electronics

MDPI

Article

The Application Mode of Multi-Dimensional Time Series Data Based on a Multi-Stage Neural Network

Ting Wang [1,2], Na Wang [3], Yunpeng Cui [1,2] and Juan Liu [1,2,*]

[1] Agricultural Information Institute, Chinese Academy of Agricultural Sciences, Beijing 100081, China
[2] Key Laboratory of Big Agri-Data, Ministry of Agriculture and Rural Areas, Beijing 100081, China
[3] Unit 96962, Beijing 102206, China
* Correspondence: liujuan@caas.cn; Tel.: +86-151-1029-5951

Abstract: How to use multi-dimensional time series data is a huge challenge for big data analysis. Multiple trajectories of medical use in electronic medical data are typical time series data. Although many artificial-intelligence techniques have been proposed to use the multiple trajectories of medical use in predicting the risk of concurrent medical use, most existing methods pay less attention to the temporal property of medical-use trajectory and the potential correlation between the different trajectories of medical use, resulting in limited concurrent multi-trajectory applications. To address the problem, we proposed a multi-stage neural network-based application mode of multi-dimensional time series data for feature learning of high-dimensional electronic medical data in adverse event prediction. We designed a synthetic factor for the multiple -trajectories of medical use with the combination of a Long Short Term Memory–Deep Auto Encoder neural network and bisecting k-means clustering method. Then, we used a deep neural network to produce two kinds of feature vectors for risk prediction and risk-related factor analysis, respectively. We conducted extensive experiments on a real-world dataset. The results showed that our proposed method increased the accuracy by 5%~10%, and reduced the false rate by 3%~5% in the risk prediction of concurrent medical use. Our proposed method contributes not only to clinical research, where it helps clinicians make effective decisions and establish appropriate therapy programs, but also to the application optimization of multi-dimensional time series data for big data analysis.

Keywords: neural network; time series data; feature learning; electronic medical data; concurrent medical use; risk prediction

Citation: Wang, T.; Wang, N.; Cui, Y.; Liu, J. The Application Mode of Multi-Dimensional Time Series Data Based on a Multi-Stage Neural Network. *Electronics* **2023**, *12*, 578. https://doi.org/10.3390/electronics12030578

Academic Editors: Hiroki Tamura, Enzo Pasquale Scilingo and Keiko Sakurai

Received: 29 November 2022
Revised: 7 January 2023
Accepted: 19 January 2023
Published: 24 January 2023

1. Introduction

Owning to the popularization of the Internet of Things and the increasing computational capacity during the last decades, nowadays, the era of big data is beginning in many different fields, such as healthcare. Additionally, with the influence of national policies and social needs, healthcare big data has been a national strategy. As a consequence, lots of time series data are generated in healthcare, some of which constitute prescriptions filling for drugs included in electronic medical data. Socio-demographics and health status are also included in electronic medical data, and we consider it as patients' static data. Therefore, how to make full use of the multi-dimensional time series data in the electronic medical use is a huge challenge for healthcare big data analysis. In this paper, we discussed the application mode of multi-dimensional time series data in electronic medical data. We used medical-use trajectory to describe the patients' prescription fills for a period, as trajectory is defined as the chain of positions through which an item passes over time. The prescription fills of different drugs constitute multi-dimensional time series data, which are called the multiple trajectories of medical use in this paper.

Feature learning is an effective method for data mining. It can be used not only as an independent task, but also as the basis of other research. Features driven based on experts is

a traditional method, but is limited because it depends too much on the expert knowledge, which cannot help to discover the complex features hidden in electronic medical data. Deep learning in neural networks has been proved to be an advanced technology for feature learning, and has already contributed to major breakthroughs in image classification, speech recognition, object detection, and so on. Deep learning is widely used without domain knowledge, so it shows great promise in learning the features of complex electronic medical data.

There are a large number of successful applications in clinical research based on deep learning, such as disease diagnosis, clinical treatment, and medical-image processing, as shown in Table 1. The classification method for nuclear cataracts [1] based on eye images, with the combination of CNN- and RNN-based deep learning models, broke a record in terms of their accuracy in the diagnosis of cataracts. The CNN-based method [2], which automatically identifies mitosis in adenocarcinoma cell images, was the champion of the ICPR (International Conference Pattern Recognition) competition owing to its high accuracy. Meanwhile, with the enormous amount of data being generated in the biological field, there also are a number of breakthroughs in using deep learning to promote bioinformatics, including sequence analysis, structure prediction, biomolecular function prediction, structure reconstruction, and biomolecule interaction prediction, as shown in Table 1. For example, Ref. [3] proposed a method based on DNN and reached dermatologist-level performance in classifying skin cancer.

Table 1. Medical applications based on deep learning.

Application Category	The Problems to Solve	Deep Learning Models
Disease diagnosis	AD/MCI classification [4–8], cancer [9–13], cataract grading [1], chronic gastritis [14,15], geriatric depression [16] Alzheimer's disease classifier [17], autism spectrum disorder [13], pulmonary classification [18]	SAE, DBN, CNN, DBN + SVM, CNN + SVM, FCN, ANN, DNN
Clinical treatment	Drug discovery [19,20]	DNN
Biomedical image processing and diagnosis	Automatic image segmentation [21,22], MRI image reconstruction [23,24]	SAE, CNN
Medical data modeling	Brain development [25,26], emotional analysis [27]	DBN
Protein structure prediction and reconstruction	Secondary structure [28,29], three-level structure [30–32]	SAE, DBN, DNN, GSN, DST-NN
Sequence analysis	DNA marking [33,34], RNA variable shear [35–37]	DNN, CNN, SAE + DNN
Biomolecule property and function prediction	Structure properties [38], function prediction [39,40]	DNN, CNN, RNN
Biomolecule interaction prediction	Microarray gene expression [41], gene–disease interaction [42], disease–disease similarity, disease-variant network [43,44]	CNN, GCN

Although a lot of progress has been made in the relevant studies, there still remain some limitations, especially in the analysis of both the multiple trajectories of medical use and the patients' static data, such as electronic medical data. Researchers always employ two modes of feature learning for electronic medical data. One is to analyze patients' medical-use trajectories and static data independently, and the other one is to consider

the medical-use trajectories as some specific statistical indicators, such as the mean value for each medical-use trajectory. However, neither mode has a correlation between the medical-use trajectories and other static data in electronic medical data, nor between the different medical-use trajectories. Therefore, their ability to mine the hidden patterns of the multiple trajectories of medical use and analyze the entirety of the electronic medical data is limited.

