

Received December 3, 2020, accepted December 15, 2020, date of publication December 28, 2020, date of current version January 7, 2021.

Digital Object Identifier 10.1109/ACCESS.2020.3047683

On-Site Identification of Counterfeit Drugs Based on Near-Infrared Spectroscopy Siamese-Network Modeling

ZHENG AN-BING¹, YANG HUI-HUA^{1,2}, PAN XI-PENG², YIN LI-HUI³, AND FENG YAN-CHUN³

¹School of Artificial Intelligence, Beijing University of Posts and Telecommunications, Beijing 100876, China

²School of Computer Science and Information Security, Guilin University of Electronic Technology, Guilin 541004, China

³China Institute for Food and Drug Control, Beijing 100050, China

Corresponding author: Yang Hui-Hua (yhh@bupt.edu.cn)

This work was supported in part by the National Natural Science Foundation of China under Grant 61906050 and Grant 21365008, and in part by the Guangxi Technology Research and Development Program under Grant 2018AD11018.


ABSTRACT Near-infrared spectroscopy (NIR) has become one of the most important methods for counterfeit drugs identification for its low cost, non-destructive, and on-site detection. However, it is often invalid for unknown samples beyond the scope of modeling samples, as well as it is not efficient in conditions of insufficient samples (insufficient number of samples within the class), unbalanced samples (large difference in the number of samples between classes), and sensitivity of identification results (different tolerance for different errors in application scenarios). To solve these problems, this paper proposes a general method for on-site identification of counterfeit drugs based on Siamese-network modeling with near-infrared spectroscopy, which especially constructs the train set and test set, learning the general knowledge of spectral differences to identify the different drugs by a costumed one-dimensional convolution neural network (1D-CNN), and finally answered the question of whether the on-site two spectra are pointed to the same drug. Based on experimental modeling samples of 1314 spectra, which are involved 9 drugs produced by 25 manufacturers, this paper has constructed and fully trained its model. Then, not known at the time of modeling, 4 drugs produced by 9 manufacturers are used for testing in the on-site application, and the accuracy rate amounts to 97.3%. For generalizing consideration, randomly divided into training and testing categories, the 32015 spectra of 135 drugs produced by 391 manufacturers in the spectral library are handled by the same processing. The generalization model is equally applicable, and the accuracy is above 97%. Compared with traditional binary classification identification methods such as SVM, PLS-DA, Auto-encoding, and one class (OC) threshold identification algorithms such as SVM-OC, SIMCA, conformity test (CT), the proposed method has the best identification ability for unknown samples in modeling.

INDEX TERMS 1D CNN, drug identification, near-infrared spectroscopy, Siamese-network.

I. INTRODUCTION

The dangers of counterfeit drugs are unquestionable. For drug supervision, the most direct and powerful means to crack down on the manufacture and sale of counterfeit drugs is to quickly and accurately identify the true and false drugs on the spot. Therefore, on-site drug detection technology has special significance.

The near-infrared spectroscopy (NIR) and its modeling, analysis technology, which is often used in rapid on-site

The associate editor coordinating the review of this manuscript and approving it for publication was Ravibabu Mulaveesala .

drug inspection, have the advantages of low instrument cost, non-destructive detection, on-site detection, etc. [1]–[4], and are suitable for rapid qualitative and quantitative analysis of the organic matter. However, these methods relied heavily on background data modeling, and models built by traditional modeling methods are often not valid for unknown samples beyond the scope of categories of the modeling phase.

There are just so many kinds of drugs that market regulators could not know exactly which drugs counterfeiters are counterfeiting. Even if it is known that a popular drug will be counterfeited, it is still unknown what methods and means the counterfeiter will use to counterfeit it. Background

database samples can not cover all the conditions, and test samples often exceed the category range. To make it worse, when modeling near-infrared spectral data with traditional methods, the problems of insufficient samples (insufficient number of samples within the class), unbalanced samples (large difference in the number of samples between classes), and cost sensitivity of identification results (different tolerance for different errors in application scenarios) often become big challenges or obstacles.

In on-site detection, immediately giving the identification results is often necessary, and large-scale, long-time data acquisition, transmission, modeling analysis cannot be conducted. Meanwhile, a simplified and universal model is needed, which should use general knowledge as humans do. Give the preliminary results quickly and take the suspected drug back to the laboratory, where we can use other types of precise equipment and methods to solve the problems.

Various algorithms under the two common ideas are usually used in near-infrared spectroscopy (NIRs) have various limitations for rapid identification.

One idea is to use binary classification methods, that is, to set genuine drugs as negative samples and counterfeit drugs as positive samples, carrying out the binary classification of genuine and counterfeit drugs, and constructing classifiers using linear classifiers such as PLS-DA, SVM [5]–[10], BP-ANN [11]–[12], or using deep-learning classifiers like various Auto-encoding methods [13], DBN [14], CNN [15]. Although this kind of analysis has high classification accuracy in the laboratory, it is almost invalid for unknown samples beyond the range of modeling samples. In other words, it needs to pre-determine the limitation of testing samples within one or several certain drugs, and it also needs to obtain sufficient labeled samples in the modeling stage, which is quite difficult in the actual scenario. Additionally, this kind of algorithm is sensitive to the category, quality, and quantity of samples in its modeling stage. If the quantity of samples is not balanced, the accuracy rate will be biased to the larger number of samples. In other words, in reality, it is easier to judge a generic drug as a genuine drug since genuine drug samples are easier to obtain, and its accuracy rate is often suspected to be false high in practice.

Another idea is to use the one-class (OC) threshold identification algorithm for genuine drugs, which first selects or extracts the characteristics of the genuine drug, and then defines a set of threshold ranges according to each characteristic, identifying drugs that exceed the threshold range as counterfeit drugs. The representative methods to realize this idea is SVM-one class [16], SIMCA, and peak-valley correlation conformity test [17]. When applied in modeling, it guarantees the quality of a drug identified as a genuine drug to an extent, even if it is a counterfeit drug. However, it still needs to collect sufficient authentic drug spectra before modeling, and it is still ineffective for the samples beyond the class range of authentic drugs in modeling. Therefore, the generality of the model is limited, and there is still the problem of identifying

unknown samples, let alone the limited effect on the imitated counterfeit drugs with similar ingredients.

Given the above situation, this paper proposes a novel method: build the Siamese-network [18] to learn the universal knowledge of contractive, and to establish a general “common sense” identification model instead of the specific ones. In the on-site inspection, only two spectra are mandatory, one genuine drug spectrum as the benchmark and the other representing the drug to test. The model answers the question of whether the two spectra are “the same drug produced by the same manufacturer” when spectra are delivered. In this way, the “unknown samples cannot be identified” problem and the related problems such as the “insufficient samples”, the “unbalanced samples”, and the “sensitive cost of identification results” can be effectively solved.

II. MATERIALS, OBJECTIVES, AND METHODS

All the near-infrared spectra in this paper were obtained from the China National Institute for Food and Drug Control (NIFDC).

Since 2006, to conduct the mobile inspection on drug quality by quickly sample the near-infrared spectra of drugs, the Chinese government has invested in equipping more than 400 drug inspection vehicles in 363 cities across the country. The near-infrared spectrometers onboard is all Bruker Matrix-F spectrometers, and the spectral measurement methods used by the staff follow uniform internal inspection specifications [24].

However, most of the near-infrared spectra of drugs sampled during the inspection are discarded after the test, and it was after 2014 that NIFDC realized that it could collect some spectra for scientific research.

