are ranked in the descending order of importance calculated by formula (9). The first 200 main features are retained as the basis for the next step of GAN feature generation.

4.3.1. Use GAN to Generate New Features. GAN is a generative model based on zero-sum game theory. It includes a generative model (G) and a discriminant model (D), both of which are based on neural networks. In the training process of G and D, G generates data similar to the true value through the noise space z. The goal of D is to distinguish between real data or generated data. Generator and discriminator are iteratively optimized with each other, so that their performance continues to improve. In the end, the two models reached a Nash equilibrium. At this time, the data generated by GAN approximates the real data [31, 32]. The evaluation formula of GAN is as follows:

$$\min_G \max_D V(D,G) = E_{x\sim P_{\text{data}}(x)}[\log D(x)] + E_{z\sim P_z(z)}[\log(1 - D(G(x)))]. \tag{10}$$

In formula (10), x is the real data, which conforms to the $P_{\text{data}}(x)$ distribution; z is the hidden space noise, which conforms to the $P_z(z)$ distribution. $V(D,G)$ represents the degree of difference between the real sample and the generated sample. Formula (10) indicates that when the discriminator maximizes the difference and the generator minimizes the difference between the real samples and the generated samples, after multiple rounds of iterative training, realistic data can be obtained.

We use two sets of neural networks to construct G and D, respectively. The key factors in the training process are gradient descent, alternate training, and back propagation. The training steps in the experiment are summarized as follows:

Step1: select some samples $\{z_1, z_2, \ldots, z_m\}$ from the input random noise $P_z(z)$.

Step2: sampling from the original training set, the number of samples $\{x_1, x_2, \ldots, x_m\}$ is the same as the noise samples.

Step3: set the parameter of D to θ_d, and use the gradient ascent algorithm in formula (11) to update the discriminator:

$$\nabla \frac{1}{m}\sum_{i=1}^{m}[\log D(x_i) + \log(1 - D(G(z_i)))]. \tag{11}$$

Step4: repeat steps 1–3 for k times, and then update G once.

Step5: set the parameter of G to θ_g, and use the gradient descent algorithm in formula (12) to update the generator:

$$\nabla \frac{1}{m}\sum_{i=1}^{m}\log(1 - D(G(z_i))). \tag{12}$$

Step6: repeat steps 1–5 until the GAN model converges.

Input the features selected in Step 4.3 (1) into GAN to generate an equal number of artificial ADR features. These generated features satisfy the real value distribution and are consistent with the original data. Use generated ADR features as real data to expand the feature space in order to enhance the risk characteristics of the drugs. Now, we obtain a feature-enhanced ADR imbalance data set.

4.4. Synthetic Minority Over-Sampling Technique (SMOTE).

After adding the generated ADR features to the data set, the number of effective features in the sample is increased, which is helpful for subsequent classification. However, the proportions of Rx, OTC-A, and OTC-B drugs in the data set are quite imbalanced. Traditional classification algorithms will seriously bias the majority class and ignore the minority class, leading to deviations in the result. Therefore, for the imbalanced data set in this experiment, we use the SMOTE algorithm to expand the minority samples before classification.

The core of SMOTE is to insert randomly generated new samples between the minority samples and their neighbor samples [33]. This can increase the number of minority samples and improve the class imbalance distribution of the data set [34]. The steps of the SMOTE are as follows:

Step 1: the number of majority samples in the data set is N^+, and the number of minority samples is N^-. Calculate the imbalance ratio IR and oversampling rate K of the original data set:

$$\text{IR} = \frac{N^+}{N^-}. \tag{13}$$

Round down IR in formula (13) to get the oversampling rate K:

$$K = \lfloor \text{IR} \rfloor, \tag{14}$$

($\lfloor\rfloor$ means rounding down the data.)

Step 2: for each minority sample x_i, calculate the Euclidean distance with other minority samples, and find the k nearest neighbors. The Euclidean distance is calculated as follows:

$$d(x_i, x_j) = \sqrt{(x_{i1} - x_{j1}) + (x_{i1} - x_{j2}) + \cdots + (x_{ip} - x_{jp})}. \tag{15}$$

Step 3: according to the oversampling rate K in formula (14) and Euclidean distance $d(x_i, x_j)$ in formula (15), K samples are randomly selected from the k nearest neighbors with replacement, and mark them as $\overline{x_i}\,(i = 1, 2, \ldots, K)$. Calculate the difference between x and $\overline{x_i}$ as $(x - \overline{x_i})$.

Step 4: use formula (16) to synthesize each new sample x^i_{new}:

$$x^i_{\text{new}} = x + \text{rand}(0,1) \times (x - \overline{x_i}), \quad i = 1, 2, \ldots, K, \tag{16}$$

($\text{rand}(0,1)$ returns a random value in the interval $(0,1)$.)

Step 5: repeat the above steps to synthesize $K \cdot N^{-}$ data artificially for the minority samples.

After the above steps, the three categories of Rx, OTC-A, and OTC-B in the data set have reached the same number, which improves the data distribution in the sample space. As a result, we obtained the feature-enhanced ADR balanced data set, which laid a data foundation for classification.

4.5. Random Forest Classifier. We use the random forest (RF) algorithm to classify the feature-enhanced ADR balanced data set. The RF algorithm is a machine-learning method proposed by Breiman in 2001 [35]. Its main idea is to build a forest containing multiple decision trees. Each decision tree adopts a random decision-making method in this process and remains independent during classification. Each decision tree in the RF will predict the outcome. Finally, all the outcomes are integrated by voting, and the class with the highest probability is selected as the classification result [36]. The steps of RF classification are as follows:

Step 1: assume that the number of samples in the training set (S) is N. We randomly select N samples from the training set with replacement as the training set S_i of the decision tree T_i. A total of K training sets are extracted to construct K decision trees.

Step 2: the dimension of the features in each sample is M. In the process of training the decision tree, m subsets are randomly selected from all the features of each node.

Step 3: the decision tree selects a node with the best splitting ability in the feature subset to split.

Step 4: each decision tree grows to the maximum extent and does not require pruning.

Step 5: all decision trees constitute the final RF, and the result of the classification is determined by voting.

4.6. Evaluation Metrics. Traditional classification algorithms use precision metric to determine the performance of the classifier on the data set. Although it is effective for balanced data, there will be obvious deviations for unbalanced data.

For example, for tumor detection data, the proportion of benign is very high, and the proportion of malignant is very low. High accuracy can be obtained by classifying all tumors as benign. However, this classification is meaningless, because for issues such as disease detection, disaster prediction, and credit fraud, the minority samples are of great significance and need to be focused on.

For a given sample set, we can get the confusion matrix by comparing the real class with the class predicted by the classifier [37]. As shown in Table 3, there are four situations:

Table 3: Classification in the confusion matrix.

	Positive	Negative
True	True positive (TP)	True negative (TN)
False	False positive (FP)	False negative (FN)

TP: the number of samples that predict the positive class as a positive class.
TN: the number of samples that predict the negative class as a negative class.
FP: the number of samples that predict a negative class as a positive class.
FN: the number of samples that predict a positive class as a negative class.

According to the confusion matrix in Table 3, the following evaluation metrics can be calculated:

(1) Precision: Tthe precision rate reflects the proportion of true positive samples in the positive class judged by the classifier.

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}}. \tag{17}$$

(2) Recall: the recall rate reflects the proportion of positive classes that are correctly classified in the total positive classes.

$$\text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}}. \tag{18}$$

(3) Accuracy: the accuracy rate reflects the classifier's ability to predict the positive and negative classes correctly.

$$\text{Accurary} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{FN} + \text{FP} + \text{TN}}. \tag{19}$$

(4) *F*-measure: F1 is the harmonic mean of precision and recall [38], and is a commonly used evaluation criterion for classification of imbalanced data sets. After obtaining Precision and Recall in formulas (17) and (18), F1 can be calculated as

$$\text{F1} = \frac{\left(1 + \beta^2\right) * \text{Recall} * \text{Precision}}{\beta^2 * \text{Recall} + \text{Precision}}, \tag{20}$$

β is the scale factor, and its usual value is 1.

(5) Macro-avg: macro average is a commonly used evaluation index for multi-classification problems, which can measure the overall situation of the classifier [39]. For formulas (17), (18), and (20), the values of each class are first calculated, and then the average values of all the classes are calculated.

$$\text{Macro}_P = \frac{1}{n} \sum_{i=1}^{n} P_i. \tag{21}$$

$$\text{Macro}_R = \frac{1}{n} \sum_{i=1}^{n} R_i. \tag{22}$$

$$\text{Macro}_{\text{F1}} = \frac{2 * \text{Macro}_P * \text{Macro}_R}{\text{Macro}_P + \text{Macro}_R}, \tag{23}$$

n represents the number of classes, i represents each class.

Macro average in formula (21)–(23) treats each class equally, and its results are more susceptible to minority samples. In other words, macro average has advantages in highlighting the classification performance of minority samples.

(6) Weighted-avg: The weighted average can comprehensively evaluate the accuracy of classification [40]. By assigning weight to each class, the average value of all classes is calculated according to Precision, Recall, and F1 in formulas (17), (18), and (20).

$$\text{Weighted}_P = \sum_{i=1}^{n} \frac{C_i}{|C|} * P_i,$$
$$\text{Weighted}_R = \sum_{i=1}^{n} \frac{C_i}{|C|} * R_i, \tag{24}$$
$$\text{Weighted}_{\text{F1}} = \sum_{i=1}^{n} \frac{C_i}{|C|} * \text{F1}_i,$$

n represents the number of classes, i represents each class, $|C|$ represents all samples, and C_i represents the samples included in one class.

(7) Receiver Operating Characteristic (ROC) Curve

Among the evaluation criteria for imbalanced data sets, the ROC curve is a generally accepted and comprehensive evaluation criterion [41]. The ROC curve has a false positive rate ($\text{FPR} = \text{FP}/(\text{FP} + \text{TN})$) on the horizontal axis and a true positive rate ($\text{TPR} = \text{TP}/(\text{TP} + \text{FN})$) on the vertical axis. Through the cross-validation method, multiple sets of point pairs (FPR, TPR) of the classifier can be obtained. Then, draw them to a plane and connect them to form the final ROC curve. The ROC curve is a very intuitive way to evaluate the classifier. The closer the curve is to the upper left corner, the better the performance of the classifier.

Area under curve (AUC) refers to the area enclosed by the ROC curve and the coordinate axis. The value of this area will not be greater than 1. Since the ROC curve is generally above the line $y = x$, the value of AUC ranges between $[0.5, 1]$. The closer the AUC is to 1, the higher the accuracy of classification.

4.7. Experiment Design. In order to observe the effect of the abovementioned methods on the classification of CFDA's spontaneous reporting data, this paper designs three comparative models.