To tackle the abovementioned limitations, we devised an application mode of multi-dimensional time series data based on a multi-stage neural network (MSNN) in electronic medical data, which shows how we use multi-trajectory of medical use data in risk prediction for concurrent medical use. This study makes the following contributions. First, for the feature learning of multiple trajectories of medical-use data, we designed neural network architecture with the combination of a Long Short-Term Memory (LSTM) neural network and a Deep Auto-encoder (DAE) neural network. Second, we used a deep neural network (DNN) to produce two feature vectors: one is to predict patients' risk of concurrent medical use, and the other one is to analyze the factors related to high risk of concurrent medical use. Third, we developed a finer-grained application model for the multi-dimensional time series data in the analysis with both time series data and static data. In risk prediction of concurrent medical use, the multiple trajectories of medical use are converted into synthetic factors to account for the temporal properties of medical-use trajectory and the correlation between different medical-use trajectories.

2. Multi-Stage Neural Network

According to the different types of electronic medical data, we constructed a LSTM-DAE neural network and DNN to perform the feature learning of the multi-dimensional medical-use data and the static data, respectively.

2.1. LSTM-DAE

The LSTM-DAE neural network is to perform feature learning of multiple trajectories of medical use, with the LSTM neural network as the hidden layer neurons of the DAE model. The LSTM neural network could handle the problem of long-term dependencies well by introducing a gate function into the cell structure. The DAE neural network is similar to Principal Component Analysis (PCA), but much more flexible and shows superior performance in feature self-learning. Then, the LSTM-DAE neural network realizes dimensional reduction, accounting for the temporal property of time series data, and performs well on the feature learning of medical-use trajectories.

2.1.1. LSTM

The LSTM neural network is a variation of a recurrent neural network (RNN). While RNN only has a hidden state to keep previous medical-use information in the trajectory and an output state to indicate the adverse events in concurrent medical use, LSTM contains a more complex gate mechanism, including an input gate, a forget gate, an output gate, and a memory cell with a hidden state. A gate is a structure controlling information moving through. A LSTM cell for input $x^{<t>}$ can be calculated as Equations (1)–(6).

Forget gate:

$$\Gamma_{\mathrm{f}} = \sigma(W_{\mathrm{f}}[a^{<t-1>}, x^{<t>}] + b_{\mathrm{f}}) \tag{1}$$

Update gate:

$$\Gamma_{\mathrm{u}} = \sigma(W_{\mathrm{u}}\left[a^{<t-1>}, x^{<t>}\right] + b_{\mathrm{u}}) \tag{2}$$

Output gate:

$$\Gamma_{\mathrm{o}} = \sigma(W_{\mathrm{o}}\left[a^{<t-1>}, x^{<t-1>}\right] + b_{\mathrm{o}}) \tag{3}$$

Memory cell:

$$c^{<t>} = \Gamma_{\mathrm{u}} * \widetilde{c}^{<t>} + \Gamma_{\mathrm{f}} * c^{<t-1>} \tag{4}$$

$$\widetilde{c}^{<t>} = \tan h(W_c\left[a^{<t-1>}, x^{<t>}\right] + b_c) \quad (5)$$

Hidden state:

$$a^{<t>} = \Gamma_o * \tan h(c^{<t>}) \quad (6)$$

$x^{<t>}$ is the new medical-use dose at each time step t in the medical-use trajectory. W, U, and b are learnable model parameters, and the values are between 0 and 1. The forget gate controls the extent to which the previous medical-use information is forgotten, the update gate controls how much medical-use information is updated in each state, and the output gate controls the proportion of medical-use information that will determine the health state. In detail, at time t, state $c^{<t>}$ and the activation $a^{<t>}$ are updated by the current input $x^{<t>}$, the previous state $c^{<t-1>}$, and previous activation $a^{<t-1>}$. Since LSTM adopts a memory cell to store the information at previous time step, it could accurately process the medical-use information by keeping the information in long-term medical use. Function $\sigma(\cdot)$ is the sigmoid activation function, defined as $e^x/(1+e^x)$. Function $\tan h(\cdot)$ is the activation function. The operator $\circ$ is element-wise multiplication. To deal with the long-term gradient vanishing and explosion problem, LSTM automatically updates and forgets the state in each cell.

2.1.2. DAE Neural Network

AE [45] extracts a feature vector in lower dimensions from a huge of unlabeled high-dimensional data and cannot be affected by humans. A standard AE is a symmetrical three-layer feedforward neural network including an input layer, a hidden layer, and an output layer. The input and output layer have the same dimensions, while the hidden layer has less dimensions. The objective of AE is to establish identical mapping and minimize the difference between the input and output, and then the hidden layer can represent the features of the input. Therefore, the mapping from the input to the feature vector is called encoding, and the mapping from the feature vector to the output is called decoding.

DAE is a usually used as a variation of AE [46]. Compared with AE, there are more hidden layers in the architecture of DAE, which make the feature vector more reliable and valid. In DAE, the middle layer that has the least neurons is considered as the feature vector. The neuron number from the input to feature vector decreases layer by layer and from the feature vector to the output is opposite. With adding the number of neuron layers and the number of neurons in each layer, turning AE into a deep network performs the feature transformation of the input layer by layer and can increase the non-linearity of the neural network.

While DAE has proven to be useful for feature extraction, it overlooks the temporal correlation between different time steps on the trajectory, e.g., the contextual and cumulative medical-use information. To explicitly extract the temporal correlation between different time steps on medical-use trajectory, LSTM can be used to replace the feedforward networks in both the encoder and decoder. Figure 1 illustrates the architecture of LSTM-DAE, whose training process is as follows:

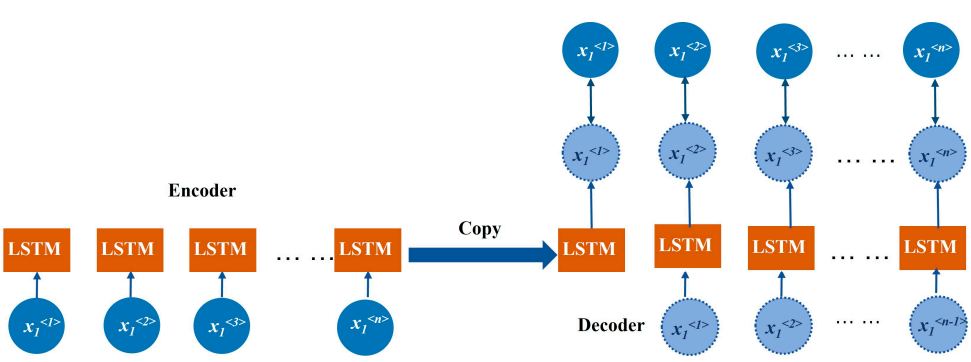

Figure 1. The architecture of LSTM-DAE.