Because the NIFDC can only passively collect the spectra obtained by each inspection vehicle, the labeling information outside the spectra is largely lost. Of the more than 700,000 spectra collected in recent years, only 32015 have been labeled with both the name of the drug and the name of the manufacturer. Although **no less than 6 samples** should be taken for each drug according to the inspection specifications, we can not guarantee the actual sample quantity of a single drug from the spectra we have.

We gradually put these 32015 spectra into the experiments:

First, 1314 spectra of 9 drugs (Cephalexin tablets, metformin hydrochloride tablets, propylthiouracil tablets, etc.) produced by 25 manufacturers, is employed in the modeling of the Siamese-network.

When testing, we use the other 369 spectra of 4 drugs (Ranitidine Hydrochloride, citicoline sodium for injection, and paracetamol tablets, etc.) produced by other 9 manufacturers, which are unknown in the modeling stage.

This total of 1683 spectra will ensure the basic functionality of the proposed model.

For generalizing consideration, this paper will use the whole of 32015 spectra within 135 drugs produced by 391 manufacturers by randomly dividing them into training

categories and test categories, and then, to find whether the model can be effective for all drugs in a wide range, we use the same method to deal with them to investigate the accuracy.

All the spectra have been adjusted to the wavelength range of 4000-11995 cm^{-1} with a resolution of 4cm^{-1} .

Table 1 shows the relevant information of 9 drugs (modeling samples) produced by 25 manufacturers, which will be used to construct the proposed model.

In Table 1, there are 25 categories of modeling samples classified by “drug-manufacturer” including tablets, granules, and capsules (but not including powder injections, which we did on purpose). There are classes with insufficient

TABLE 1. NIR data for constructing the model

Drug Name	Company	Package	Spectra
Levonorgestrel Tablets	Shanghai Xinyi Kangjie Pharmaceutical Co., Ltd	Tablets	21
	Shenyang No.1 pharmaceutical factory of Northeast Pharmaceutical Group Company	Tablets	54
	Chengdu Di'ao Pharmaceutical Group Co., Ltd	Tablets	36
Yinghuang	Lunan Houpu Pharmaceutical Co., Ltd	Tablets	24
	Zhongshan Zhongzhi Pharmaceutical Co., Ltd	Granule	88
	Jiangxi Jimin Xinxin Pharmaceutical Co., Ltd	Granule	44
Metformin Hydrochloride Tablets	Shanghai Hengshan Pharmaceutical Co., Ltd	Tablets	54
	Shanghai Squibb Pharmaceutical Co., Ltd	Tablets	37
	Beijing Zhonghui Pharmaceutical Co., Ltd	Tablets	75
Cefaclor Capsules	Shandong Zibo Xinda Pharmaceutical Co., Ltd	Capsules	36
	Guangzhou Nanxin Pharmaceutical Co., Ltd.	Capsules	90
Cefalexin Tablets	Changchun Xin'an Pharmaceutical Co., Ltd	Tablets	84
	Jilin yimintang Pharmaceutical Co., Ltd	Tablets	36
	Jilin Daojun Pharmaceutical Co., Ltd	Tablets	48
Roxithromycin Capsules	Yangtze River Pharm Pharmaceuticals Co., Ltd.	Capsules	209
	Hunan Qianjin Xiangjiang Pharmaceutical Co., Ltd.	Capsules	66
Ibuprofen	Changchun Overseas Pharmaceutical Group Co., Ltd.	Capsules	54
	Huainan Jia League Pharmaceutical Co., Ltd.	Tablets	36
Propylthiouracil Tablets	Shanghai Xinyi Pharmaceutical Co., Ltd.	Tablets	30
	Nantong Essence Pharmaceutical Co., Ltd.	Tablets	36
Diclofenac sodium Tablets	Jilin Province Seven Star Yampo Pharmaceutical Co., Ltd.	Tablets	36
	Jilin Province Broad Weiye Pharmaceutical Co., Ltd.	Tablets	48
	Liaoyuan Real Rain Pharmaceutical Co., Ltd.	Tablets	30
	Liaoyuan Dikang Pharmaceutical Co., Ltd.	Tablets	42
Total	9 drugs produced by 25 manufacturers.		1314

spectra (some classes only have 21 spectra), and all categories of spectra are unbalanced, with only 21 spectra at the lowest and 209 spectra at the highest.

After the modeling is complete, four other drugs produced by 13 other different manufacturers (test samples) will be tested for identification.

Drug-related information is shown in Table 2.

TABLE 2. NIR data for testing

Drug Name	Company	Package	Spectra
Ranitidine hydrochloride	Jilin Xianfeng technology Pharmaceutical Co., Ltd	Capsules	72
	Guangzhou Ouhua Pharmaceutical Co., Ltd	Tablets	108
	Shanghai shikangte Pharmaceutical Co., Ltd	Capsules	54
Citicoline sodium for injection	Henan Furen huaiqingtang Pharmaceutical Co., Ltd	Powder injection	27
	Zhejiang Asia Pacific Pharmaceutical Co., Ltd	Powder injection	27
Azithromycin	Siping emer Pharmaceutical Co., Ltd	Tablets	36
	Shanghai Xinyi Jiahua Pharmaceutical Co., Ltd	Tablets	15
Shexiang Jiegu capsule	Jilin tianqiang Pharmaceutical Co., Ltd	Capsules	12
	Tangshan jingzhongshan Pharmaceutical Co., Ltd	Capsules	18
Total	4 drugs produced by 9 manufacturers.		369

As can be seen in Table 2, according to “drug-manufacturer” classifying, neither drugs nor manufacturers have been found in the training samples in Table 1. Therefore, for the training course, the test samples are totally “unknown samples”. We also intentionally added the powder injection package to the test samples that were not covered in the training samples. If the experimental results are good, we will add the medicinal ferrous sulfate produced by two manufacturers at the end, which is an inorganic substance, and its identification method is significantly different from that of ordinary drugs. This will be a strict test for the model when identifying “unknown samples”.

Examining the spectra of modeling samples and test samples, these spectra have the following characteristics:

1) There is little difference between spectral classes. For example, the spectra of diclofenac sodium tablets produced by different manufacturers shown in Figure 1 describe that the important spectral sections (peak valley sections) of the two spectra overlap greatly, and there is almost no difference in a manual inspection.

2) Inner the same drug, there are great differences among the spectra, especially Chinese patent medicines. As shown in Figure 2, the Yinhuang series has a wide range of spectral differences within its class, and they can be easily divided erroneously into different categories in general modeling, causing the “judge the genuine drug as counterfeit drug” result.

It can be seen from Figure 1 and Figure 2 that the information contained in the modeling samples and test samples

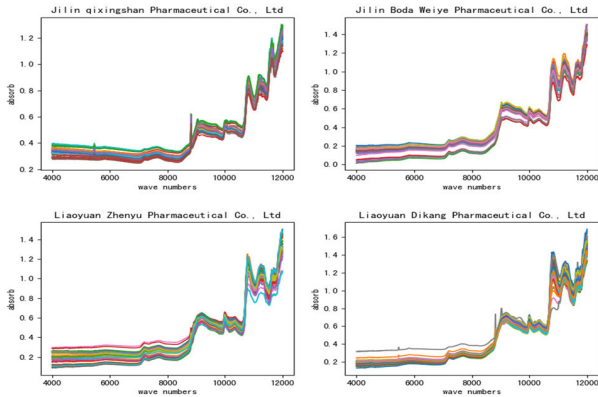


FIGURE 1. The spectrum of diclofenac sodium tablets produced by different manufacturers shows little difference among different categories.