In the first model (*Model 1. RF*), we use the data set after PRR signal detection as the basis (ADR imbalance data set). The total number of three categories of drugs is 1047, including 887 Rx drugs (label = 0), 113 OTC-A drugs (label = 1), and 47 OTC-B drugs (label = 2). The data space contains 751 features (ADRs). Use traditional RF algorithm for classification.

In the second model (*Model 2. SMOTE + RF*), we use the SMOTE algorithm to expand the data set after PRR signal detection, so that the quantity of each category reaches a balance (ADR balanced data set). The total number of drugs is 2661, and the number of Rx (label = 0), OTC-A (label = 1), OTC-B (label = 2) drugs are equal, all of which are 887. The data space contains 751 features. Then, use RF for classification.

In the third model (*Model 3. FS_GAN + SMOTE + RF*), we will use the model proposed in this paper. The ADR data set used by this model has also been improved in terms of samples and features (feature-enhanced ADR balanced data set). The total number of drugs is 2661, and the number of Rx (label = 0), OTC-A (label = 1), and OTC-B (label = 2) drugs are equal, all of which are 887. The data space contains 951 features (751 original + 200 generated). Finally, the RF algorithm is used for classification.

The experiment in this article consists of two sections. In the first section, the above three models all use 70% of the sample space as the training set, and the remaining 30% as the test set. Then, we observe the classification results based on the test set. In the second section, we input the actual ADR data collected by CFDA into all three trained models and observe the results. Furthermore, it means that the three models use the same actual ADR data after PRR (1047 samples) as the test set.

5. Results

5.1. Results of the Classifiers Using the Test Set. In this section, the three models use 70% of their sample space for training, and use the remaining 30% as the test set. The sample size in the test set of each model is calculated as follows:

Sample size (model 1) = 1047 × 30% = 315
Sample size (model 2) = 2661 × 30% = 799
Sample size (model 3) = 2661 × 30% = 799

Therefore, the sample size of test sets used by each model is 315 (Model 1), 799 (Model 2), and 799 (Model 3). The confusion matrices obtained by classification are shown in Figure 2:

Figure 2 shows three confusion matrices of three models, from which it can be seen that Model 3 (FS_GAN + SMOTE + RF) has the largest proportion of results on the diagonal, which means it has the highest accuracy of classification. More detailed evaluation indicators are shown in Table 4.

Table 4 shows the evaluation metrics of the three models based on their test set. Model 1 is biased towards the majority class, so the prediction results for OTC-A (label = 1) and OTC-B (label = 2) are very poor, and its accuracy is the lowest, only 84.44%. Model 2 balances the data set and can predict most of the OTC drugs (label = 1, 2), with an accuracy rate of 91.99%. Model 3 has the highest prediction accuracy for minority samples, reaching 96.25%. From the macro average and weighted average metrics, Model 1 is the worst, Model 2 ranks second, and Model 3 has the best performance.

5.2. Validation Results Based on Actual ADR Data. In the verification section, the actual ADR data are used as the test set to validate the prediction results of the three trained

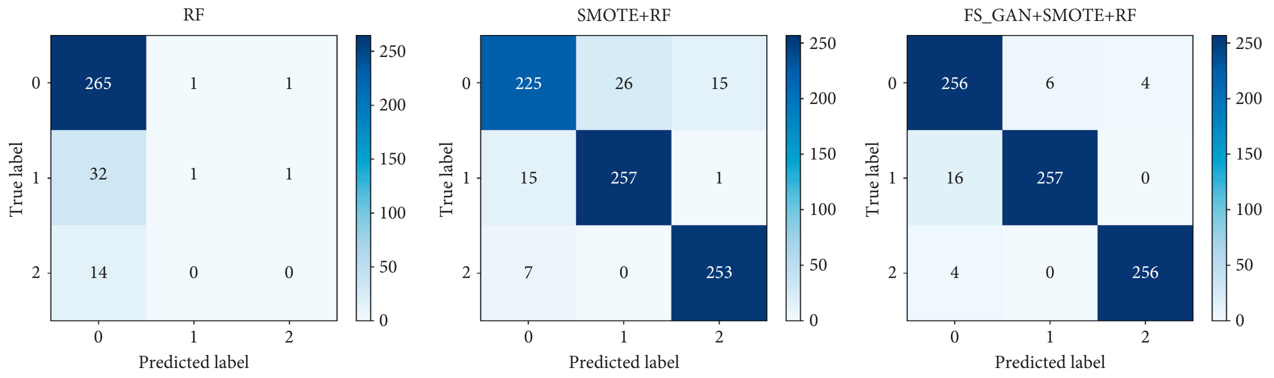


Figure 2: Confusion matrices based on the test set.

Table 4: Evaluation metrics based on the test set.

Classifier	Label	Precision	Recall	F1	Accuracy (%)
Model 1 (RF)	0	0.85	0.99	0.92	84.44
	1	0.50	0.03	0.06	
	2	0.00	0.00	0.00	
	Macro-avg	0.45	0.34	0.32	
	Weighted-avg	0.78	0.84	0.78	
Model 2 (SMOTE + RF)	0	0.91	0.85	0.88	91.99
	1	0.91	0.94	0.92	
	2	0.94	0.97	0.96	
	Macro-avg	0.92	0.92	0.92	
	Weighted-avg	0.92	0.92	0.92	
Model 3 (FS_GAN + SMOTE + RF)	0	0.93	0.96	0.94	96.25
	1	0.98	0.94	0.96	
	2	0.98	0.98	0.98	
	Macro-avg	0.96	0.96	0.96	
	Weighted-avg	0.96	0.96	0.96	

models in Section5.1. The test sample size of the three models equals to that of the actual ADR data, which is 1047.

The confusion matrices corresponding to the three models are shown in Figure 3.

Figure 3 illustrates the confusion matrices of the three models using the actual ADR data after PRR signal detection as the input ($\text{sample size}_{1,2,3} = 1047$). In the confusion matrix, the blocks on the diagonal indicate the number of correctly classified labels. For each model, the sum of correctly predicted data is calculated as follows:

$$\text{Sum (Model 1)} = 874 + 41 + 13 = 928$$

$$\text{Sum (Model 2)} = 845 + 102 + 44 = 991$$

$$\text{Sum (Model 3)} = 876 + 105 + 44 = 1025$$

Under the verification of the same data set, the number of samples correctly predicted by Model 3 is the largest, reaching 1025, which is higher than 928 of Model 1 and 991 of Model 2. This result means that Model 3 has the highest prediction accuracy. More detailed evaluation metrics are shown in Table 5.

Table 5 shows the evaluation metrics of the three models using the same actual ADR data. Among them, the accuracy of Model 1 is the lowest, only 88.63%. Model 2 has significantly improved its ability to recognize minority classes, with an accuracy rate of 94.65%. The results indicate that Model 3, which uses the combination of feature enhancement (FS_GAN) and SMOTE, has a higher accuracy than Model 2, which only uses SMOTE, reaching 97.90%. The other metrics such as macro average and weighted average also indicate that the performance of Model 3 is the best.

Figure 4 shows the ROC curves and AUC values of the three models. It indicates that Model 1 using only the RF algorithm has the worst classification result for the imbalanced ADR data set, and its AUC value of 0.85 is also the lowest. For the latter two models after SMOTE, the ROC curve of Model 3 with feature enhancement (FS_GAN) is closer to the (0, 1) point. The AUC value of Model 3 is also the highest among them, reaching 0.99.

6. Discussion

The classification results on CFDA's actual ADR data show that the accuracy of Model 1 reaches 88.63%, which seems to be a good result. However, by observing the index of the recall rate of Model 1, we can find that the recall rate of label 1 is 0.36, and the recall rate of label 2 is 0.28. In other words, Model 1 predicts most of the samples as the majority class (label = 0), so it obtains high accuracy. As mentioned in Part 4, such classification is meaningless, because the minority classes are not identified. The latter two models use SMOTE

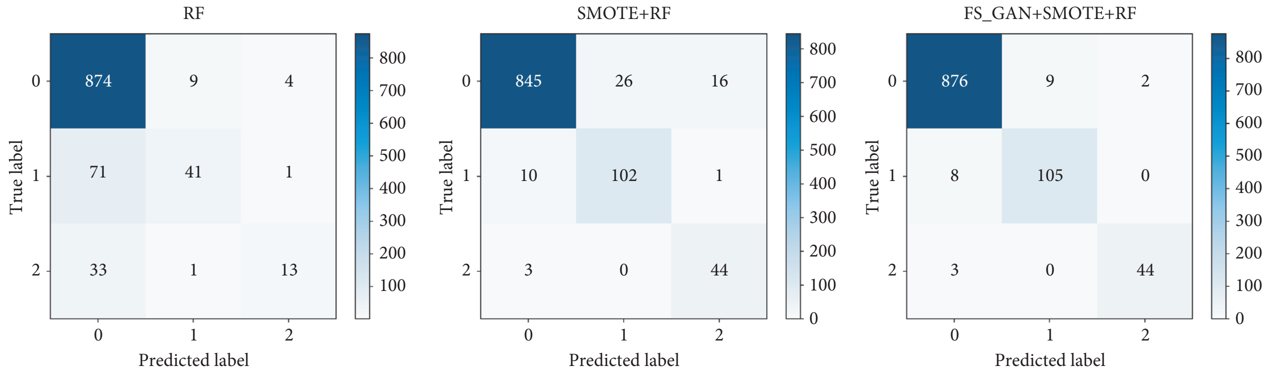


Figure 3: Confusion matrices based on the actual ADR data.

Table 5: Evaluation metrics based on the actual ADR data.