(1) Take the medical-use trajectory as input for the encoder network of LSTM-DAE one by one, then consider $a^{<m>}$ obtained by $x^{<m>}$, which is the last one, as the representation of input. Take Relu as the activation function.

(2) Take $a^{<m>}$ as the input for the decoder network of LSTM-DAE. Mean Squared Error (MSE) is the optimization function once the network is trained, which is calculated as Equation (7). Kullback–Leibler divergence is used to perform the sparse reconstruction of the encoder network, calculated as Equations (8)–(10).

$$v_{\mathrm{MSE}} = \sum_{m=1}^{\mathrm{M}} \| a^{<m>} - \widetilde{a}^{<m>} \| \tag{7}$$

$$L_{\mathrm{sparse}} = v_{\mathrm{MSE}} + \beta \sum_{j=1}^{\mathrm{J}} \mathrm{KL}(\rho || \overline{\rho}) \tag{8}$$

$$\overline{\rho} = \frac{1}{\mathrm{N}} \sum_{n=1}^{\mathrm{N}} h_i \tag{9}$$

$$\mathrm{KL}(\rho \| \overline{\rho}) = \rho \log \frac{\rho}{\overline{\rho}} + (1-\rho) \log \frac{1-\rho}{1-\overline{\rho}} \tag{10}$$

β is the weight parameter that controls the sparseness, $\overline{\rho}$ is the average activity of the hidden layer, and $\mathrm{KL}(\rho \| \overline{\rho})$ is the relative entropy between the two variables.

2.2. DNN

Deep neural network (DNN) is a typical deep learning architecture that is widely used in literature and technology. It is a feedforward network, where the input from the input layer flows though hidden layers and reaches the output layer. The output of each layer is calculated as Equation (11).

$$\mathrm{y} = g(Wx + b) \tag{11}$$

x, g, W, and b are the input, the activation function, the weight matrix, and the bias vector, respectively. The cross-entropy loss function, including the regularization item, was used as the optimization once the network was trained.

$$\mathrm{J} = \frac{1}{\mathrm{N}} \sum_{n=1}^{\mathrm{N}} \sum_{k=1}^{\mathrm{K}} \mathrm{L}(y_{nk}, z_{nk}) + \frac{1}{2\mathrm{N}} \sum_{k=1}^{\mathrm{K}} \sum_{m=1}^{\mathrm{M}} (w_{km})^2 \tag{12}$$

N, K, and M are the numbers of training samples, data features, and the neurons, respectively. y_{nk} and z_{nk} are the expected value and the observed value of the output, respectively.

2.3. The MSNN Method

Taking the different types of data in electronic medical data into account, we propose the MSNN method, constructing LSTM-DAE and DNN, to perform feature learning of multiple trajectories of medical-use data and multi-dimensional static data, respectively. Then, we predict the risk of concurrent medical use and analysis its relative factors. The architecture of the MSNN method is shown in Figure 2, consisting of the following sections: the construction of the synthetic factor for multiple trajectories of medical use, data merging, the construction of feature vectors for risk prediction of concurrent medical use and its related factors, the risk prediction for adverse events, and the analysis of its related factors.

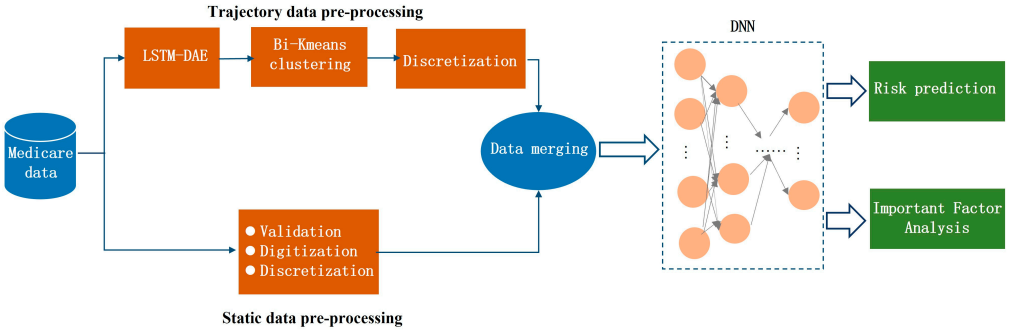

Figure 2. Architecture of the MSNN method.

2.3.1. Synthetic Factor for Multiple Trajectories of Medical Use

Based on feature learning for multiple trajectories of medical use with LSTM-DAE, bisecting kmeans (bi-kmeans) clustering method was used to construct the synthetic factor for multiple trajectories of medical use. The process consists of four steps as follows: (1) calculating the distance between different patients' multi-trajectories of medical use, (2) k value selection in bi-kmeans clustering, (3) bi-kmeans clustering, (4) constructing the synthetic factor for multiple trajectories of medical use.