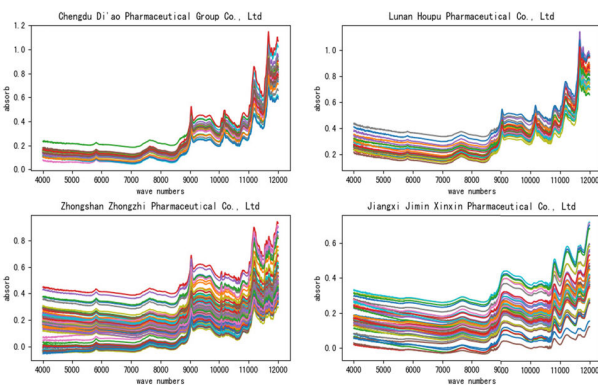


FIGURE 2. Spectra of Yinhuang series drugs produced by the same manufacturer show a great difference within the category.

selected in this paper coincides with the following two main difficulties in the on-site identification practice of genuine and counterfeit drugs:

1) THE DIFFICULTY OF DISTINGUISHING “SELL SECONDS AT BEST QUALITY PRICES”

In the practice of drug identification, there are often some kinds of drugs, all of them have similar ingredients and have a certain curative effect on a certain disease, but the prices are very different. Manufacturers may use cheap drugs instead of genuine drugs at higher prices. If these drugs can be compatible, counterfeiters even add a small number of genuine drugs into their cheap medicines to increase the difficulties of identification.

It is possible to make the spectral difference between genuine and counterfeit drugs smaller than the spectral differences obtained by different samples, different measurement methods, and different measuring equipment inside the genuine drug.

2) THE DIFFICULTY OF IDENTIFYING IMITATION AND COUNTERFEITING BETWEEN DIFFERENT MANUFACTURERS OF THE SAME DRUG

The near-infrared spectra of the same drug produced by different manufacturers according to the same drug standard are

very similar, and the important spectral bands (peak valley positions) are even overlapped. This situation often exists between generic drugs and original drugs.

Generally, the time from R&D to the final registration and marketing of the original drug is as long as 15 years, and it costs a lot to go through four phases of clinical trials. Such drugs cannot be copied before the patent expires, and enjoy the protection of separate pricing and other policies. However, generic drugs only copy the main components of the original research drug. Even if a large amount of money is invested in the imitation process, the cost is only about 1/3 or even 1/6 of the original research drug. Therefore, from the perspective of the counterfeiter, the generic drug and the original research drug should be as consistent as possible without being distinguished. It is very possible to label generic drugs as original drugs for sale.

Identification of generic drugs posing as the original drug sales, or, identifying the same drug posing as a well-known brand, because of its composition differences near the trace, identification is more difficult.

Referring to Table 1, Table 2, Figure 1, and Figure 2, it can be seen that the information contained in the spectra we have can reflect the main difficulties in real identification scenarios. The modeling and testing process based on these data covered the four issues as “need to detect unknown samples beyond the scope of the modeling sample category,” “insufficient samples”, “samples imbalance”, and “model application of error-sensitive”.

III. ALGORITHM DESCRIPTION

In this paper, we mainly use Siamese-network to build the model.

In recent years, the Siamese-network has been widely used in handwriting font identification, face ID authentication, dynamic object tracking, and other graphic image and video tasks [18]–[22], and achieved good results. In the mineral (inorganic) analysis by Raman spectroscopy, Jinchao Liu *et al.* used a twin neural network to compare the unknown spectrum with the known spectrum in the spectral library [23] and achieved better similarity measurement results than cosine distance and LMNN (large margin nearest neighbor) methods. However, there is no report about Siamese-network application in the field of near-infrared spectroscopy of organic compounds (including drugs).

The modeling process of the Siamese-network used in this paper is shown in Figure 3:

In Figure 3, we first construct the required data set for the Siamese-network by the method of the left block diagram, then input the data into the model on the right for training. After the training is successful, the model will output a D_w value for any input spectrum pair (usually one is the genuine drug spectrum and the other is the spectrum to be tested), to judge whether the D_w is less than 0.5, the answer of “whether the drug category is the same” can be obtained so that the genuine and counterfeit drugs can be identified on-site.

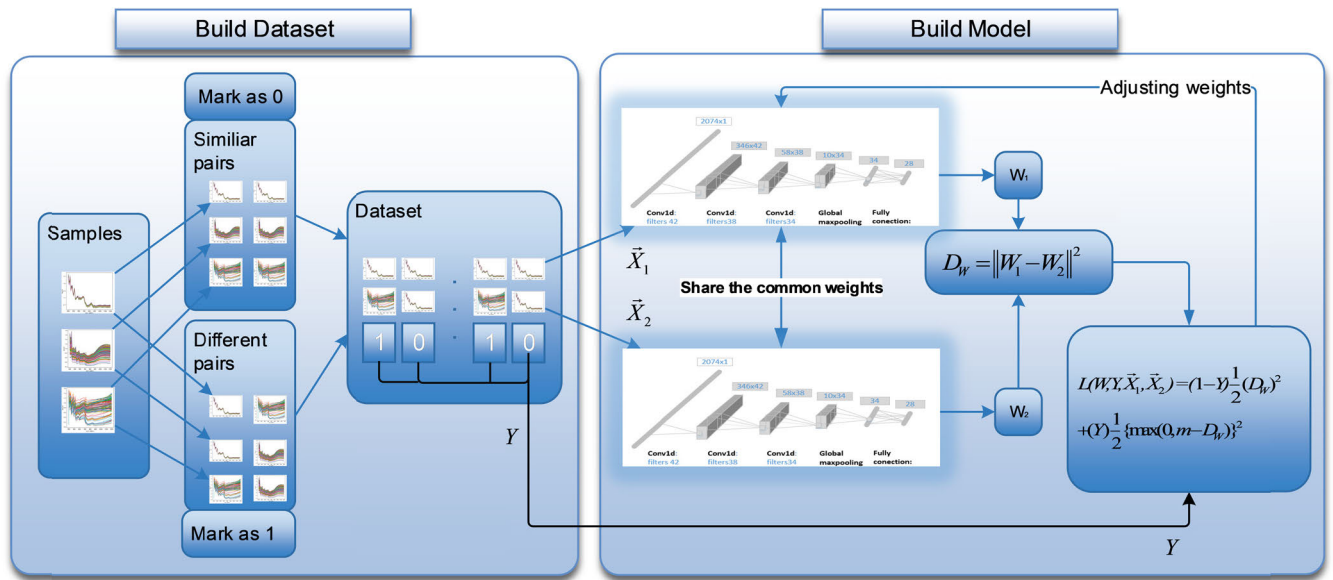


FIGURE 3. Model construction and training process.

A. CONSTRUCTION OF DATA SET

1) PAIRING

In this paper, because the input of training and testing is an identical or heterogeneous pair, rather than a single spectrum, therefore, training samples and testing samples are not members of the training set and test set. We need to use the pairing method to build a training set and test set (collectively called data sets).

We use the bootstrap method to construct the pairs: for every two extractions in the same class, the similar pair counted once, and it labeled as $y = 0$. For heterogeneous pairs, one spectrum is randomly selected from one class, and the other is randomly selected from other classes, followed by the different pair counted once, its label marked as $y = 1$. In this way, s_1 rounds are extracted from the same kind of pairing, and s_2 rounds are extracted from the heterogeneous pairing, then $s_1 + s_2$ pairs are obtained. Compound the pairs, the train and test set will be formed. Since bootstrapping method sampling with replacement, limits of spectra could be sampled unlimited rounds. The problem of insufficient spectral numbers within a class could be solved to a certain extent.