Classifier	Label	Precision	Recall	F1	Accuracy (%)
Model 1 (RF)	0	0.89	0.99	0.94	88.63
	1	0.80	0.36	0.50	
	2	0.72	0.28	0.40	
	Macro-avg	0.81	0.54	0.61	
	Weighted-avg	0.88	0.89	0.87	
Model 2 (SMOTE + RF)	0	0.98	0.95	0.97	94.65
	1	0.80	0.90	0.85	
	2	0.72	0.94	0.81	
	Macro-avg	0.83	0.93	0.88	
	Weighted-avg	0.95	0.95	0.95	
Model 3 (FS_GAN + SMOTE + RF)	0	0.99	0.99	0.99	97.90
	1	0.92	0.93	0.93	
	2	0.96	0.94	0.95	
	Macro-avg	0.96	0.95	0.95	
	Weighted-avg	0.98	0.98	0.98	

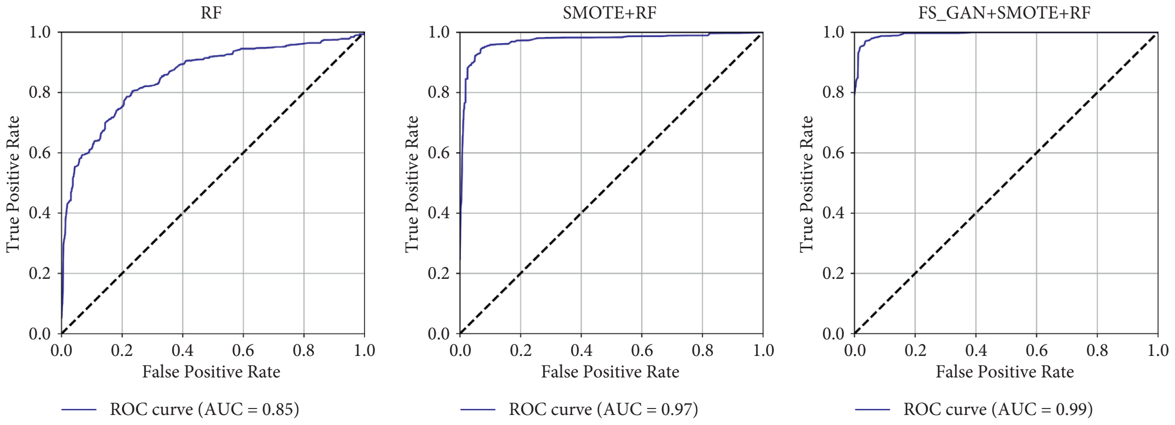


Figure 4: ROC curves and AUC values based on the actual ADR data.

to expand the minority samples, and the number of Rx (label 0), OTC-A (label = 1), and OTC-B (label = 2) drugs reached a balance. Therefore, the recall rate and F1 index are both very high, which indicates that they have a good classification effect on the three categories of drugs.

By comparing the three models, we found that Model 3 is the best, with an accuracy of 97.90%. The Precision, Recall, and F1 index corresponding to the three categories of labels in Model 3 are all higher than Model 2. Especially, for the recognition of minority classes (label = 1 and label = 2), their prediction success rates in Model 3 have been greatly improved.

From the perspective of macro averaging, Model 3 has achieved excellent performance. The macro average takes the arithmetic average of all classes, which means that each class is treated equally during classification, so that the impact of small samples on the results can be more clearly highlighted. The macro-average value of Model 3 is higher than Model 1 and Model 2, so Model 3 is more suitable for the classification of imbalanced samples.

Compared with the macroaverage, the weighted average is more inclined to be affected by the majority class, because the majority category accounts for a larger proportion of the entire samples, and the corresponding weight is also larger. The weighted average of each metric of Model 3 is 0.98, which is the highest among all classifiers.

From the perspective of the ROC curve, the ROC curve of Model 3 is closest to the $(0, 1)$ point among the three, which indicates that Model 3 has the highest classification accuracy rate for imbalanced data sets. This result is confirmed again from the perspective of AUC. The AUC value of Model 3 is 0.99, which is higher than 0.85 of Model 1 and 0.97 of Model 2.

Based on the same CFDA's ADR data, we compared the model proposed in this paper (Model 3) with the model established by our previous work in multiple evaluation indicators. Previously, we compared the prediction results of four machine-learning algorithms, including RF, gradient boost (GB), logistic regression (LR), and AdaBoost (ADA), in the steps of PRR signal detection and SMOTE oversampling, and finally obtained the optimal combination PRR-SMOTE-RF. Through the comparison of experimental results, the accuracy of Model 3 proposed in this paper is 0.98, which is higher than the 0.95 of the previous model PRR-SMOTE-RF. This comparison shows that for ADR samples with obscure features, Model 3 will achieve better prediction results. From the perspective of ROC curves, Model 3 in this paper also has better performance. The AUC value of Model 3 reached 0.99, higher than the 0.97 of PRR-SMOTE-RF, which means that Model 3 has better classification performance for imbalanced data sets. Finally, we can determine that the model with feature enhancement proposed in this paper has better performance on actual ADR data, and has a higher accuracy rate for drug risk prediction.

For the high-dimensional ADR feature space, it is difficult for us to remove redundant features to improve the classification accuracy. The reasons mainly include the following two points. On the one hand, the feature space contains the adverse reactions corresponding to the drugs. If a part of the features that have no effect on the classification are deleted, the potential risks of some drugs may be ignored, leading to deviations in the classification of drugs. For drugs with serious adverse reactions, ignoring their ADR features is fatal, which will cause great harm to patients in the future. On the other hand, additional experiments prove that deleting some redundant features does not improve the classification accuracy very well. We hope to add some effective data that are helpful for classification in the feature space to achieve feature enhancement. In addition, GAN has great advantages in data generation. Through multiple training iterations, GAN can learn about the potential data distribution in the samples and generate similar artificial data. Therefore, when the number of samples is sufficient, GAN-based feature enhancement is an efficient method to solve such problems.

The experimental results prove that it is effective to use feature enhancement technology and minority oversampling at the same time for high-dimensional imbalanced data sets. Compared with the previous PRR-SMOTE-RF framework that does not use feature enhancement, the model proposed in this paper has a higher classification accuracy on the same ADR data set. Other evaluation indicators also confirmed this conclusion. Furthermore, the results indicate that it is effective to use GAN to generate artificial data to improve the overall data distribution in the feature space. In other words, on the basis of minority oversampling of imbalanced data sets, feature enhancement can help achieve more accurate classification. At the same time, this method retains all existing ADR features, thus avoiding the risk evaluation deviation caused by lack of features.

Furthermore, we compare the artificial data generated by GAN with the real data in the ADR data set. Since the data set contains a variety of ADR symptoms, and a drug causes only a small part of the adverse reactions, the data matrix after PRR signal detection is high-dimensional and sparse. The proportion of nonzero elements in the original ADR imbalanced data set is 1.73%. For the top 200 features screened by FS, the proportion of nonzero elements is 5.02%; while for the artificial data generated by GAN, its proportion is 4.85%. This result indicates that artificial data and real data have a high degree of similarity in numerical form. The artificial features generated by GAN satisfy the spatial distribution characteristics of the original data. More specifically, the data distribution of artificial data and real data is similar. Therefore, adding artificial features to the ADR imbalanced data set can improve the sparsity of its feature space. This once again verified that it is feasible to use GAN to achieve feature enhancement.

Through the above analysis, we can draw conclusion that Model 3 (FS_GAN + SMOTE + RF) is more suitable for the prediction of CFDA's spontaneous report data. When choosing this model to evaluate drug risks, we need to conduct further analysis on misclassified drugs. On the one hand, the proposed model has deviation, which can make some medicines misclassified. On the other hand, the adverse reaction corresponding to the drug does not match the class it belongs to, which leads to the wrong classification. In view of the above two situations, experts will reevaluate the misclassified drugs. For drugs that do not match their category, they need to switch among Rx, OTC-A, and OTC-B to control the risks of drugs.

However, this study has several limitations including the following:

(1) Sample size: the research used 985,960 spontaneous reports from 2011 to 2018 provided by CFDA in Jiangsu Province as experimental data, which can visually verify the effectiveness of the proposed model. However, it is difficult to verify the model's evaluation results of drug risks on a larger scale because the sample size is not sufficient. Since Chinese spontaneous report database is not open to the public, we cannot further obtain more updated samples. This results in the limited availability of

relevant data due in part to the high cost of collection of such specialized data. During the preprocessing stage, we deleted a large amount of incomplete and worthless ADR data, which led to a reduction in the sample size. At the same time, the time span and the quality of the SRS are also key factors affecting sample size.

(2) Feature enhancement: in the process of feature selection, we use the relative importance score to rank the ADR features. This selection method may cause some features that have an important impact on the classification to be ranked lower, or even be obscured. We discussed the characteristics of artificial data and real data above, and proved the similarity between the two in terms of data distribution. However, how to further measure the difference between artificial data and real data requires in-depth research in the follow-up work.

(3) Drug interaction: in this study, aspects such as adverse reactions caused by drug interactions are not investigated as these factors are beyond the scope of this research. Potentially, the analysis of ADRs caused by the interaction of different drugs involved in the collected spontaneous reports will help us understand the process of adverse reactions and further clarify the risks of drugs.

In summary, the results of this study indicate that it is feasible to use GAN and SMOTE to classify imbalanced ADR data from CFDA's spontaneous reporting database. This classification can help us understand the applicable population of drugs. Through the evaluation of the classification results, we can further identify the drug risks faced by consumers in a variety of situations, so as to reduce the occurrence of unexpected problems. At the same time, the evaluation of drug risks may help to develop new interventions to deal with adverse reactions after medication.

The main contributions of this study include the accurate classification of actual ADR data, as well as the GAN and SMOTE methods used in this process, in an effort to realize the feature enhancement and minority oversampling. We verify that the model combining PRR, feature enhancement (FS_GAN), SMOTE, and RF classification is optimal for CFDA's spontaneous reporting data, and gives the evaluation metrics suitable for imbalanced data set. This study also provides reference for medical experts on the risk evaluation and status switch of post-marketing drugs.

7. Conclusions

This paper proposes a model combining feature enhancement (FS_GAN) and SMOTE for drug risk evaluation in CFDA's spontaneous reporting data. Based on the comparison of three sets of models, the classification accuracy of the proposed model is nearly 98%. The results suggest that the combination of PRR, FS_GAN, SMOTE, and RF method is determined to be the optimal framework for class-imbalance problems in ADR data. At the same time, the effective features generated by GAN have a significant contribution to the classification performance. This means GAN can be used in more classification scenarios to obtain better results.

This model has the potential to be generalized to more drug regulatory agencies, because it can provide a convenient and reliable way for the ADR signal detection and drug classification. The results will serve as a strong basis for experts to evaluate potential risk of drugs and help them make more judgmatic decisions for the switch of drug status. In the future, it is necessary to pay attention to the adverse reactions caused by the mutual influence of multiple drugs, which will help to further explore the relationship between different ingredients and reduce the risk of medication.

Data Availability

All ADR spontaneous reporting data in this study are licensed by the CFDA. The data sets are not publicly available due to the policy of confidentiality of the CFDA but are available from the corresponding author on reasonable request and with permission of the CFDA.

Conflicts of Interest

The authors declare no conflicts of interest.

Acknowledgments

This study was funded by the National Social Science Foundation of China (no. 17CTQ022), the Major Project of Philosophy and Social Science Research in Jiangsu Universities (no. 2020SJZDA102), and the Postgraduate Research and Practice Innovation Program of Jiangsu Province (no. SJCX20_0223).

References

[1] S. G. Schneeweiss, M. Goettler, and J. Hasford, "Adverse drug events in hospitalized patients," *Jama the Journal of the American Medical Association*, vol. 277, no. 17, pp. 1351-1352, 1998.