1. Calculating the distance between different multi-trajectories of medical use

Taking into consideration that the value domains of different trajectories will affect the distance measure between different patients' multiple trajectories of medical use, we carried out the data normalization first, calculated as Equations (13) and (14).

$$x_1 = \frac{x_1 - \min(x_1)}{\max(x_1) - \min(x_1)} \tag{13}$$

$$x_2 = \frac{x_2 - \min(x_2)}{\max(x_2) - \min(x_2)} \tag{14}$$

We used Euclidean distance to measure the distance between different patients' multiple trajectories of medical use, which is calculated as Equations (15)–(17). For each dimensional trajectory in multiple trajectories of medical use, we computed the distance between two patients. Then, we treated the average of the two distances as the final distance between two patients.

$$\mathrm{d}(x,y) = \frac{\mathrm{d}_{\mathrm{euc}}(x_1,y_1) + \mathrm{d}_{\mathrm{euc}}(x_2,y_2)}{2} \tag{15}$$

$$\mathrm{d}_{\mathrm{euc}}(x_1,y_1) = \sqrt{\sum_{i=1}^{\mathrm{N}} (x_{1i} - y_{1i})^2} \tag{16}$$

$$\mathrm{d}_{\mathrm{euc}}(x_2,y_2) = \sqrt{\sum_{i=1}^{\mathrm{N}} (x_{2i} - y_{2i})^2} \tag{17}$$

2. k value selection

We decided to use the k value in bi-kmeans clustering by using the gap statistic method, which is calculated as Equation (18).

$$W_n = \sum_{l=1}^{n} \sum_{q,r \in C_l} \frac{1}{2|\mathrm{C_l}|} d(q,r) \tag{18}$$

$$\mathrm{Gap}_n = \frac{1}{\mathrm{B}} \sum_{b=1}^{\mathrm{B}} \log(W_{nb}) - \log(W_n) \tag{19}$$

$$d_n = \sqrt{\frac{1}{\mathrm{B}}\sum_{b=1}^{\mathrm{B}}(\log(W_{nb}) - \bar{l})} \tag{20}$$

$$\bar{l} = \frac{1}{\mathrm{B}}\sum_{b=1}^{\mathrm{B}}\log(W_{nb}) \tag{21}$$

$$s_n = \sqrt{1 + \frac{1}{\mathrm{B}}}d_n \tag{22}$$

$$k = \max(\mathrm{Gap}(n) - (\mathrm{Gap}(n+1) - s_{n+1})) \tag{23}$$

where n is the number of clusters, $|\mathrm{C_l}|$ is the number of samples in $\mathrm{C_l}$, $d(q,r)$ is the distance between q and r, W_n is cluster dispersion, and B is the number of data sets. The processing of k selection is as follows [47]:

(1) Set the domain of k as 1~N.
(2) For each value within the value domain, randomly generate samples with the same number as initial samples according to a uniform distribution, then perform bi-kmeans clustering and obtain the value of W_k.; repeat this step 100 times.
(3) Select the maximum value calculated by $\mathrm{Gap}(n) - (\mathrm{Gap}(n+1) - s_{n+1})$ as the optimal value of k.

3. Bi-kmeans clustering

Bi-kmeans algorithm, as a classic unsupervised learning approach, is a modification of the k-means clustering algorithm. It has been widely used for identifying groups in many domains, aiming to group data samples that are similar to each other into the same group. Compared with other clustering algorithms, bi-kmeans is very effective and less susceptible to initialization problems.

(1) Bisect all data samples using bi-kmeans clustering.
(2) Select the cluster with the minimum value of sum of squared errors (SSE), which is calculated as Equation (24), to perform bisecting.
(3) Repeat the second step until the number of clusters meets k as decided in the section of k value selection.

$$v_{\mathrm{SSE}} = \sum_{l=1}^{k}\sum_{j\in \mathrm{C_l}}(x_j - \overline{x_l})^2 \tag{24}$$

4. Synthetic factor construction for multiple trajectories of medical use

After the multiple trajectories of medical use in electronic medical data are grouped into k groups by the bi-kmeans clustering algorithm, we denoted each group with one-hot encoding respectively, which is considered as a synthetic factor for multiple trajectories of medical use.

2.3.2. Data Merging

First, discretize the string data in static data with one-hot encoding, e.g., map 'male' and 'female', which belong to gender characteristics, to 01 and 10, respectively. Then, merge the static data with the synthetic factor for multiple trajectories of medical use as the input data trained in DNN.

2.3.3. Feature Vectors of Electronic Medical Data

In order to predict the risk of concurrent medical use and analyze its related factors, we applied DNN to construct two kinds of feature vectors of electronic medical data: feature vector of risk prediction and feature vector of risk-related factor analysis. The first one constitutes the new feature expression of electronic medical data, and the second one represents the association between patients' characteristics with the risk of concurrent medical use.

- Feature vector of risk prediction. It consists of new feature values constructed by all neurons on the last layer of DNN.
- Feature vector of risk-related factors. The weight values of all hidden layers in DNN were used to construct a weight network of characteristics in electronic medical data, and the weight value of each node was calculated by summing products of the values of its upper node and corresponding edge. The structure of the weight network is shown as Figure 3.

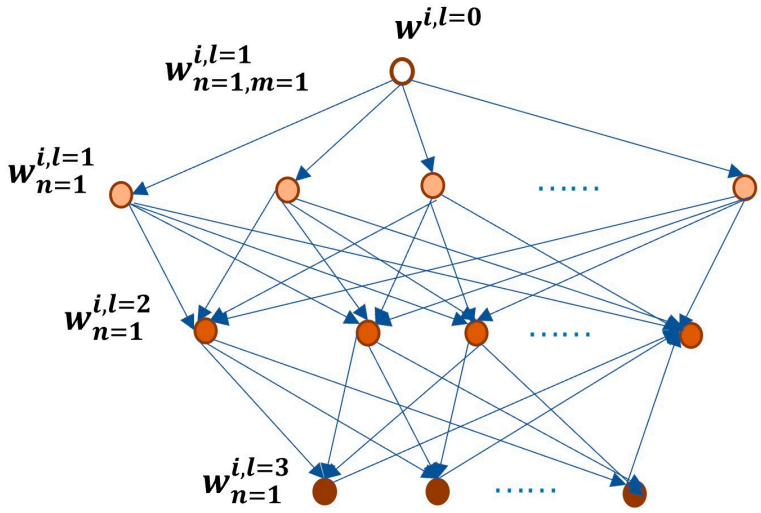

Figure 3. Weight network structure of patients' characteristics.

L, n, m are the sequence number of the node on the current layer and the upper one in the network, $w^{i,l=0}$ is the initial weight value of the patient's characteristic I, which is defined as 1, $w^{i,l}_n$ is the weight value of the lth layer and the nth node, and $w^{i,l}_{n,m}$ is the weight value of the edge between the nth node on the ith layer and the mth node on the lth layer. The processing is as follows:

① $l = 1$. $w^{i,l=1}_{n=1}$ is the weight value of the edge between the current node and the initial node, which is $w^{i,l=1}_{n=1,m=1} \times 1$.