If the total number of spectra is n , the number of different pairs is n^2 , which is a large number. At the same time, if the number of classes is k and the number of spectra in each class is m , then the number of similar pairs is km^2 , and the number of different pairs is $k^2m^2 - km^2$. The number of similar pairs is only $1 / (k-1)$ of the number of different pairs. The chosen spaces of similar pairs and different pairs are extremely asymmetric.

To ensure the typicality and uniformity of the subsequent data sets, a fair sampling strategy must be adopted in the sampling pairing in a large and unbalanced space.

In this paper, a fair sampling strategy is designed for this purpose:

Firstly, the total number of pairs N , the number of classes k , and the sampling ratio α of the similar and different pairs should pre-determined by users.

Then, according to formula (1), we can get the s_1 and s_2 . For each class, according to the counting rounds s_1 and s_2 , training data and testing data can be obtained by randomly selecting.

$$\begin{cases} s_1 = \frac{N}{k + \frac{k}{\alpha}} \\ s_2 = \frac{N}{k\alpha + k} \end{cases} \quad (1)$$

This sampling method will have the following advantages: first, $s_1 + s_2 = n/k$, which can ensure that each class (whether the spectra number deviates too much from the average spectra number or not) can get an equal share of attention, thus improving the impact of insufficient samples within the class and unbalanced samples between classes. Second, it can solve the problem of cost sensitivity which is described below.

2) COST-SENSITIVE PROBLEM HANDLING

In the practice, the genuine drug samples are easy to obtain, while the counterfeit drug samples can only get one or two cases each time when the counterfeiter has been caught. The genuine and counterfeit drug samples are often uneven. In this way, when the detection error occurs, it is easy to judge the counterfeit drug as a genuine drug, but it is not easy to judge the genuine drug as a counterfeit drug.

The cost or risk of identifying a genuine drug as a counterfeit drug is not the same as identifying a counterfeit drug as a genuine drug. In this scenario, when the “genuine drug is identified as a counterfeit drug” error occurs, it can often

be taken back to the laboratory for further analysis and determination. However, if the “counterfeit drug is identified as a genuine drug” error occurs, the counterfeit drug usually has to be released, thus cause serious consequences.

In this paper, the space of heterogeneous sampling is much larger than that of the same class sampling. In this way, the cost-sensitive factors α can be taken into account by setting the sampling ratio of the same and different classes in the construction of the data set.

From our experiences, α taking 0.125-0.5, that stands similar sampling: different sampling between 1:2 to 1:8 will play a better role than others. It can ensure that the cost-sensitive problem can be solved to a certain extent, and at the same time, it will not sacrifice the accuracy to ensure the purpose. We set it as 0.333, that is, similar sampling: different sampling = 1:3.

3) CONSTRUCTION OF TRAINING SET AND TEST SET

This paper focuses on solving the problem of unknown samples. Therefore, it is necessary to completely isolate the training process from the acquisition of test process information. The training samples should not be involved in any spectrum of test samples in the process of heterogeneous pairing. In the same way, the test samples should not involve any training spectra when they are doing heterogeneous pairing either.

After the isolation principle is defined, a fixed number of pairings are selected from the training samples as the training set, and another fixed number of pairs are selected from the test samples as the test set. In this paper, 400 * 60 pairs of the training set and 80 * 60 pairs of the test set are extracted.

B. CONSTRUCTION OF 1D-CNN

In Siamese-network, for two groups of values in pairing, a neural network that can extract its features must be provided, and the two neural networks can share the same weights. In the implementation, the two neural networks are often combined into one. After receiving two inputs, the combined network calculates the distance, loss, and then process optimizing progress (adjusts the weights). In this paper, a one-dimensional convolutional neural network, which can effectively extract the near-infrared spectra of drugs, is constructed to realize Siamese-network. The parameters of the network have been marked in Figure 4. After many rounds of exploration and repeated debugging, these parameters have been confirmed to be the optimal results in our experience.

For the identification scenario in this paper, it is the best choice to select 1D-CNN as the pre-treatment network of Siamese network preprocessing. The alternative can also be Sparse Auto-encoder (SAE), Multi-Layer Perceptron (MLP), and other artificial neural networks.

To evaluate the effectiveness of 1D-CNN as the basic preprocessing network of the Siamese network, we construct SAE, MLP as alternative neural networks, and then test their effects when they replace 1D-CNN.

The SAE network is constructed using the method provided in reference [14], choosing the 2074-120-60-120-2074

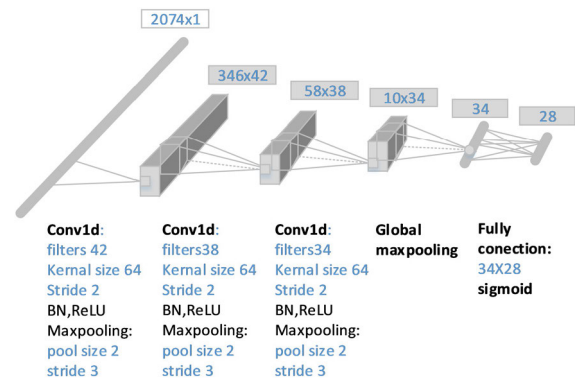


FIGURE 4. One dimensional convolution neural network and its main parameters for near-infrared spectrum feature extraction of drugs constructed in this paper.

structure and using cross entropy as its loss function to calculate reconstruction loss. SAE needs to be pre-trained, and its trained encoder is connected to the Siamese network through a fully connected layer.

MLP directly replace 1D-CNN with the 2074-150-28 structure.

Experiments (described in section IV. D) show that at least in our basic experiments, 1D-CNN is better than the other two networks. Therefore, this paper uses 1D-CNN as the pre-treatment network.

In this paper, Euclidean distance is used to measure the features extracted from a one-dimensional convolution network, and the contractive loss is used as the loss function to optimize the network [22]. Namely:

$$\begin{cases} D_W = \|W_1 - W_2\|^2 \\ L(W, Y, \vec{X}_1, \vec{X}_2) = (1 - Y) \frac{1}{2} (D_W)^2 + (Y) \\ \quad \times \frac{1}{2} \{\max(0, m - D_W)\}^2 \end{cases} \quad (2)$$

Among them, D_w is Euclidean distance, which is calculated according to the output of the Siamese-network. L is the training loss, and Y is the label generated by the pairing process. m is the margin value, which adjusts the distance between two different spectra within $0 \sim m$. In the experiments, we set $m = 1$.

IV. EXPERIMENT AND DISCUSSION

To test the effectiveness and superiority of the method in this paper, three experiments are carried out. One is the basic experiment related to the training and test operation of the model in this paper; after the basic experiment is successful, expand the modeling samples and test samples to our full spectra library; the last experiment is a comparative experiment, which verifies the superiority of the model by comparing with the current mainstream drug identification algorithms.

A. EXPERIMENTAL ENVIRONMENT

This paper uses the following hardware and software environment for data modeling experiment.