[2] I. R. Edwards and J. K. Aronson, "Adverse drug reactions: definitions, diagnosis, and management," *The Lancet*, vol. 356, no. 9237, pp. 1255–1259, 2000.

[3] E. H. Creyer, I. Hrsistodoulakis, and C. A. Cole, "Changing a drug from Rx to OTC status: the consumer behavior and public policy implications of switch drugs," *The Journal of Product and Brand Management*, vol. 10, no. 1, pp. 52–64, 2001.

[4] C. G. Liu, F. Jin, and G. H. Yuan, "The history and expectation of classification management of drugs in China," *Chinese Journal of Pharmacovigilance*, vol. 10, no. 6, pp. 348–351, 2013.

[5] S.-A. Francis, N. Barnett, and M. Denham, "Switching of prescription drugs to over-the-counter status," *Drugs & Aging*, vol. 22, no. 5, pp. 361–370, 2005.

[6] J. A. Rizzo, R. J. Ozminkowski, and R. Z. Goetzel, "Prescription to over-the-counter switching of drugs: methodological issues and implications for non-sedating antihistamines," *Disease Management and Health Outcomes*, vol. 13, no. 13, pp. 83–92, 2005.

[7] E. P. Brass, “Changing the status of drugs from prescription to over-the-counter availability,” *New England Journal of Medicine*, vol. 345, no. 11, pp. 810–816, 2001.

[8] P. Heijden, E. Puijenbroek, S. V. Buuren et al., “On the assessment of adverse drug reactions from spontaneous reporting systems,” *Statistics in Medicine*, vol. 21, no. 14, pp. 2027–2044, 2015.

[9] R. Harpaz, K. Haerian, H. S. Chase, and C. Friedman, “Statistical mining of potential drug interaction adverse effects in FDA’s spontaneous reporting system,” *AMIA. Annual symposium proceedings. AMIA Symposium*, vol. 2010, no. 7, pp. 281–285, 2010.

[10] C. Tantikul, N. Dhana, K. Jongjarearnprasert et al., “The utility of the world health organization-the Uppsala monitoring Centre (WHO-UMC) system for the assessment of adverse drug reactions in hospitalized children,” *Asian Pacific Journal of Allergy and Immunology/Launched by the Allergy and Immunology Society of Thailand*, vol. 26, no. 2-3, pp. 77–82, 2009.

[11] M. Sarangdhar, S. Tabar, C. Schmidt et al., “Data mining differential clinical outcomes associated with drug regimens using adverse event reporting data,” *Nature Biotechnology*, vol. 34, no. 7, pp. 697–700, 2016.

[12] P. M. Coloma, P. Avillach, F. Salvo et al., “A reference standard for evaluation of methods for drug safety signal detection using electronic healthcare record databases,” *Drug Safety*, vol. 36, no. 1, pp. 13–23, 2013.

[13] B. Z. Luo, Y. F. Qian, X. F. Ye et al., “Present status and future prospect of signal detection methods for adverse drug reaction,” *Pharmaceutical Care and Research*, vol. 9, no. 4, pp. 255–260, 2009.

[14] K. Nag and N. R. Pal, “A multiobjective genetic programming-based ensemble for simultaneous feature selection and classification,” *IEEE Trans Cybern*, vol. 46, no. 2, pp. 499–510, 2017.

[15] I. J. Goodfellow, J. Pouget-Abadie, M. Mirza et al., “Generative adversarial networks,” *Advances in Neural Information Processing Systems*, vol. 3, pp. 2672–2680, 2014.

[16] N. V. Chawla, A. Lazarevic, L. O. Hall et al., “SMOTEBoost: improving prediction of the minority class in boosting,” in *Proceedings of the European Conference on Knowledge Discovery in Databases*, Cavtat-Dubrovnik, Croatia, September 2003.

[17] Y. Pouliot, A. P. Chiang, and A. J. Butte, “Predicting adverse drug reactions using publicly available PubChem BioAssay data,” *Clinical Pharmacology & Therapeutics*, vol. 90, no. 1, pp. 90–9, 2011.

[18] V. Santiago, H. Rave, H. S. Chase et al., “Facilitating adverse drug event detection in pharmacovigilance databases using molecular structure similarity: application to rhabdomyolysis,” *Journal of the American Medical Informatics Association Jamia*, vol. 18, 2011.

[19] X. Chen, X. Liu, X. Jia et al., “Network characteristic analysis of adr-related proteins and identification of adr-adr associations,” *Scientific Reports*, vol. 3, 2013.

[20] G. Roberto, C. Piccinni, R. D’Alessandro, and E. Poluzzi, “Triptans and serious adverse vascular events: data mining of the FDA Adverse Event Reporting System database,” *Cephalalgia*, vol. 34, no. 1, pp. 5–13, 2013.

[21] F. Mai, H. Tomoya, H. Kouichi et al., “Association between statin use and cancer: data mining of a spontaneous reporting database and a claims database,” *International Journal of Medical Sciences*, vol. 12, no. 3, pp. 223–233, 2015.

[22] J. H. G. Scholl, F. P. A. M. van Hunsel, E. Hak, and E. P. van Puijenbroek, “A prediction model-based algorithm for computer-assisted database screening of adverse drug reactions in The Netherlands,” *Pharmacoepidemiology and Drug Safety*, vol. 27, no. 2, pp. 199–205, 2018.

[23] I. S. Alimova and E. V. Tutubalina, “Entity-level classification of adverse drug reactions: a comparison of neural network models,” *Proceedings of the Institute for System Programming of the RAS*, vol. 30, no. 5, pp. 177–196, 2018.

[24] A. Martocchia, V. Spuntarelli, F. Aiello et al., “Using INTERCheck® to evaluate the incidence of adverse events and Drug–Drug Interactions in Out- and Inpatients Exposed to Polypharmacy,” *Drugs - Real World Outcomes*, vol. 7, no. 67, 2020.

[25] N. P. Giangreco and N. P. Tatonetti, “Evaluating risk detection methods to uncover ontogenic-mediated adverse drug effect mechanisms in children,” *BioData Mining*, vol. 14, no. 1, 2021.

[26] H. B. Mehta, L. Wang, I. Malagaris et al., “More than two-dozen prescription drug-based risk scores are available for risk adjustment: a systematic review,” *Journal of Clinical Epidemiology*, vol. 137, no. 5, 2021.

[27] C. Ge, Y. Zhang, H. Duan, and H. Li, “Identification of adverse drug events in Chinese clinical narrative text,” *Ubiquitous Computing Application and Wireless Sensor*, vol. 331, pp. 605–612, 2015.

[28] J. Wei, Z. Lu, K. Qiu, P. Li, and H. Sun, “Predicting drug risk level from adverse drug reactions using SMOTE and machine learning approaches,” *IEEE Access*, vol. 8, pp. 185761–185775, 2020.

[29] N. Moore, G. Hall, M. Sturkenboom, R. Mann, R. Lagnaoui, and B. Begaud, “Biases affecting the proportional reporting ratio (PRR) in spontaneous reports pharmacovigilance databases: the example of sertindole,” *Pharmacoepidemiology and Drug Safety*, vol. 12, no. 4, pp. 271–281, 2003.

[30] S. J. W. Evans, P. C. Waller, and S. Davis, “Use of proportional reporting ratios (PRRs) for signal generation from spontaneous adverse drug reaction reports,” *Pharmacoepidemiology and Drug Safety*, vol. 10, no. 6, pp. 483–486, 2001.

[31] D. Georgios and B. Fernando, “Effective data generation for imbalanced learning using conditional generative adversarial networks,” *Expert Systems with Applications*, vol. 91, 2018.

[32] L. Xu, M. Skoularidou, A. Cuesta-Infante, and K. Veeramachaneni, “Modeling tabular data using conditional GAN,” 2019, https://arxiv.org/abs/1907.00503.

[33] T. Maciejewski and J. Stefanowski, “Local neighbourhood extension of smote for mining imbalanced data,” in *Proceedings of the 2011 IEEE Symposium on Computational Intelligence and Data Mining (CIDM)*, Paris, France, April 2011.

[34] R. Blagus and L. Lusa, “SMOTE for high-dimensional class-imbalanced data,” *BMC Bioinformatics*, vol. 14, no. 1, pp. 106–116, 2013.

[35] L. Breiman, “Random forest,” *Machine Learning*, vol. 45, no. 1, pp. 5–32, 2001.

[36] V. F. Rodriguez-Galiano, B. Ghimire, J. Rogan, M. Chica-Olmo, and J. P. Rigol-Sanchez, “An assessment of the effectiveness of a random forest classifier for land-cover classification,” *ISPRS Journal of Photogrammetry and Remote Sensing*, vol. 67, no. Jan, pp. 93–104, 2012.

[37] X. Deng, Q. Liu, Y. Deng, and S. Mahadevan, “An improved method to construct basic probability assignment based on the confusion matrix for classification problem,” *Information Sciences*, vol. 340-341, pp. 250–261, 2016.

[38] G. Hripcsak and A. S. Rothschild, “Agreement, the F-measure, and reliability in information retrieval,” *Journal of the*

American Medical Informatics Association, vol. 12, no. 3, pp. 296–298, 2005.

[39] Y. Yang, "An evaluation of statistical approaches to text categorization," *Proc Amia Annu Fall Symp*, vol. 1, no. 1-2, pp. 358–362, 1999.

[40] G. De Luca and J. R. Magnus, "Bayesian model averaging and weighted-average least squares: equivariance, stability, and numerical issues," *STATA Journal: Promoting communications on statistics and Stata*, vol. 11, no. 4, pp. 518–544, 2012.

[41] O. Komori, "A boosting method for maximization of the area under the ROC curve," *Annals of the Institute of Statistical Mathematics*, vol. 63, no. 5, pp. 961–979, 2011.

Hindawi
Journal of Healthcare Engineering
Volume 2021, Article ID 9963576, 9 pages
https://doi.org/10.1155/2021/9963576

Research Article

Intelligent Disease Prediagnosis Only Based on Symptoms

Fangfang Luo[1] **and Xu Luo**[2]

[1] *School of Nursing, Zunyi Medical University, Zunyi 563000, China*
[2] *Department of Information Engineering, Zunyi Medical University, Zunyi 563000, China*

Correspondence should be addressed to Xu Luo; silyaseln@live.cn

Received 14 March 2021; Accepted 9 July 2021; Published 2 August 2021

Academic Editor: Daniel Espino

People often concern the relationships between symptoms and diseases when seeking medical advices. In this paper, medical data are divided into three copies, records related to main disease categories, records related to subclass disease types, and records of specific diseases firstly; then two disease recognition methods only based on symptoms for the main disease category identification, subclass disease type identification, and specific disease identification are given. In the methods, a neural network and a support vector machine (SVM) algorithms are adopted, respectively. In the method validation part, accuracy of the two diagnosis methods is tested and compared. Results show that automatic disease prediction only based on symptoms is possible for intelligent medical triage and common disease diagnosis.