② $l = 2$. $w^{i,l=2}_{n=1}$ is the sum of products of the weight values of the current node and the edge between the current node and its upper node, which is $\sum_{m=1}^{M} w^{i,l=1}_{n=1} \times w^{i,l=2}_{n=1,m}$.

③ $l = 3$. Accordingly, the weight value of the nth neuron is $w^{i,l=3}_n = \sum_{m=1}^{M} w^{i,l=2}_{n,m} \times w^{i,l=3}_{n,m}$.

The weight values of all the neurons on the last layer constitutes feature vector of risk-related factor analysis.

3. Experiment and Results

3.1. Data Sources

We performed our analysis on the Medicare claim data from a 5% national representative sample of Medicare beneficiaries from 2013 to 2016. Medicare claim data include master beneficiary summary files with socio-demographics, health status, and part D drug event files. Socio-demographic factors include age, sex, race/ethnicity, disability status, and receipt of low-income subsidy and dual Medicaid eligibility. Health status factors include patient comorbidity, serious mental illness, anxiety disorders, and numbers of prescription fills for drugs and other prescriptions not mentioned above. We initially identified 10,000 random Medicare beneficiaries who filled at least one opioid (OPI), and about 23.5% patients have OPI overdose events. Due to lack of complete medical information in the Medicare data, 6171 patients remained. Then for each patient, we construct multiple trajectories for OPI and benzodiazepine (BZD) utilization during the 90 days prior

to the overdose date or the end date. We calculated the single daily dose (SDD) for OPI using a morphine milligram equivalent (MME), and for BZD used the minimum effective daily dose, which was defined as DDD. We calculated MME for each OPI prescription by multiplying the quantity dispensed by strength in milligrams and a conversion factor provided by the Centers for Disease Control and Prevention (CDC). We created a daily diary of opioid and benzodiazepine use for each patient by summing up the total daily SDD for opioids and benzodiazepines, respectively. For example, the total SDD for BZD for a day is two if a person had a 1 mg alprazolam prescription overlapped with a 2.5 mg lorazepam prescription on the same day.

Then each patient has 139 dimensions of characteristic information, including 90 dimensions of medical-use trajectory and 49 dimensions of static information. The bi-kmeans algorithm grouped all patients into four groups. Then the four groups of patients were denoted with one-hot encoding, respectively, 0001, 0010, 0100, and 1000, which is considered as synthetic factor for multi-trajectory medical use. After data merging, the Medicare data were converted from 139 dimensions to 62 dimensions.

Of the Medicare data, 80% were selected as a training set, and 20% were used as test sets. Taking the patients' characteristics as the input, whether the patient has overdose events as the output, and 62: 40:20:2 as the training network structure of the DNN model.

3.2. Performance Measurement Metrics

We adopted silhouette coefficient (SC) and consensus clustering method to assess the clustering results. Additionally, we adopted precision rate (PR), recall rate (RR), false rate (FR), and F_1 score as the evaluation metrics, which are commonly used in binary predictive studies.

- Silhouette coefficient.

$$v_{\mathrm{sc}} = \overline{s(i)} = \frac{1}{\mathrm{N}} \sum_{i=1}^{\mathrm{N}} s(i) \tag{25}$$

$$s(i) = \frac{b(i) - a(i)}{\max\{a(i), b(i)\}} \tag{26}$$

$a(i)$ is the dissimilarity in one cluster, which is calculated as the average distance between sample i and other samples in the same cluster, and $b(i)$ is the dissimilarity between different clusters, which is calculated as the average distance between sample i and other samples in other clusters.

- Precision assessed the proportion of patients the model classified as high risk among those who actually suffer from overdose events.

$$v_{\mathrm{PR}} = \frac{n_{\mathrm{TP}}}{n_{\mathrm{TP}} + n_{\mathrm{FP}}} \tag{27}$$

- The false rate assessed the proportion of patients which the prediction method classified as high risk among those that did not actually have overdose events.

$$v_{\mathrm{FR}} = \frac{n_{\mathrm{FP}}}{n_{\mathrm{TP}} + n_{\mathrm{FP}}} \tag{28}$$

- Recall assessed the proportion of patients identified by the prediction method among those that actually had overdose events in the dataset.

$$v_{\mathrm{RR}} = \frac{n_{\mathrm{TP}}}{n_{\mathrm{TP}} + n_{\mathrm{FN}}} \tag{29}$$

- F_1 score is a comprehensive measure of accuracy integrating both precision and recall.

$$v_{F_1} = \frac{2n_{TP}}{2n_{TP} + n_{FP} + n_{FN}} \tag{30}$$

n_{TP} is the number of patients classified as high risk but who actually had overdose events, n_{TN} is the number of patients classified as no risk, n_{FN} is the number of patients classified as no risk but actually had overdose events, and n_{FP} is the number of patients classified as high risk but actually have not had overdose events.

3.3. Utilization Pattern Analysis of Concurrent Medical Use

3.3.1. DNN Parameter Analysis

We analyzed the parameters of learning rate (α) and the number of hidden neurons (m) used in LSTM. As shown in Figure 4, The loss function of L_{sparse} reached the minimum value at $\alpha = 0.00001$, and it achieved the fastest convergence rate at $m = 75$ while $\alpha = 0.00001$. Consequently, we chose $\alpha = 0.00001$ and $m = 75$ in the LSTM training.