TABLE 3. Experimental results of test samples

Drug Name	Company	Precision	Recall score	F1	Accuracy
Ranitidine hydrochloride	Jilin Xianfeng technology Pharmaceutical Co., Ltd	1	1	1	1
	Guangzhou Ouhua Pharmaceutical Co., Ltd	1	1	1	1
	Shanghai shikangte Pharmaceutical Co., Ltd	1	1	1	1
Citicoline sodium for injection	Henan Furen huaiqingtang Pharmaceutical Co., Ltd	0.941	0.916	0.922	0.948
	Zhejiang Asia Pacific Pharmaceutical Co., Ltd	0.940	0.935	0.933	0.927
Azithromycin	Siping emer Pharmaceutical Co., Ltd	1	1	1	1
	Shanghai Xinyi Jiahua Pharmaceutical Co., Ltd	1	1	1	1
Shexiang Jiegu capsule	Jilin tianqiang Pharmaceutical Co., Ltd	1	1	1	1
	Tangshan jingzhongshan Pharmaceutical Co., Ltd	1	1	1	1
Average (by pairs)		0.973	0.972	0.973	0.973

Hardware environment: CPU Xeon 2698v4 (20 cores, 40 threads), memory 96GB, SSD 1TB, GPU NVIDIA Tesla V100.

Software environment: operating system Ubuntu 20.04.1 LTS, NVIDIA driver version 440.33.01, CUDA v10.0, cudnn v7.6.5, keras-gpu 2.3.1, tensorflow-gpu 1.15.0, sci-kit learn 0.23.2.

B. EXPERIMENTAL PREPARATION

In the experiment, in addition to the network parameters already set in Figure 4, the following super parameters are given in this paper:

When sampling pairing, α is set to be 1:3. The number of training pairs is 400 times the batch size to form the training set, while the number of test pairs is 80 times the batch size to form the test set. Training pairs are extracted and generated in training classes, while test pairs are extracted and generated in test classes without interference.

Model training uses RMSprop optimizer with batch size set to 60.

Each experiment trained 60 epochs. Before each epoch began, the input data (pairs) were generated in real-time from the sampling-pairing process.

C. BASIC EXPERIMENT

The basic experiment is based on the samples given in Table 1 (modeling samples) and Table 2 (test samples). The experimental results of test samples are shown in Table 3.

As can be seen from Table 3, the model in this paper has a good identification effect. Although the test sample is completely different from the training sample in terms of drug name or manufacturer, except citicoline sodium for injection, the precision, recall, accuracy, and F1 score of other test samples (ranitidine hydrochloride, azithromycin series, Shexiang Jiegu capsule) are all 100%. The model has complete discrimination ability to most test samples.

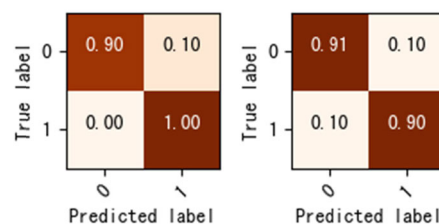


FIGURE 5. Confusion matrix of identification results of citicoline sodium for injection produced by two manufacturers.

Although there are errors in the identification of citicoline sodium for injection, its identification ability is still available, and the accuracy rate is above 92%.

The confusion matrix of citicoline sodium for injection was drawn as shown in Figure 5 (left is manufactured by Henan Furen huaiqingtang Pharmaceutical Co., Ltd, and right is manufactured by Zhejiang Asia Pacific Pharmaceutical Co., Ltd).

As can be seen from Figure 5, the error does not appear in the focus of “mistakenly classifying counterfeit drugs into genuine drugs” (when the true label is 1 and prediction is 0), it appears in the situation of “mistakenly classifying genuine drugs into counterfeit drugs”. This shows that we set the similar and different sampling ratio as 1:3 (pay more attention to “do not mistakenly classify counterfeit drugs into genuine drugs”) to play its due role. While maintaining a high accuracy rate (at least 93%), the cost-sensitive problem has been effectively solved. The experimental results have achieved our pre-set goal.

Combined with the description of citicoline sodium for injection in Table 2 of section 2, it is found that citicoline sodium for injection is a powder injection, but our training sample does not contain the powder injection. Even for human teaching, we must first teach the difference between the spectrum of tablets, capsules, and powder injection to have discrimination. The small error caused by the lack of information can be understood.

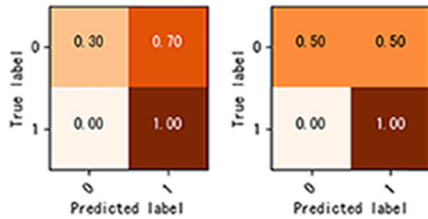


FIGURE 6. Confusion matrix of identification results of ferrous sulfate tablets produced by two manufacturers.

We intentionally expand the missing information by adding ferrous sulfate tablets produced by two manufacturers into the test samples and draw the confusion matrix of the experimental results according to the trial results in Figure 6. The left one is manufactured by Taiyuan Satellite Pharmaceutical Co., Ltd, and the right is manufactured by Shanghai Huanghai Pharmaceutical Co., Ltd.

As the main component of ferrous sulfate tablets is inorganic, its spectral difference is significantly different from that of various organic substances in training samples. As can be seen from Figure 6, the error is significantly enlarged. When the error is the largest, 70% of the genuine drugs are judged to be counterfeit drugs. However, the model still does not take counterfeit drugs as genuine drugs.

D. PRE-TREATMENT EXPERIMENT

To test the effect of other pre-treatment networks, we replaced 1D-CNN with SAE and MLP respectively, and carried out the same experiment.

The SAE network is only used as a pre-training model in this experiment. After the training is over, it connects to a fully connected layer (28 outputs) through its feature layer (the 60 part of the 2074-120-60-120-2074 structure), which will provide the hidden 28 features for the Siamese-network to calculate the D_w value. The super parameters are set as follows:

The ‘RELU’ activation function is used inside the Auto-encoder, and the sigmoid activation function is used in the full connection layer. During the pre-training course, the optimizer is set to Adam, the learning rate is initialized to 0.003. β_1 is set to 0.9, β_2 is set to 0.999, the batch size is set to 60, and the decay value is set to 10^{-5} . 150 epochs were pre-trained.

Alternatively, MLP using 2074-150-28 structure with Keras’ ‘Dense()’ function, and is directly connected to Siamese-network with ‘sigmoid’ activation function.

Comparing the effect of Siamese network constructed by 1D CNN, SAE, and MLP, the experimental results are shown in Table 4:

As can be seen from Table 4, 1D-CNN has the best effect, and its accuracy rate is about 0.8 and 1.2 percent higher than the other two networks respectively. The basic pre-treatment network is mainly used to extract the features of the spectrum. When the main structure (Siamese-network) is determined, the change of the feature extraction method will

TABLE 4. Experimental results under different pre-treatment networks

pre-treat network	Precise	Recall	F1	Accuracy
1D CNN	0.973±0.005	0.972±0.005	0.973±0.005	0.973±0.005
SAE	0.965±0.003	0.965±0.003	0.965±0.003	0.965±0.003
MLP	0.960±0.005	0.959±0.004	0.959±0.003	0.961±0.004

also affect the experimental results, but its influence is quite limited relatively.

E. EXTENDED EXPERIMENT

To avoid large errors caused by lack of information, based on the success of the above experiments, we used 32015 spectra of 135 drugs produced by 391 manufacturers in the spectra library (the source of spectra library has already been described in section II) to expand the experiment. The purpose is to include most of the information needed for general drug identification in the training samples so that the established model can be applied to most of the drugs in the market.

There are 472 classes of drugs in the spectra library according to “drug-manufacturer”. According to 9:1, 8:2, 7:3, 6:4, 5:5, 4:6, 3:7, 2:8, and 1:9, the training samples and test samples are randomly divided, and the training set and test set are generated according to the same rules of the basic experiment. The same modeling training and testing process is carried out just like the basic experiment. The experimental results are shown in Table 5.