1. Introduction

At present, there is shortage of per capita medical resources, and high-quality medical resources are concentrated in large cities and large hospitals. In China, many patients have strong health awareness, even if their symptoms are not serious; they also flock to large hospitals to seek quality medical services. Constraints and conflicts between medical resource supply and demand are a long-standing phenomenon.

In medical consultations, what people intuitively care about are the relationships between symptoms and diseases. Nowadays, many people provide symptoms online to obtain prediagnosis results, and their objective is to screen critical illnesses and seek an advice for further accurate medical treatment.

An intelligent information system, which can automatically perform prediagnosis based on the symptoms provided by patients, can alleviate the problem of medical resource shortage. In this paper, such diagnosis methods are proposed. Through these methods, preliminary diagnoses can be provided for specialized diseases, and it can help medical workers in areas having underdeveloped medical resources implement medical triage and provide consultation services for people who will seek precise treatment in big hospitals. Additionally, a common disease diagnosis service can be realized for people who can seek medical treatment by themselves.

2. Related Works

Computer aided diagnosis research has begun since the last century. Most intelligent disease diagnosis researches focus on a certain type disease or only a specific disease. The contents mostly are intelligent diagnosis using machine learning algorithms based on pathology data, influencing factors, examination data, physiological performance, or images when disease types are known previously [1–7]. Some exploratory works have discussed the disease diagnosis only based on the symptoms provided by patients. A simple method is to compare the symptoms provided by a patient to record symptoms in each data item, and the disease in the most similar entry is an output result. In [8], the user gives out feathers related to the diseases such as gender, age, affected part, and related symptoms firstly. Jackcard similarities are calculated based on symptom matrixes, and the similarities are arranged in descending order. Diseases in the first 3 items are selected as alternative recommended answers. In [9], the similarities, which are evaluated by differences between a symptom vector provided by the user and characteristic symptom sets of different diseases, are calculated. The similarities are also arranged in descending order, and the diseases in the first 3 selected items are alternative recommended answers. Disease diagnosis only based on symptoms and without disease type limitation is a general practice (GP) problem. If the above methods are used to solve this kind of diagnosis problem, the efficiency is extremely low, and repetition calculations are involved in each diagnosis case. In related works [10, 11], automatic disease diagnoses based on machine learning algorithms are proposed; in these works, symptoms are extracted firstly, and then, the diagnosis is implemented using deep learning algorithms. There are many diseases, while all proposed methods are limited to discussions on few diseases in the above papers.

Without detailed medical examination data and pathology support, accuracy of diagnosis methods based on symptoms cannot be guaranteed, while, in current online applications, reports, and documents, diagnosis only based on symptoms can be a disease screening method and used to help fast disease type recognition and disease triage in hospital. The key problem is the adaptability of this kind of diagnosis methods. At present, there is no discussion about which disease type levels or which diseases this kind of diagnosis methods is suitable for. To fill this gap, in this paper, this issue is considered.

Disease prediagnosis based on symptoms, which are contained in consultation words, is indeed a text classification problem. In these works, the first step would mostly be lexical feather extraction, and then classification based on different feather properties is implemented [12–14]. Considering the particularity in clinic and immature Chinese word segmentations, in this paper, we only discuss the core prediagnosis problem, and the symptoms, which are also disease feathers, have been extracted according to clinical experience previously. A hierarchical frame is provided in this paper. Firstly, the diseases are divided into major categories and then are divided into several subtypes. Furthermore, specific diseases are filled into subclass disease types. In this paper, two automatic diagnosis methods using a neural network technology and a support vector machine (SVM) technology, respectively, are given to solve this general practice (GP) problem. In the methods, the first is the major disease category identification, and then it is based on the results to identify disease subtypes. Further process is the training for specific disease identification. To observe the effectiveness, the two diagnosis methods are tested and compared.

3. Problem Statement and Theories in This Paper

3.1. The Diagnosis Problem in This Paper. The intelligent diagnosis problem to be solved in this paper includes two aspects. The first one is seeking diagnosis experience according to the relationships between symptoms and diseases. Here, supervised machine learning methods are adopted. The second one is disease prediction based on the symptoms provided by visitors. The first one is the main problem.

In our research, symptoms have been extracted in data preprocessing. Consider that samples with respect to the same disease type are in a hyperplane and linearly separable, and a different symptom may make two similar samples refer to different disease types; the support vector machine (SVM) algorithm is an appropriate method. As the neural network is a generic method in multiclassification problems, this method is also adopted in this paper and compared with SVM [15].

3.2. The Neural Network in This Paper. To describe this method, symbolic notations are given firstly:

N: there are N symptoms in each data item

Γ: the number of nodes in the output layer of a neural network

Y: the number of nodes in the hidden layer of a neural network is Y, and $Y = 10 + \sqrt{N + T}$

$hn = (hn_1, hn_2, hn_3, \ldots, hn_{\mathrm{Y}})$: the input vector of the hidden layer is hn, and the input of the pth hidden layer unit is hn_p

$ho = (ho_1, ho_2, ho_3, \ldots, ho_{\mathrm{Y}})$: the output vector of the hidden layer is ho, and the output of the pth hidden layer unit is ho_{P}

$yn = (yn_1, yn_2, yn_3, \ldots, yn_\Gamma)$: the input vector of the output layer is yn, and the input of the qth output layer unit is yn_q

$yo = (yo_1, yo_2, yo_3, \ldots, yo_\Gamma)$: the output vector of the output layer is yo, and the output of the qth output layer unit is yo_q

w_{np}: the connection weight between the nth input layer unit and the pth hidden layer unit

ϖ_{pq}: the connection weight between the pth hidden layer unit and the qth output layer unit

b_p: the threshold value of the pth hidden layer

b_q: the threshold value of the qth output layer

$x_k = (x_1^{(k)}, x_2^{(k)}, \ldots, x_N^{(k)})$: the kth symptom record, which contains N components, is x_k, and each component represents a different symptom

$d_k = (d_1^{(k)}, d_2^{(k)}, \ldots, d_\Gamma^{(k)})$: the expected output when x_k is input to the neural network is d_k, and if this record is about the r th disease or disease type, the component $d_r^{(k)} = 1$, other components $d_j^{(k)} = 0\,(j \neq r, j \in \{1, 2, 3, \ldots, \Gamma\})$

$o(k) = (x_k, d_k) = ((x_1^{(k)}, x_2^{(k)}, \ldots, x_N^{(k)}), (d_1^{(k)}, d_2^{(k)}, \ldots, d_\Gamma^{(k)}))$: represents the kth training sample

In the neural network, each nerve cell is actually an activation function. For the pth hidden layer unit, if sample $o(k)$ is used, the input is

$$hn_p^{(k)} = \sum_{n=1}^{N} w_{np} x_n^{(k)} - b_p. \tag{1}$$

A sigmoid function is used as the activation function, and the output is

$$h(k)_p^{(k)} = f\left(hn_p^{(k)}\right) = \frac{1}{\left(1 + \exp\left(-hn_p^{(k)}\right)\right)}, \tag{2}$$

where exp() is an exponential function. An output cell of the hidden layer is an input cell of the output layer, and for the qth output layer unit, if sample $o(k)$ is used, the input is

$$yn_q^{(k)} = \sum_{p=1}^{Y} \varpi_{pq} ho_p^{(k)} - \theta_q, \tag{3}$$

and a softmax output is

$$y(k)_q^{(k)} = f_2\left(yn_q^{(k)}\right) = \frac{\exp\left(yn_q^{(k)}\right)}{\sum_{q=1}^{\Gamma} \exp\left(yn_q^{(k)}\right)}. \tag{4}$$

Furthermore, in the neural network, a cross-entropy loss function is adopted:

$$e^{(k)} = f_3\left(yo_q^{(k)}\right) = -\sum_{q=1}^{\Gamma} d_q^{(k)} \ln y(k)_q^{(k)}. \tag{5}$$

In the neural network, some important differential equations are also involved. The first is the partial differential of error function $e^{(k)}$ with respect to ϖ_{pq}, and it is

$$\frac{\partial e^{(k)}}{\partial \varpi_{pq}} = \frac{\partial e^{(k)}}{\partial yn_q^{(k)}} \cdot \frac{\partial yn_q^{(k)}}{\partial \varpi_{pq}}. \tag{6}$$

Considering formula (3) and that the processing procedure is focused on the connection weight between the pth specific hidden layer unit and the qth specific output layer unit, the following formula can be obtained:

$$\frac{\partial yn_q^{(k)}}{\partial \varpi_{pq}} = \frac{\partial\left(\sum_{p=1}^{Y} \varpi_{pq} h(k)_p^{(k)} - \theta_q\right)}{\partial \varpi_{pq}} = ho_p^{(k)}. \tag{7}$$

Further, based on formulas (4) and (5), there is

$$\frac{\partial e^{(k)}}{\partial yn_q^{(k)}} = \frac{\partial e^{(k)}}{\partial yo_q^{(k)}} \cdot \frac{\partial yo_q^{(k)}}{\partial yn_q^{(k)}} = -d_q^{(k)} + y(k)_q^{(k)} \sum_{q=1}^{\Gamma} d_q^{(k)}. \tag{8}$$

Here, this result is marked as $\delta_q^{(k)}$.

If $\delta_q^{(k)}$ is obtained, it can be used to renew the weight between a hidden layer unit and an output layer unit, and the update rule is

$$\varpi_{pq} = \varpi_{pq} + \eta \frac{\partial e^{(k)}}{\partial \varpi_{pq}} = \varpi_{pq} + \eta \delta_q^{(k)} h(k)_p^{(k)}. \tag{9}$$

The connection weight between the pth hidden layer unit and the qth output layer unit in the next training process is the connection weight at present combined with the partial differential $\delta_q^{(k)}$ and output $ho_p^{(k)}$. η is a given learning rate.

In the concrete implementation process, the parameter values of k, p, and q are given in operations with respect to a particular neuron unit.