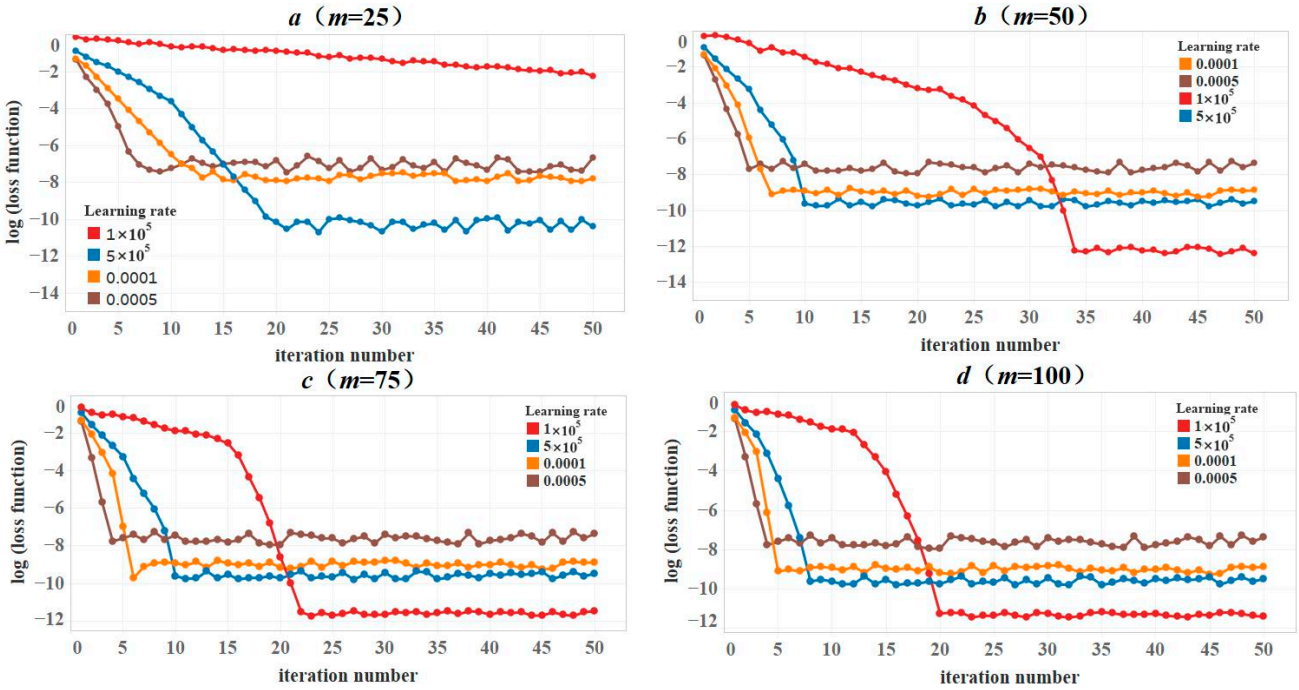

Figure 4. The changes of log(L_{sparse}) with different parameters in LSTM.

3.3.2. Clusters Quality Assessment

We used SC to evaluate the quality of clustering. SC ranges between −1 and 1, where a higher value indicates more coherent and better performance of the clustering result. The results are from multiple iterations ($n = 5$) of an inner-loop bi-kmeans clustering, which is shown as Figure 5. Figure 5A shows the SC values at different k values, and Figure 5B is the SC value distribution at $k = 4$. We can see that the SC value reached the maximum at $k = 4$, and the SC values of most patients at $k = 4$ were higher than 0.75, which performed best among all the k values. It shows that $k = 4$ is the optimal value in our study.

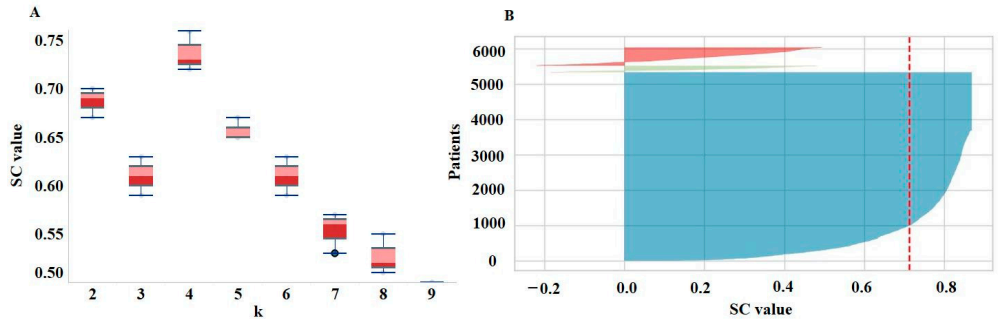

Figure 5. SC analysis.

We compared the clustering performance of bi-kmeans (BK) to methods built with five classic models: a hierarchical clustering method (HCM), kmeans, kmedoids (KM), and bi-kmeans without data normalization (BW). Additionally, in the comparison, we used four kinds of distance-measurement methods: Euclidean, Pearson, longest common subsequence (LCSS), dynamic time wrapping (DTW) and edit distance on real sequence (EDR). It can be seen from Table 2 that the BC value of the BK method is significantly higher than other methods and increased by at least 15%; the BC value of the BK method is higher than the BW method without data normalization. Finally, compared with other distance-measurement methods, the BW method with Euclidean distance measurements is the highest.

Table 2. SC values of different clustering methods.

Distance	HCM	Kmeans	KM	BW	BK
Euclidean	0.65	0.56	0.63	0.63	0.88
Pearson	0.41	0.49	0.59	0.61	0.72
LCSS	0.55	0.52	0.53	0.59	0.70
DTW	0.63	0.54	0.47	0.53	0.67
EDR	0.57	0.58	0.51	0.55	0.66

3.3.3. Clustering Stability Assessment

In order to assess clustering stability, we used the consensus clustering method to incorporate results from multiple iterations (n = 50) of an inner-loop bi-kmeans clustering algorithm. Consensus clustering method generates a consensus value for each pair of patients by calculating the proportion of times that two patients are assigned to the same group among total number iterations of bi-kmeans clustering executed. Figure 6A is a heatmap of the consensus matrix for k = 4. The values in the consensus matrix, shown as rectangles in the figure, range from 0 to 1 marked by white to blue. The color of white shows the proportion of these memberships clustered together is 0, and the blue shows the proportion of these memberships clustered together is 1. Clearer heatmaps indicate more stable clustering. From the figure we can see that the separation of colors in the heatmap is very clear, which indicates good stability of the clustering by the MSNN method. Figure 6B shows the cumulative distribution functions (CDF) of the consensus matrix for each k, which is indicated by different colors. We can see that the CDF reached an approximate maximum at k = 4, where consensus and cluster confidence at a maximum. Figure 6C shows the relative change in area under the CDF curve comparing k and $k - 1$. Figure 6D shows the cluster-consensus value of clusters at each k, which is the average of all pairwise consensus values between members in the same group. Higher values indicate a cluster has higher stability. From all of the figures, we can see that the clustering result at k = 4 had better stability than other k values.