As can be seen from Table 5, the modeling also achieved a good identification effect. In the above cases, the average of scores and accuracies were above 96%.

The training process was investigated from the 9:1 experiment, and the curves of accuracy and loss with the increase of epochs are extracted as shown in Figure 7.

It can be seen from Figure 7 that although we set epoch to 60, in fact, with only 9 epochs, the increase of experimental accuracy and the decrease of loss has come to an available point. The optimal values (the highest point of accuracy and the lowest point of loss) within 0 to 9 epochs are not much different from the optimal value after 60 epochs. When we test or use the model, we usually take the model generated

TABLE 5. Experimental results under different training and test set partitions

Train: Test	Precise	Recall	F1	Accuracy
9:1	0.973±0.005	0.970±0.005	0.971±0.005	0.970±0.005
8:2	0.979±0.003	0.979±0.003	0.979±0.003	0.979±0.003
7:3	0.973±0.008	0.972±0.009	0.973±0.008	0.972±0.090
6:4	0.977±0.005	0.977±0.005	0.977±0.005	0.977±0.005
5:5	0.970±0.008	0.970±0.008	0.969±0.009	0.970±0.008
4:6	0.974±0.005	0.974±0.005	0.974±0.005	0.974±0.005
3:7	0.964±0.012	0.964±0.012	0.964±0.012	0.964±0.012
2:8	0.965±0.011	0.965±0.011	0.965±0.011	0.965±0.011
1:9	0.963±0.011	0.963±0.011	0.963±0.011	0.963±0.011

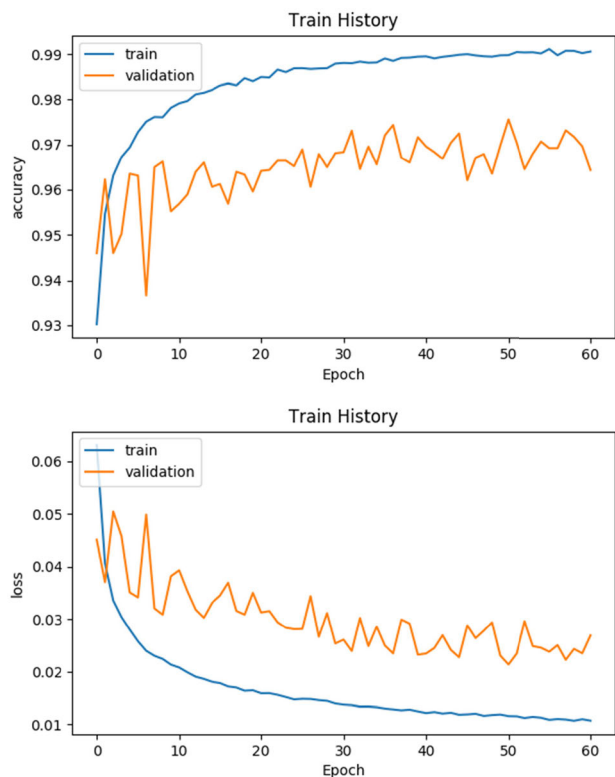


FIGURE 7. Changes in accuracy and loss during training.

under the optimal value. Then after trained 9 epoch, we can assert that the model is available, although it is not globally optimal. Since data will be randomly paired to form a training set before the epoch starts, the growth of epochs essentially means that the training data is randomly expanded. Nine epochs are available, which means that there is not much training data demanded in the training course. Therefore, the problem of insufficient and unbalanced samples is solved.

The confusion matrix of each experiment is drawn in Figure 8.

As can be seen from Figure 8, the probability of “counterfeit drug being classified as the genuine drug” (the true label is 0 and prediction is 1) is significantly less than that of “genuine drug being classified as the counterfeit drug” (the true label is 1 and the prediction is 0), it shows that we set the same and different sampling ratio as 1:3 (paying more attention to “not mistakenly classifying the counterfeit drugs into genuine drugs”) in large data sets also take effect. The cost-sensitivity problem is thus resolved to a certain extent.

F. COMPARATIVE EXPERIMENT

In this paper, extended experiments are used to compare the accuracy of six algorithms in two categories of traditional true and false drug identification practice. Three of them are binary classification algorithms: RBF-SVM, PLS-DA, Sparse auto-encoder (SAE). The other is one class identification algorithms: SVM one-class (SVM-OC), SIMCA, conformity test (CT) [17]. Because the preconditions of

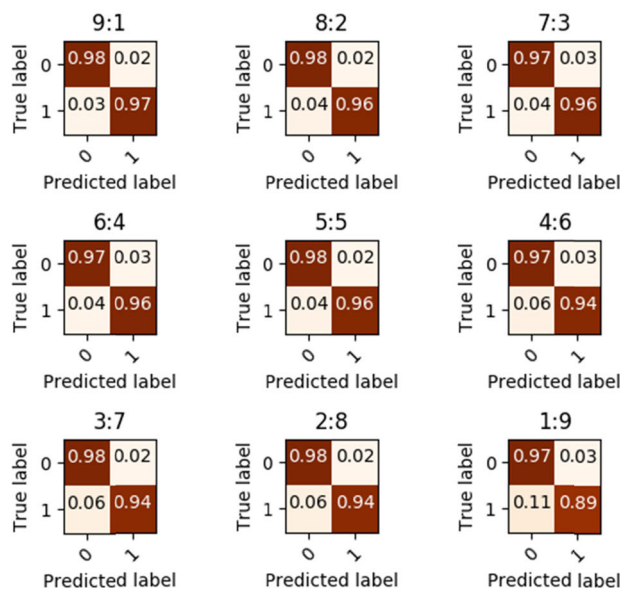


FIGURE 8. Confusion matrix of each experiment.

these algorithms are different from our method in this paper, we need to construct training sets and test sets for these algorithms according to the common logic.

For each class, we assume that the in-class spectra are genuine samples, while spectra in the rest outer classes are counterfeit samples. Based on this consideration, the training and test sets of each comparative algorithm are established.

The parameters of RBF-SVM, PLS-DA, and SAE are set as follows:

The gamma value of RBF SVM is 0.001, the C value is 1, and the number of PLS-DA components is 28. Other parameters are the default values provided by scikit-learn software.

SAE uses the 2074-120-28-120-2074 structure to set up its networks, and RELU is used as its activate function, 28-150-logistic regression structure is used to design the network for the classifier, and sigmoid activation function is used in the progress of classifying. By using Adam optimizer and setting the batch size to 60, the training is divided into two stages. In the first stage, the learning rate is initialized to 0.003, β_1 is set to 0.9, β_2 is set to 0.999, and the decay parameter is set to 10-5, trained 150 epochs. In the second stage, the learning rate is initialized to 0.000007, and the rest of the parameters are the same in the first stage.

SVM one class (SVM-OC), SIMCA, and consistency test (CT) are calculated by NIDFC’s customized product software, which is developed from OPUS, the official working software of the spectrometer manufacture Bruker.

The results of comparative experiments are shown in Table 6.