In the neural network, the partial differential of error function $e^{(k)}$ with respect to w_{np} is also involved, and it is shown as follows:

$$\frac{\partial e^{(k)}}{\partial w_{np}} = \frac{\partial e^{(k)}}{\partial h(k)_p^{(k)}} \cdot \frac{\partial hn_p^{(k)}}{\partial w_{np}}. \tag{10}$$

Similarly, considering formula (2) and that the processing procedure is focused on the connection weight between the nth specific input layer unit and the pth specific hidden layer unit, the following formula can be obtained:

$$\frac{\partial hn_p^{(k)}}{\partial w_{np}} = \frac{\partial\left(\sum_{i=1}^{N} w_{np} x_n^{(k)} - b_p\right)}{\partial w_{np}} = x_n^{(k)}. \tag{11}$$

Further, based on formulas (2)–(5), there is

$$\begin{aligned} \frac{\partial e^{(k)}}{\partial hn_p^{(k)}} &= \frac{\partial e^{(k)}}{\partial yn_q^{(k)}} \cdot \frac{\partial yn_q^{(k)}}{\partial ho_p^{(k)}} \cdot \frac{\partial ho_p^{(k)}}{\partial hn_p^{(k)}} = -\left(\sum_{q=1}^{\Gamma} \delta_q^{(k)} \varpi_{pq}\right) f_1'\left(hn_p^{(k)}\right), \\ &= -\left(\sum_{q=1}^{\Gamma} \delta_q^{(k)} \varpi_{pq}\right) \exp\left(-hn_p^{(k)}\right)\left(1 + \exp\left(-hn_p^{(k)}\right)\right)^{-2}. \end{aligned} \tag{12}$$

Here, this result is marked as $\sigma_p^{(k)}$.

If $\sigma_p^{(k)}$ is obtained, it can be used to renew the weight between a hidden layer unit and an output layer unit, and the update rule is

$$w_{np} = w_{np} + \eta \frac{\partial e^{(k)}}{\partial w_{np}} = w_{np} + \eta \sigma_p^{(k)} x_n^{(k)}. \tag{13}$$

The connection weight between the nth hidden layer unit and the pth output layer unit in the next training process is the connection weight at present combined with the partial differential $\sigma_p^{(k)}$ and input $x_n^{(k)}$. η is also a given learning rate.

3.3. The Support Vector Machine (SVM) in This Paper.

In this paper, a disease sample is $x^{(k)} = (x_1^{(k)}, x_2^{(k)}, \ldots, x_N^{(k)}, x_{N+1}^{(k)})$, where $(x_1^{(k)}, x_2^{(k)}, \ldots, x_N^{(k)})$ represents different symptoms and $x_{N+1}^{(k)}$ is a disease or disease type.

The hyperplane separating samples are depicted as follows:

$$f(x) = \boldsymbol{\omega}^T x + c. \tag{14}$$

The purpose is to get the classification parameters $\boldsymbol{\omega}$ and c. If there are $T(b)$ samples in the sample space, the specific problem should be solved:

$$\begin{aligned} &\min_{(\omega,c)} \frac{1}{2}\|\boldsymbol{\omega}\|^2, \\ &\text{s.t. } x_{N+1}^{(k)}\left(\boldsymbol{\omega}\left[x_1^{(k)}, \ldots, x_N^{(k)}\right]^T + c\right) \geq 1, \quad k = 1, 2, \ldots, T(b). \end{aligned} \tag{15}$$

If the classification parameters have been obtained, and there is a symptom vector $\mathbf{k} = (\kappa_1, \kappa_2, \ldots, \kappa_N)$, while

$$\boldsymbol{\omega\kappa} + c = \omega_1\kappa_1 + \omega_2\kappa_2 + \cdots + \omega_N\kappa_N + c > 0, \tag{16}$$

it can be determined that κ belongs to the disease category I, while

$$\boldsymbol{\omega\kappa} + c = \omega_1\kappa_1 + \omega_2\kappa_2 + \cdots + \omega_N\kappa_N + c < 0, \tag{17}$$

and it can be determined that $\boldsymbol{\kappa}$ does not belong to the disease category I.

In learning procedures, a one-against-the-rest SVM method [16] based on this basic form can be adopted to implement multiclassification.

4. Disease Identification Methods

4.1. Preconditions. Suppose that a preprocess step has been implemented on existing electronic medical records. Disease symptoms, disease types, and relations between the two are known clearly.

4.2. Labelling. Number the N disease symptoms in the database, and the symptoms are numbered as $1, 2, 3, 4, \ldots, N$, respectively. Considering that the same symptoms in different gender patients are often with regard to different common diseases or disease types, gender is deemed as a default "symptom," which is labelled 1. Diseases in the database are divided into B main categories, which are numbered as $N * 10 + 1$, $N * 10 + 2$, $\ldots$, $N * 10 + B$. Each main disease category is further divided into several subclasses and numbered. There are $T(b)$ subtype diseases under the main disease category $N * 10 + b$, and they are numbered as $N * 10 + b + 1$, $N * 10 + b + 2, \ldots, N * 10 + b + T(b)$, $b = 1, 2, 3, \ldots, B$. $T^{(b,j)}$ diseases are related to the disease type $(N * 10 + b) * 10 + j$ and numbered as $((N * 10 + b) * 10 + j) * 10 + 1$, $((N * 10 + b) * 10 + j) * 10 + 2, \ldots$, $((N * 10 + b) * 10 + j) * 10 + T^{(b,j)}$, $b = 1, 2, 3, \ldots, B$, $j = 1, 2, 3, \ldots, T(b)$.

Establish a data relationship list, in which the data structure is (Symptom 1, Symptom 2, Symptom 3, …, Symptom N, Disease). Each entry contains N symptoms. If symptom n does exist in the item of a disease, the value below "Symptom n " is 1, or else, the value is 0.

For example, suppose that there are only $N = 11$ symptoms in the current medical study records, the symptoms are male, fever, ulcer, pain, aching and limp, nasal congestion, diarrhea, bleeding, tumor, drowsiness, and face yellowing, and the label values of these symptoms are 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, and 11. There are only $B = 10$ major disease categories in the studied medical records, tumor disease, infectious disease, blood disease, cardiovascular disease, digestive disease, endocrine system disease, respiratory disease, urinary system disease, ophthalmic disease, and otolaryngology disease, and labelled numbers of these disease types are $101\,(N * 10 + 1)$, $102\,(N * 10 + 2)$, $103\,(N * 10 + 3)$, $104\,(N * 10 + 4)$, $105\,(N * 10 + 5)$, $106\,(N * 10 + 6)$, $107\,(N * 10 + 7)$, $108\,(N * 10 + 8)$, $109\,(N * 10 +9)$, $110\,(N * 10 + 10)$, respectively. Furthermore, suppose that there are $T(1) = 3$ subtype diseases, benign tumor $1011\,((N * 10 + 1) * 10 + 1)$, borderline tumor $1012\,((N * 10 + 1) * 10 + 2)$, and malignant tumor $1013\,((N * 10 + 1) * 10 + 3)$ in tumor diseases. And $T^{(1,1)} = 5$ diseases, which are squamous cell carcinoma $10111\,(((N * 10 + 1) * 10 + 1) * 10 + 1)$, adenocarcinoma $10112\,(((N * 10 + 1) * 10 + 1) * 10 + 2)$, basal cell carcinoma $10113\,(((N * 10 + 1) * 10 + 1) * 10 + 3)$, transitional cell carcinoma $10114\,(((N * 10 + 1) * 10 + 1) * 10 + 4)$ and sarcoma $10115\,(((N * 10 + 1) * 10 + 1) * 10 + 5)$, are in the benign tumor disease. If there is a medical record about the squamous cell carcinoma disease, and the symptoms are fever, ulcers, pain, and tumor, there are three data items that are related to this case and shown in Table 1.

A BP neural network that is shown in Figure 1 is used for the disease type and specific disease identification. There are N input layer nodes, Γ output layer nodes, and $Y = 10 + \sqrt{N + \Gamma}$ hidden layer nodes. K training symptom samples $o(k) = (x_k, d_k)$, $k = 1, 2, 3, \ldots, K$ are known. One medical record is related to a sample. When symptom $x_n^{(k)}$ appears in the record $x^{(k)}$, $x_n^{(k)} = 1$, or else $x_n^{(k)} = 0$. When a medical record is about the disease type $d_\varsigma^{(k)}$, $d_\varsigma^{(k)} = 1$, and the rest items are zero, that is, $d_{q\neq\varsigma}^{(k)} = 0$. The $(N + 1)$ th input layer unit with an input value "−1" and the $(Y + 1)$th hidden layer unit also with an input value "−1" are used to generate threshold values, and connection weights $\omega_{(N+1)p}$ and $\bar{\omega}_{(Y+1)q}$ are used as thresholds b_p and θ_q, respectively.

Specific training procedures are implemented according to formulas (1)–(13) in Section 3.2. Based on the data form in Table 1, the value of $x_n^{(k)}$ is 0 or 1, $n = 1, 2, \ldots, N$. If $d_j^{(k)}$ is in $\{(N * 10 + 1), (N * 10 + 2), \ldots, (N * 10 + B)\}$, it is the training procedure to identify main disease categories. An identification neural network NT is obtained. If $d_j^{(k)}$ is in $\{(N * 10 + b) * 10 + 1, (N * 10 + b) * 10 + 2, \ldots, (N * 10+ b) * 10 + T(b)\}$, it is the training procedure to identify subclass disease types under the main disease category $N * 10 + b$. Identification neural networks $\mathrm{NT} - b$, $b = 1, 2, 3, \ldots, B$ are obtained. If $d_j^{(k)}$ is in $\{((N * 10+ b) * 10 + j) * 10 + 1, ((N * 10 + b) * 10 + j) * 10 + 2, \ldots, ((N * 10 + b) * 10 + j) * 10 + T\ (b, j)\}$, it is the training procedure to identify specific diseases under the subclass disease type $(N * 10 + b) * 10 + j$. Identification neural networks $\mathrm{NT} - (b, j)$, $b = 1, 2, 3, \ldots, B$, $j = 1, 2, 3, \ldots, T(b)$ are obtained.

While the SVM method mentioned in Section 3.3 is used, classification parameters with respect to major disease types

$$(\boldsymbol{\omega}, c)^1, (\boldsymbol{\omega}, c)^2, \ldots, (\boldsymbol{\omega}, c)^B, \tag{18}$$

classification parameters with respect to subcategory disease types

Table 1: An example of disease records.