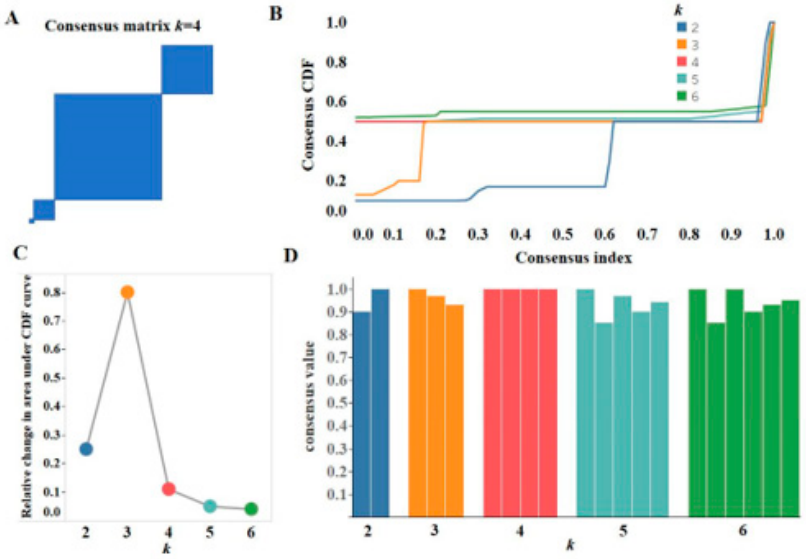

Figure 6. Consensus clustering method with bi-kmeans results.

3.4. Pattern Analysis of Concurrent Medical Use

1. Concurrent medical-use patterns

Four distinct trajectories of medical use were identified, which incorporate both medical dose and duration changes over time. Figure 7 illustrates the results along with the predicted probability of each multi-trajectory subgroup of medical use within 90 days preceding the opioid overdose events or the first OPI prescription day. The value at a certain time of the trajectory is the centroid of the data points in the same group. The Y axis (above) indicates the daily average dose of OPI in different groups, and the Y axis (below) indicates the daily average dose of different groups for BZD There were distinct patterns in terms of the dose level of concurrent opioid and benzodiazepine use prior to the first OPI prescription day, and each one is denoted as the synthetic factor of concurrent medical use in the form of one hot, which is shown as follows.

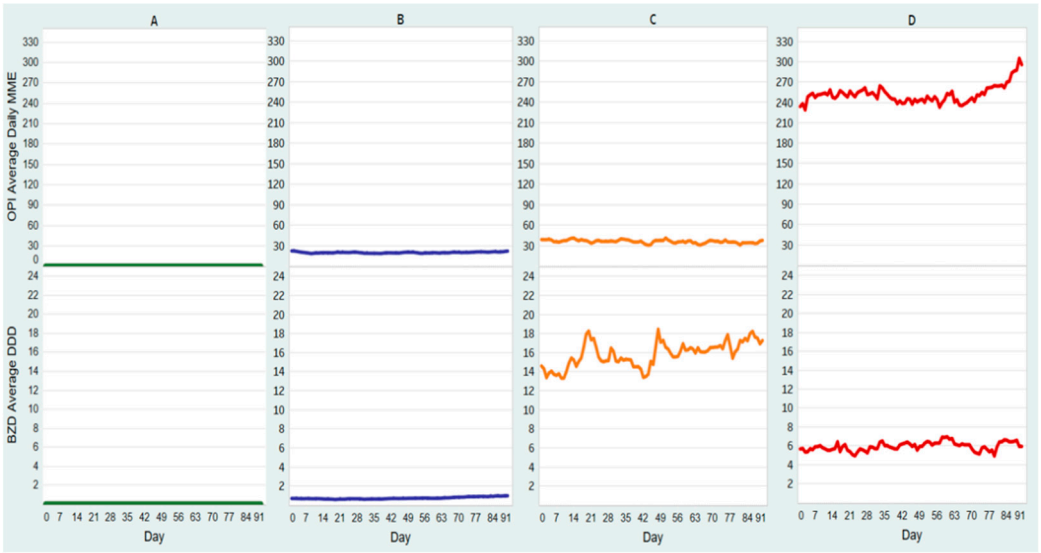

Figure 7. Multi-trajectory patterns of medical use.

- Pattern A, whose synthetic factor is denoted as 0001. the majority (76.6% of the patients) only had minimal opioid and benzodiazepine use (MME < 30 and benzodiazepine SDD < 1).
- Pattern B, whose synthetic factor is denoted as 0010. 5.2% patients had consistent low-dose opioid and very-high-dose benzodiazepine use (30 $\leq$ MME < 50 and 6 < benzodiazepine SDD < 8).
- Pattern C, whose synthetic factor is denoted as 0100. 12.2% patients had consistent high-dose opioid use and low-dose benzodiazepine use (60 $\leq$ MME < 90 and benzodiazepine SDD < 1).
- Pattern D, whose synthetic factor is denoted as 1000. 6.0% patients had consistent very-high-dose opioid and high-dose benzodiazepine use (MME > 210 and benzodiazepine SDD > 2).

2. Comparison of characteristics that contribute to different multi-trajectories

Compared to patients with minimal opioid and benzodiazepine use, patients in pattern C were more likely to be over 65 years old (40.2% vs. 32.8%), and to have more alcohol use disorders (7.1% vs. 4.3%). Additionally, the patients in the other three patterns are more likely to be white, and to have more opioid use disorders. All these differences were statistically significant ($p < 0.0001$). Figure 8 shows the ranked characteristics by the mean standardized proportional distribution difference between the concurrent medical-use pattern pairs. The pattern results are from both the MSNN method and group-based multi-trajectory models (GBTM). The proportions of the characteristic (e.g., COAG) in the individual pattern and both patterns in the graph as a whole are calculated separately. A value of 0.1 for the variable value of proportional distribution (x-axis) signifies that the proportion of the characteristic in the pattern was 10% higher than the proportion in both patterns as a whole. Figure 8A shows the difference between multi-trajectory A and B, Figure 8B shows the difference between multi-trajectory A and C, Figure 8B shows the differences between multi-trajectory A and D. It is seen that the overall results from both methods yielded similar findings and the characteristics of patients in different patterns are distinguishable, which validate the effectiveness of our proposed clustering procedure in the MSNN method.