It can be seen from Table 6:

1) Generally speaking, although the test rules are obviously biased towards comparative algorithms (for example, our method is tested with unknown categories, and the comparative algorithms are at least genuine drugs within the training

TABLE 6. Accuracy of various multi-class classification algorithms

<i>Train: Test</i>	<i>Siamese-network</i>	<i>RBF-SVM</i>	<i>PLS-DA</i>	<i>SAE</i>	<i>SVM-OC</i>	<i>SIMCA</i>	<i>UT</i>
9:1	0.973±0.005	0.723±0.102	0.632±0.153	0.784±0.072	0.791±0.163	0.618±0.168	0.695±0.231
8:2	0.979±0.003	0.701±0.013	0.577±0.192	0.704±0.121	0.732±0.138	0.612±0.187	0.701±0.206
7:3	0.973±0.005	0.653±0.021	0.513±0.251	0.682±0.138	0.712±0.106	0.543±0.225	0.412±0.301
6:4	0.977±0.005	0.642±0.101	0.591±0.203	0.657±0.171	0.700±0.117	0.501±0.203	0.513±0.275
5:5	0.970±0.008	0.587±0.203	0.416±0.206	0.601±0.172	0.631±0.114	0.483±0.279	0.425±0.279
4:6	0.974±0.005	0.589±0.201	0.384±0.313	0.532±0.105	0.578±0.212	0.542±0.221	0.612±0.283
3:7	0.964±0.012	0.473±0.251	0.544±0.307	0.581±0.132	0.501±0.208	0.375±0.306	0.511±0.301
2:8	0.965±0.011	0.577±0.204	0.513±0.301	0.578±0.118	0.492±0.273	0.512±0.237	0.521±0.363
1:9	0.963±0.011	0.511±0.271	0.448±0.254	0.513±0.174	0.533±0.391	0.402±0.401	0.528±0.401

TABLE 7. Training and inferring time of each algorithm (in second, training time/inferring time)

<i>Train: Test</i>	<i>Siamese-network</i>	<i>RBF-SVM</i>	<i>PLS-DA</i>	<i>SAE</i>	<i>SVM-OC</i>	<i>SIMCA</i>	<i>UT</i>
9:1	172.0/0.2	592.4/3.3	210.8/1.1	416.3/2.1	207.7/2.7	239.8/1.5	257.6/2.6
8:2	171.8/0.2	551.3/4.5	202.9/2.0	408.2/2.4	204.7/2.9	240.3/2.4	221.3/2.4
7:3	171.7/0.2	531.1/5.8	197.5/2.3	401.2/2.3	201.3/3.2	229.5/2.6	211.0/2.3
6:4	172.0/0.3	502.5/6.9	183.6/2.4	398.7/2.8	197.6/3.5	219.2/2.6	209.5/2.3
5:5	170.9/0.2	473.3/8.0	180.3/2.9	379.3/3.2	195.0/3.7	209.2/2.6	208.7/2.2
4:6	171.1/0.3	425.2/9.1	180.5/2.4	358.4/3.3	191.0/4.2	208.8/2.7	204.3/2.2
3:7	170.8/0.2	378.1/9.9	177.2/3.0	341.1/3.5	187.4/4.3	208.0/3.9	201.6/2.1
2:8	169.9/0.2	342.0/10.8	175.0/3.2	339.0/3.9	184.2/5.1	200.9/5.0	198.0/1.7

set), the algorithm in this paper still has better performance and its accuracy is far higher than any other algorithm.

2) The accuracy of each comparative algorithm decreases with the increase of test set share. When the proportion is less than 5:5, the accuracy of the algorithm is almost the same as that of the random answer, sometimes even worse. However, our method is stable in all ratios from 9:1 to 1:9, and the variance is the smallest among all algorithms.

The training time and inferring time of each algorithm are shown in Table 7. Siamese network, SAE uses GPU for training and inferring, and CPU is used for other algorithms.

It can be seen from the table that:

No matter training time cost or inferring time cost, our algorithm is far lower than other algorithms.

There are three reasons for the long-running time of other algorithms.

Firstly, this method (Siamese-network) and SAE use the TensorFlow-GPU module for training and inferring, and it can effectively use more than 3000 CUDA cores in GPU, while other algorithms using CPU modules for training and inferring. For example, RBF-SVM, PLS-DA use scikit-learn as their backbone, and scikit-learn can not use GPU, worse still, it can use only single-core CPU utilization unless the customized parallel modules are explicitly designed. Therefore, although the experimental equipment includes a powerful GPU and a 12 cores CPU, its efficiency does not play out on these algorithms.

Secondly, because this paper is applied in the identification cases, RBF-SVM, PLS-DA, and SAE are used for binary classification. This means that among the 472

“drug-manufacturer” classes, these algorithms need to establish 472 sub-models, and the data were loaded and combined frequently from the database to form the training set and test set required by each sub-model. Therefore much time had to spend on input and output processing. During the courses, the performance of the database itself also consuming some experimental time. Compared with our algorithm, our algorithm does not need to consider these problems because it only does these things once.

Finally, although each trained comparative model only needs input one inspecting spectrum to put into utilization (our method needs to input two spectra, one for reference and one for inspection), it needs to determine which model to use in application from 472 choices. After that, it also needs to unload the old model and its parameters, select and load the new model and its parameters from a total of 472 models. Therefore, it takes too much additional time in its inference course. However, our method replaces the complex courses with the “inputting two spectra” strategy, which can avoid all the troubles above.

Therefore, our algorithm is better than other traditional algorithms in terms of simplicity and ease of use.

V. CONCLUSION

In this paper, based on Siamese-network, a universal near-infrared spectroscopy identification model was established. Compared with various traditional methods, it has many advantages, such as easy to use, strong generalization, can deal with unknown samples beyond the scope of modeling samples, and can solve the problems of insufficient

samples within the class, unbalanced samples between classes, and cost-sensitive problems of error occurrence.

In the process of modeling, aiming at the main problem of “unknown samples beyond the range of modeling samples”, this paper uniquely designs the training set division and sampling pairing method and constructs the Siamese-network model through the customized one-dimensional convolutional neural network. Through the performance test and the comparison of 6 traditional drug classification and identification algorithms, the effectiveness of the model is analyzed and verified.

The strong identification ability of the network in the scene of near-infrared spectrum identification of drugs provides a useful reference for readers to identify and classify drugs in multi-variety, multi-manufacturer, insufficient, unbalanced samples, and cost-sensitive application scenarios. This method can also be applied to the identification of crude oil, organic chemical industry, and other similar scenes, and has broad application prospects.