Sequence number	Gender	Fever	Ulcer	Pain	Aching and limp	Nasal congestion	Diarrhea	Bleeding	Tumor	Drowsiness	Face yellowing	Disease
1	1	1	1	1	0	0	0	0	1	0	0	101
2	1	1	1	1	0	0	0	0	1	0	0	1013
3	1	1	1	1	0	0	0	0	1	0	0	10131

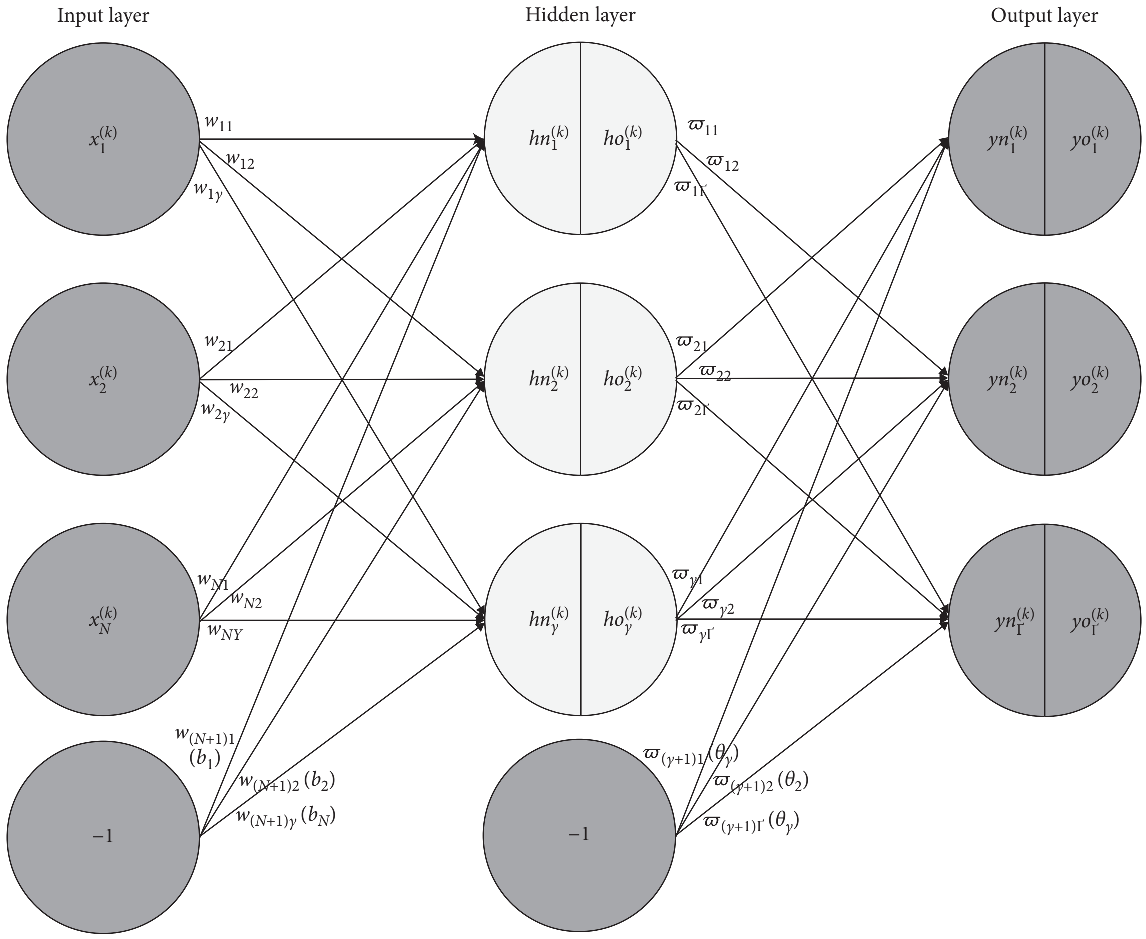


Figure 1: BP neural network in this paper.

$$\begin{aligned}(\boldsymbol{\omega}, c)^{(1,1)}, (\boldsymbol{\omega}, c)^{(1,2)}, \ldots, (\boldsymbol{\omega}, c)^{(1,T(1))}, (\boldsymbol{\omega}, c)^{(2,1)}, (\boldsymbol{\omega}, c)^{(2,2)}, \ldots, (\boldsymbol{\omega}, c)^{(2,T(2))},\\ \cdots\\ (\boldsymbol{\omega}, c)^{(B,1)}, (\boldsymbol{\omega}, c)^{(B,2)}, \ldots, (\boldsymbol{\omega}, c)^{(B,T(B))}.\end{aligned} \tag{19}$$

and classification parameters with respect to specific diseases

$$\begin{aligned}(\boldsymbol{\omega}, c)^{(1,1,1)}, (\boldsymbol{\omega}, c)^{(1,1,2)}, \ldots, (\boldsymbol{\omega}, c)^{\left(1,1,T^{(1,1)}\right)}, (\boldsymbol{\omega}, c)^{(1,2,1)}, (\boldsymbol{\omega}, c)^{(1,2,2)}, \ldots, (\boldsymbol{\omega}, c)^{\left(1,2,T^{(1,2)}\right)},\\ \cdots\\ (\boldsymbol{\omega}, c)^{(B,T(B),1)}, (\boldsymbol{\omega}, c)^{(B,T(B),2)}, \ldots, (\boldsymbol{\omega}, c)^{\left(B,T(B),T^{(B,T(B))}\right)},\end{aligned} \tag{20}$$

can be obtained.

5. Diagnosis Implementations

5.1. Identification of Main Disease Categories. The symptoms, which are provided by a patient, are $\boldsymbol{\kappa} = (\kappa_1, \kappa_2, \ldots, \kappa_N)$.

(1) Identification based on the neutral network: put $\boldsymbol{\kappa}$ into the neutral network NT to estimate which main disease category the symptoms refer to

(2) Identification based on SVM: identify whether the disease category is based on vectors $(\boldsymbol{\omega}, c)^b$, $b = 1, 2, 3, \ldots, B$ by SVM

5.2. Identification of Subclass Disease Types. If the main disease category is $b = \varsigma$ and the symptoms provided by a patient are $\boldsymbol{\kappa} = (\kappa_1, \kappa_2, \ldots, \kappa_N)$, based on the neutral network $\mathrm{NT} - \varsigma$ and SVM identification parameters $(\boldsymbol{\omega}, c)^{(\varsigma,\tau)}$, $\tau = 1, 2, 3, \ldots, T(\varsigma)$ to identify subclass disease types.

(1) Subclass disease type identification based on the neutral network

Put $\boldsymbol{\kappa} = (\kappa_1, \kappa_2, \ldots, \kappa_N)$ into the neutral network $\mathrm{NT} - \varsigma$ to estimate which subclass disease type the symptoms refer to.

(2) Subclass disease type identification based on SVM

Step 1: Initial value is $\tau = 1$.

Step 2: Identify whether the disease type is τ based on vector $(\boldsymbol{\omega}, c)^{(\varsigma,\tau)}$ in the SVM classification method. If the disease type is τ, go to Step 3, or else, make $\tau = \tau + 1$. Verify that whether $\tau > T(\varsigma)$, and if it is, quit out the whole procedure, or else loop through Step 2.

Step 3: The subclass disease type τ is the output result.

5.3. Identification of Specific Diseases. Suppose that the main disease category is $b = \varsigma$ and the subclass type is $j = \tau$. Based on the neutral network $\mathrm{NT} - (\varsigma, \tau)$ and SVM identification parameters $(\boldsymbol{\omega}, c)^{(\varsigma,\tau,\upsilon)}$, $\upsilon = 1, 2, 3, \ldots, T^{(\varsigma,\tau)}$ to identify specific diseases.

(1) Disease identification based on the neural network

Put $\boldsymbol{\kappa} = (\kappa_1, \kappa_2, \ldots, \kappa_N)$ into the neutral network $\mathrm{NT} - (\varsigma, \tau)$ to estimate what disease it is.

(2) Disease identification based on SVM

Step 1: Initial value is $\upsilon = 1$.

Step 2: Identify whether the disease is υ based on vector $(\boldsymbol{\omega}, c)^{(\varsigma,\tau,\upsilon)}$ in the SVM classification method. If the disease is υ, go to Step 3, or else make $\upsilon = \upsilon + 1$. Verify that whether $\upsilon > T^{(\varsigma,\tau)}$, and if it is, quit out the whole procedure, or else loop through Step 2.

Step 3: Disease υ is the output result.

6. Method Tests

In this part, the diagnosis methods are tested. The tests in this paper are implemented in digestive diseases, respiratory diseases, and urinary diseases and used as examples.

6.1. Leave-One-Out Cross Validation. The neural network disease identification method and the support vector machine (SVM) disease identification method are compared.

Example 1. If a test sample is given, distinguish it as a digestive disease, a respiratory disease, or a urinary disease. Test results are shown in Table 2.

In Table 2, the accuracy of the main disease category identification is 94.4%.

Disease triage is to estimate which subclass disease type consulting symptoms provided by the user refer to. Disease triage is tested in the following examples.

Example 2. Tests are implemented in cases. Case 1: if a test sample has been diagnosed as a respiratory disease, distinguish it as a pulmonary disease, a respiratory tract infection, a chest disease, or a mediastinal disease. Case 2: if a test sample has been diagnosed as a digestive disease, distinguish it as an intestinal disease, a hepatic and gall disease, an epityphlon and pancreas disease, or a stomach disease. Case 3: If a test sample has been diagnosed as a urinary system disease, distinguish it as a bladder disease, a kidney disease, or an ureteral disease. The results are shown in Table 3.

In Table 3, the accuracy in disease subtype identification is higher than 80%, but lower than the accuracy in the main disease category identification.

Specific disease identification tests are carried on in Example 3 and Example 4. In Example 3, binary classification tests are executed. Samples about a disease are one class, samples not related to this disease are “the other” one. Example 4 is a multiclassification test, and samples related to different diseases are different categories.

Example 3. Tests are implemented in such cases. Case 1: Gastritis identification in stomach diseases; Case 2: Duodenal ulcer identification in stomach diseases. Case 3: Common cold identification in respiratory tract infections. Case 4: Pharyngitis disease identification in respiratory tract diseases; Case 5: Asthma identification in respiratory tract diseases. Case 6: Pneumonia identification in pulmonary diseases. Case 7: Pulmonary tuberculosis identification in pulmonary diseases. Case 8: Enteritis identification in intestinal diseases. Case 9: Intestinal obstruction identification in intestinal diseases. Case 10: Hepatitis identification in hepatic and gall diseases. Case 11: Gallstone identification in hepatic and gall diseases. Test results are shown in Table 4.

Table 2: Diagnosis of main disease categories.

Number of test samples	Wrong identified samples	Methods
169	9	SVM
169	9	Neural network

Table 3: Disease triage.

Cases	Number of test samples	Wrong identified samples	Methods
Case 1	76	15	SVM
	76	15	Neural network
Case 2	60	12	SVM
	60	11	Neural network
Case 3	32	4	SVM
	32	4	Neural network

Table 4: Diagnosis of specific diseases in binary classifications.