REFERENCES

- [1] P. Prajapati, R. Solanki, V. Modi, and T. Basuri, “A brief review on NIR spectroscopy and its pharmaceutical applications,” *Int. J. Pharmaceutical Chem. Anal.*, vol. 3, no. 3, p. 117, 2016.
- [2] R. Deidda, P.-Y. Sacre, M. Clavaud, L. Coïc, H. Avohou, P. Hubert, and E. Ziemons, “Vibrational spectroscopy in analysis of pharmaceuticals: Critical review of innovative portable and handheld NIR and Raman spectrophotometers,” *TrAC Trends Anal. Chem.*, vol. 114, pp. 251–259, May 2019.
- [3] D. A. Burns and E. W. Ciurczak, *Handbook of Near-Infrared Analysis, Revised and Expanded*, 3rd ed. Boca Raton, FL, USA: CRC Press, 2007.
- [4] X. L. Chu, *Molecular Spectroscopy Analytical Technology Combined With Chemometrics and Its Applications*. Beijing, China: Chemi-Cal Industry Press, 2011, p. 95.
- [5] F. Sun, Y. Chen, K.-Y. Wang, S.-M. Wang, and S.-W. Liang, “Identification of genuine and adulterated *Pinellia ternata* by mid-infrared (MIR) and near-infrared (NIR) spectroscopy with partial least squares–discriminant analysis (PLS-DA),” *Anal. Lett.*, vol. 53, no. 6, pp. 937–959, Apr. 2020.
- [6] H.-Y. Fu, D.-C. Huang, T.-M. Yang, Y.-B. She, and H. Zhang, “Rapid recognition of Chinese herbal pieces of areca catechu by different conducted processes using Fourier transform mid-infrared and near-infrared spectroscopy combined with partial least-squares discriminant analysis,” *Chin. Chem. Lett.*, vol. 24, no. 7, pp. 639–642, Jul. 2013.
- [7] T. E. Elizarova, S. V. Shtyleva, and T. V. Pleteneva, “Using near-infrared spectrophotometry for the identification of pharmaceuticals and drugs,” *Pharmaceutical Chem. J.*, vol. 42, no. 7, pp. 432–434, Jul. 2008.
- [8] W. Xinxin, M. Danzhuo, and Y. Yongjian, “Rapid qualitative analysis model for cetirizine hydrochloride Tablets by NIR using chemometric methods,” *Comput. Appl. Chem.*, vol. 29, no. 8, pp. 995–998, 2012.
- [9] L. P. Gong, W.-J. Wang, N. Yang, Z.-H. Zhang, and Y.-C. Xie, “Development of NIR method for rapid determination of cefalexin tablet,” *Chin. J. Pharmaceutical Anal.*, vol. 31, no. 8, pp. 1571–1574, 2011.
- [10] O. Y. Rodionova, A. V. Titova, K. S. Balyklova, and A. L. Pomerantsev, “Detection of counterfeit and substandard tablets using non-invasive NIR and chemometrics—A conceptual framework for a big screening system,” *Talanta*, vol. 205, Dec. 2019, Art. no. 120150.
- [11] E. Byvatov, U. Fechner, J. Sadowski, and G. Schneider, “Comparison of support vector machine and artificial neural network systems for drug/nondrug classification,” *J. Chem. Inf. Comput. Sci.*, vol. 43, no. 6, pp. 1882–1889, Nov. 2003.
- [12] W. Wu and D. L. Massart, “Artificial neural networks in classification of NIR spectral data: Selection of the input,” *Chemometric Intell. Lab. Syst.*, vol. 35, no. 1, pp. 127–135, Nov. 1996.
- [13] W.-D. Zhang, L. Ling-Qiao, H. Jin-Quan, F. Yan-Chun, Y. Li-Hui, H. Chang-Qin, and Y. Hui-Hua, “Drug discrimination by near infrared spectroscopy based on stacked sparse auto-encoders combined with kernel extreme learning machine,” *Chin. J. Anal. Chem.*, vol. 46, no. 9, pp. 1446–1454, 2018.
- [14] H. Yang, B. Hu, X. Pan, S. Yan, Y. Feng, X. Zhang, L. Yin, and C. Hu, “Deep belief network-based drug identification using near infrared spectroscopy,” *J. Innov. Opt. Health Sci.*, vol. 10, no. 2, Mar. 2017, Art. no. 1630011, doi: 10.1142/S1793545816300111.
- [15] L. Ling-Qiao, P. Xi-Peng, F. Yan-Chun, Y. Li-Hui, H. Chang-Qin, and Y. Hui-Hua, “Deep convolution network application in identification of multi-variety and multi-manufacturer pharmaceutical,” *Spectrosc. Spectral Anal.*, vol. 39, no. 11, pp. 3606–3613, 2019.
- [16] Z. Tong, Y. Zhou, and J. Wang, “Identifying potential drug targets in hepatocellular carcinoma based on network analysis and one-class support vector machine,” *Sci. Rep.*, vol. 9, no. 1, pp. 1–9, Dec. 2019.
- [17] Z. Xue-Bo and Y. Li-Hui, “Study of near-infrared spectroscopic conformity test for rapid examination of drug quality,” *Chin. J. Pharmaceutical Anal.*, vol. 31, no. 3, pp. 603–608, 2011.
- [18] J. Bromley, J. W. Bentz, L. Bottou, I. Guyon, Y. LeCun, C. Moore, E. Säckinger, and R. Shah, “Signature verification using a ‘Siamese’ time delay neural network,” in *Proc. 7th NIPS Conf. Adv. Neural Inf. Process. Syst.*, vol. 6. Denver, CO, USA. San Mateo, CA, USA: Morgan Kaufmann, 1993.
- [19] R. Hadsell, S. Chopra, and Y. LeCun, “Dimensionality reduction by learning an invariant mapping,” in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit. (CVPR)*, vol. 2, Jun. 2006, pp. 1735–1742.
- [20] S. Chopra, R. Hadsell, and Y. LeCun, “Learning a similarity metric discriminatively, with application to face verification,” in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2005, pp. 539–546.
- [21] Z. He, W. Su, Z. Bi, M. Wei, Y. Dong, and G. Xu, “The improved siamese network in face recognition,” in *Proc. Int. Conf. Intell. Comput., Autom. Syst. (ICICAS)*, Dec. 2019, pp. 443–446.
- [22] M. Gao, L. Jin, Y. Jiang, and B. Guo, “Manifold siamese network: A novel visual tracking ConvNet for autonomous vehicles,” *IEEE Trans. Intell. Transp. Syst.*, vol. 21, no. 4, pp. 1612–1623, Apr. 2020.
- [23] J. Liu, S. J. Gibson, J. Mills, and M. Osadchy, “Dynamic spectrum matching with one-shot learning,” *Chemometric Intell. Lab. Syst.*, vol. 184, pp. 175–181, Jan. 2019.
- [24] NIFDC. Beijing, China. *Technical Specification for Near-Infrared Spectrum Acquisition and Modeling 2014*. [Online]. Available: <https://wenku.baidu.com/view/00ac7da09e31433238689331.html>



ZHENG AN-BING received the M.S. degree from the Guilin University of Technology, China, in 2008. He is currently pursuing the Ph.D. degree with the Beijing University of Posts and Telecommunications, China. He is also working with the Big Data and Cloud Platform Project Group of the State Administration of Taxation, China. His research interests include machine learning, artificial intelligence, econometrics, and chemometrics.



YANG HUI-HUA received the Ph.D. degree from the East China University of Science and Technology, China, in 2005. He was a Postdoctoral Research Fellow with Tsinghua University from 2005 to 2007. He is currently a Professor with the School of Artificial Intelligence, Beijing University of Posts and Telecommunications, China. He has published more than 60 articles. His research interests include machine learning, image processing, and spectrum analysis. He serves as

the Director of China Instrument and Control Society (CICS) and the Vice Director of NIR Division of CICS. He is also a Senior Member of CCF and a member of ACM.



PAN XI-PENG received the Ph.D. degree from the Beijing University of Posts and Telecommunications, China, in 2019. He is currently an Assistant Researcher with the School of Computer Science and Information Security, Guilin University of Electronic Technology, China. He has published more than ten articles. His research interests include machine learning, near-infrared spectroscopy analysis, and medical image processing. He is also a member of CCF and ISAIR.



FENG YAN-CHUN received the Ph.D. degree from the Institute of Medicinal Biotechnology, Chinese Academy of Medical Sciences, in 2006. She is currently working as a Professor with the National Institutes for Food and Drug Control. Her research interests include chemical drug analysis and NIR in pharmaceutical application.

...



YIN LI-HUI has been involved in the quality analysis and quality evaluation of antibiotics, the application and research of instrumental analysis technology in drug quality control, since 1997, as well as the research of drug rapid detection technology. He is currently the Executive Director of the near-infrared spectroscopy branch of China Instrument and Instrument Society and a member of the National Mobile Laboratory Standardization Technical Committee Processing.