Cases	Accuracy (%)	Methods
Case 1	80.48	SVM
	82.93	Neural network
Case 2	85.36	SVM
	85.36	Neural network
Case 3	91.43	SVM
	88.57	Neural network
Case 4	97.14	SVM
	97.14	Neural network
Case 5	88.00	SVM
	88.00	Neural network
Case 6	80.00	SVM
	84.00	Neural network
Case 7	80.00	SVM
	80.00	Neural network
Case 8	86.00	SVM
	86.00	Neural network
Case 9	88.00	SVM
	90.00	Neural network
Case 10	95.00	SVM
	93.33	Neural network
Case 11	83.33	SVM
	83.33	Neural network

Example 4. Tests are implemented in such cases: Case 1: Gastritis, upper gastrointestinal bleeding, duodenal ulcer, and gastric ulcer identifications in stomach diseases; Case 2: Intestinal obstruction, intussusception, ulcerative colitis, common enteritis, and lower gastrointestinal bleeding identifications in intestinal diseases; Case 3: Viral hepatitis, cholangitis, gallstones, cholecystitis, liver abscess, and cirrhosis identifications in hepatic and gall diseases; Case 4: Pneumonia, emphysema, lung abscess, pulmonary thrombosis, and tuberculosis identifications in pulmonary diseases; Case 5: Upper respiratory tract infection and lower respiratory tract infection identifications in respiratory tract infections; Case 6: Renal failure, glomerulonephritis, pyelonephritis, kidney stones, and nephrotic syndrome identifications in kidney diseases. Test results are shown in Table 5.

Comparing the results in Tables 4 and 5, if the specific disease diagnosis is put into binary classifications, the accuracy is higher than 80%, and when it is put into multiclassification modules, the results are unsatisfactory. Without the support of detailed pathology data, specialized diseases actually cannot be accurately diagnosed by methods only based on symptoms. However, considering the result supports in Table 4, identifications of common diseases such as gastritis, common cold, pharyngitis, and common enteritis, which always do not need the support of detailed pathology data, can be provided to the user in an automatic disease diagnosis system.

6.2. Diagnosis with Weight Samples. In clinic, some diseases have high relational discrepancy symptoms. In a common disease diagnosis experiment, which we have carried out, sample weights are assigned to some samples artificially according to clinical experience, and these weights are added into loss functions in machine learning procedures [17]. In the test, binary classification results using samples with weights and without weights are similar, and the precision difference is less than 4%. Thus, high relational discrepancy degree samples are suggested to be put into test sample sets in validation procedures of machine learning methods.

6.3. Multitype Diseases Diagnosis. A person may have more than 1 disease, and these diseases refer to different types, and results also can be obtained when $\boldsymbol{\kappa} = (\kappa_1, \kappa_2, \ldots, \kappa_N)$ is put into classification modules identifying 1, 2, 3, ..., N concurrence diseases successively.

Example 5. Suppose that a patient has two or three diseases, and these diseases belong to different disease subtypes. Case 1: If the diseases belong to digestive diseases, identify it as a concurrence case of intestinal disease, and hepatic and gall disease, a concurrence case of intestinal disease and stomach disease, or a concurrence case of stomach disease, and hepatic

Table 5: Diagnosis of specific diseases in multiple classifications.

Cases	Accuracy (%)	Methods
Case 1	41.46	SVM
	34.15	Neural network
Case 2	78.00	SVM
	70.00	Neural network
Case 3	66.67	SVM
	63.33	Neural network
Case 4	60.00	SVM
	56.00	Neural network
Case 5	76.00	SVM
	70.00	Neural network
Case 6	70.00	SVM
	60.00	Neural network

Table 6: Diagnosis of multiple diseases.

Cases	Number of test samples	Wrong identified samples	Methods
Case 1	150	54	SVM
	150	52	Neural network
Case 2	150	36	SVM
	150	36	Neural network
Case 3	180	135	SVM
	180	133	Neural network
Case 4	200	134	SVM
	200	136	Neural network

and gall disease. Case 2: If the diseases belong to respiratory diseases, identify it as a concurrence case of pulmonary disease and chest disease, a concurrence case of chest disease and respiratory tract infection, or a concurrence case of pulmonary disease and respiratory tract infection. Case 3: If the diseases belong to respiratory diseases, identify it as a concurrence case of pulmonary disease, upper respiratory tract disease, and trachea and bronchi disease, a concurrence case of pulmonary disease, upper respiratory tract disease, and pleura and chest disease, or a concurrence case of pulmonary disease, trachea and bronchi disease, and pleura and chest disease. Case 4: If the diseases belong to digestive diseases, identify it as a concurrence case of intestinal disease, hepatic and gall disease, and epityphlon and pancreas disease, a concurrence case of stomach disease, intestinal disease, and hepatic and gall disease, or a concurrence case of epityphlon and pancreas disease, stomach disease, and intestinal disease. The above cases are about disease subtype identifications, and the results are shown in Table 6.

From the results in Table 6, it can be seen that, in the identification of concurrence of multiple disease types, the accuracy of machine learning methods is dropping. When there is a concurrence of more than three disease types, the identification accuracy would be much lower.

6.4. Discussion on Test Results

(1) For lacking pathologic support, the accuracy of the GP diagnosis methods based on symptoms for specific diseases is limited. In our tests, it is shown that this kind of methods can be used in the diagnosis of common diseases, such as cold, enteritis, and rhinitis, and for specialized diseases such as asthma, liver cancer, and psoriasis, these methods can be used to predict disease types and provide disease triage. Diagnosis methods, which identify disease types in this paper, can also be used in hospital guides.

(2) In consideration of sample characteristics, the neutral network and SVM machine learning methods are appropriate choices for the automatic prediagnosis problem in this paper. In our experiments, the accuracy of the neural network is close to that of SVM. Sometimes, the neutral network performs a litter better, and sometimes, it is the SVM. A corollary is that the accuracy of this kind of diagnosis methods is limited by the problem itself, and even another practicable machine learning method is adopted, and the performance is also similar with the neutral network and SVM method.

(3) From the experiment results, it can be seen that automatic prediagnosis methods only based on symptom data are suitable for single disease type identification, and it is also not difficult to infer that these methods are also only suitable for a specific common disease identification. If a symptom record is related to multiple disease types or multitype diseases, the availability is low.

(4) In our experiments, the feasibilities of diagnosis only based on symptoms using machine learning methods are tested. Even tests are carried out in digestive diseases, respiratory diseases, and urinary diseases, and without loss of generality, it can be deduced that this kind of diagnosis methods can be used in other disease categories. Test results would also be observed further in more kinds of disease types except for the cases in this paper.

7. Conclusions

In this paper, neural network and SVM machine learning methods are given to solve the automatic disease diagnosis problem only based on symptoms. In our methods, each symptom is a feature. The methods work in three layers, which are main disease category identification, subclass disease type identification, and specific disease identification. The methods are suitable for the diagnosis of common diseases and disease triage for specialized diseases. The availability in practice is proved and analyzed in the experiments of this paper. In addition, future research is also required to investigate automatic symptom extraction and discuss the maximum number size of symptoms.

Data Availability

The data used to support the findings of this study are included within the supplementary information file.

Disclosure

The funding sponsors had no role in the design of the study; in the collection, analyses, or interpretation of the data; in the writing of the manuscript; and in the decision to publish the results.

Conflicts of Interest

The authors declare that there are no conflicts of interest regarding the publication of this article.

Acknowledgments

This work was supported by National Natural Science Foundation of China under Grant 61861047, Science and Technology Program of Zunyi City under Grant HZ(2019) 22, and Innovation and Entrepreneurship Fund of Zunyi Medical University under Grant ZYDC2019118.

Supplementary Materials

Data file “prototype data.pdf.” (*Supplementary Materials*)

References

[1] K. Vanisree and J. Singaraju, “Decision support system for congenital heart disease diagnosis based on signs and symptoms using neural networks,” *International Journal of Computer Application*, vol. 19, no. 6, pp. 6–12, 2011.

[2] J. Shiraishi, Q. Li, D. Appelbaum, and K. Doi, “Computer-aided diagnosis and artificial intelligence in clinical imaging,” *Seminars in Nuclear Medicine*, vol. 41, no. 6, pp. 449–462, 2011.

[3] W. Gao, H. Liang, W. Zhong, and J. Lv, “Differential diagnosis of neonatal necrotizing enter colitis based on machine learning,” *China Digital Medicine*, vol. 14, no. 3, pp. 50–52, 2019.

[4] K. Matjaz, K. Igor, G. Ciril, K. Katarina, and F. Jure, “Analyzing and improving the diagnosis of ischaemic heart disease with machine learning,” *Artificial Intelligence in Medicine*, vol. 16, no. 1, pp. 25–50, 1999.

[5] K. Igor, “Machine learning for medical diagnosis: history, state of the art and perspective,” *Artificial Intelligence in Medicine*, vol. 23, no. 1, pp. 89–109, 2001.

[6] M. Manish, D. Damini, S. Daniel et al., “Machine learning for prediction of all-cause mortality in patients with suspected coronary artery disease: a 5-year multicenter prospective registry analysis,” *European Heart Journal*, vol. 38, no. 7, pp. 500–507, 2017.

[7] D. lin, A. V. Vasilakos, Y. Tang, and Y. Yao, “Neural networks for computer-aided diagnosis in medicine: a review,” *Neurocomputing*, vol. 216, no. 5, pp. 700–708, 2016.

[8] J. Wang, C. Su, and T. Ren, “Design and realization of the intelligent hospital guide,” *Journal of Medical Informatics*, vol. 39, no. 8, pp. 29–32, 2018.

[9] X. Luo, Y. Chang, and J. Yang, “An automatic disease diagnosis method based on big medical data,” in *Proceedings Of the 2015 International Conference On Information Science And Security*, pp. 252–254, IEEE, Seoul, Korea(South), December 2015.

[10] R. Xu, *“The Research of Intelligence Auxiliary Disease Guidance Based on Text Mining Technology,” Master Dissertation*, Beijing University of Posts and Telecommunications, Beijing, China, 2015.

[11] C. Li, *“Research and application on intelligent disease guidance and medical question answering method”*, Dalian University of Technology, Dalian, China, 2016.

[12] A. Onan, S. Korukoğlu, and H. Bulut, “Ensemble of keyword extraction methods and classifiers in text classification,” *Expert Systems with Applications*, vol. 57, pp. 232–247, 2016.

[13] A. Onan and S. Korukoğlu, “A feature selection model based on genetic rank aggregation for text sentiment classification,” *Journal of Information Science*, vol. 43, no. 1, pp. 25–38, 2017.

[14] A. Onan, “Two-stage topic extraction model for bibliometric data analysis based on word embeddings and clustering,” *IEEE Access*, vol. 7, pp. 145614–145633, 2019.

[15] Z. Zhou, *Machine Learning*, Tsinghua University Press, Beijing, China, 2016.

[16] Z. Liu, D. Li, Q. Qin, and W. Shi, “An anlytical overview of methods for multi-category support vector machine,” *Computer Engineering and Applications*, vol. 46, no. 7, pp. 10–13+65, 2004.

[17] F. Pedregosa, G. Varoquaux, A. Gramfort et al., “Scikit-learn: machine learning in Python,” *Journal of Machine Learning Research*, vol. 12, no. 85, pp. 2825–2830, 2011.