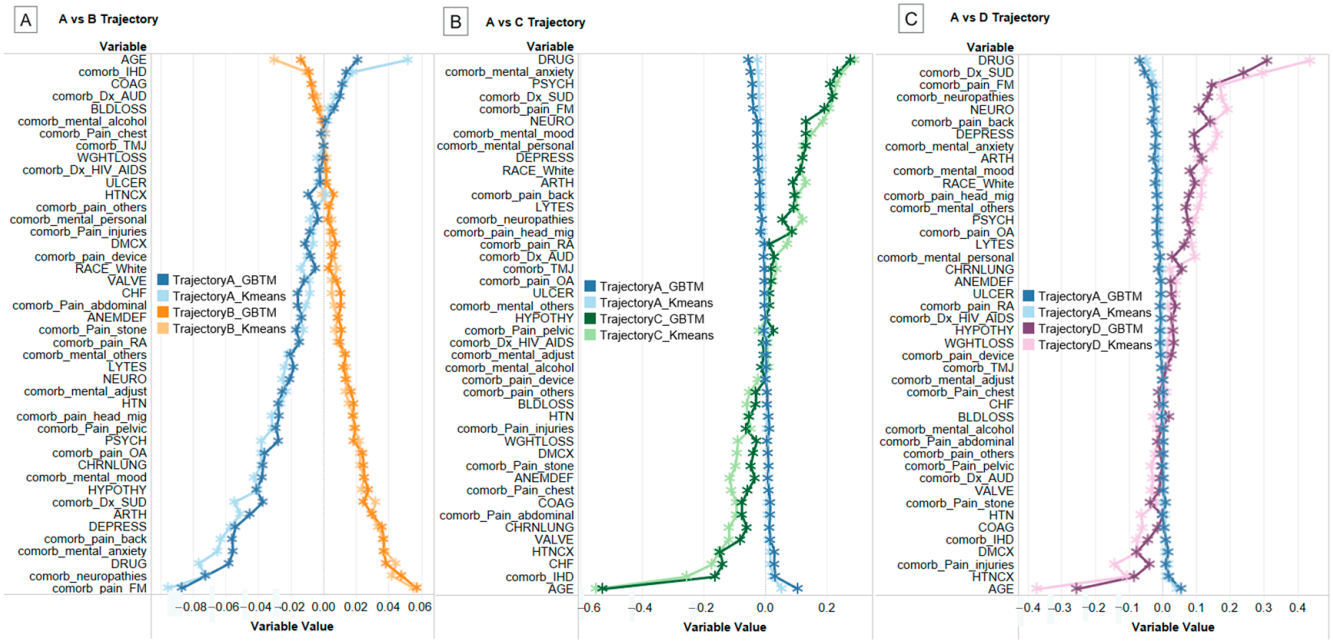

Figure 8. Comparison of va ttriables that contribute to trajectories.

3.5. Risk Prediction and the Related Factors

Based on the feature vector of risk prediction and the feature vector of risk-related factors analysis, we predict the patients' risk of overdose events, and analysis the correlation between the patients' characteristics and overdose events.

3.5.1. Risk Prediction

We used DNN to predict patients' risk of overdose events in the MSNN method. To show the effectiveness of our design, we benchmark with other three methods: Random Forest (RF), RNN, SVM, and all of the four methods were performed both with and without synthetic factor of concurrent medical use. Figure 9 shows the methods' performance, the color represents different prediction methods, and the triangles and strips represent the model with and without synthetic factor of concurrent medical use respectively. The performance of methods with synthetic factor of concurrent medical use significantly outperformed the methods without synthetic factor in F_1 score, prediction, and recall by the range between 5% and 10%, among which the value of DNN is the highest. Additionally, the FR of the methods with synthetic factors was reduced by a range between 3% and 5%, among which the value of DNN is the lowest. It shows that the mode we proposed in the MSNN method significantly increased the prediction accuracy by converting multiple trajectories of medical use into synthetic factor for multiple trajectories of medical use. The MSNN method is effective and optimizes the risk prediction for concurrent medical use. The doctors can predict the risk of adverse events with the patients' electronic medical data, and adjust the medical refill based on the results, which can effectively reduce risk of adverse events.

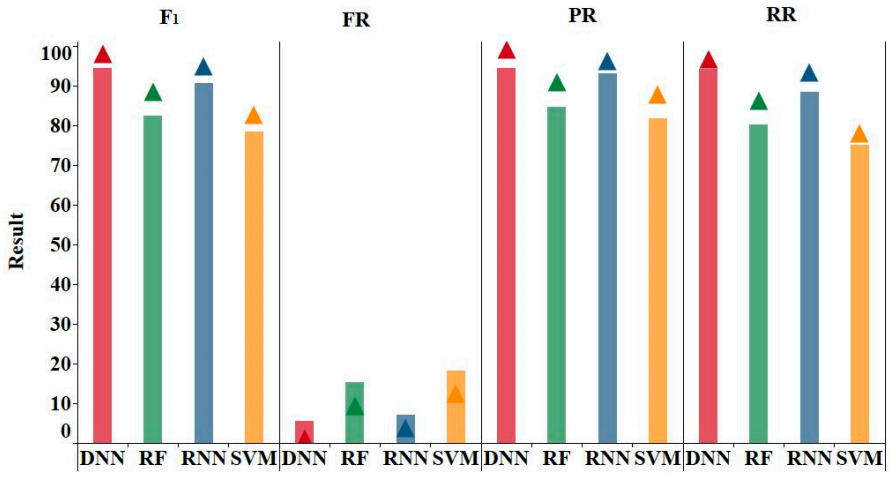

Figure 9. Risk prediction for concurrent medical use.

3.5.2. Correlation between Concurrent Medical-Use Patterns and Adverse Event Risk

Based on the weight values of the patients' different characteristics in the weight network of patients' characteristics, as shown in Section 2.3.3, we analyze the correlation between concurrent medical-use patterns and overdose events risk, as Figure 10 shows. Orange and blue dots represent concurrent medical-use patterns and other characteristics respectively. The greater the absolute value is, the more the correlation between this characteristic and the risk of overdose events. Then if the value of the concurrent pattern is greater, the patients in the relative group take more risk of overdose events. It can be seen from the figure that patients in the concurrent medical-use pattern D take the most risk, followed by B and C, and A. The different levels of risk may be valuable to identifying the concurrent medical-use type most associated with the risk of adverse events to better guide clinical care and target interventions.

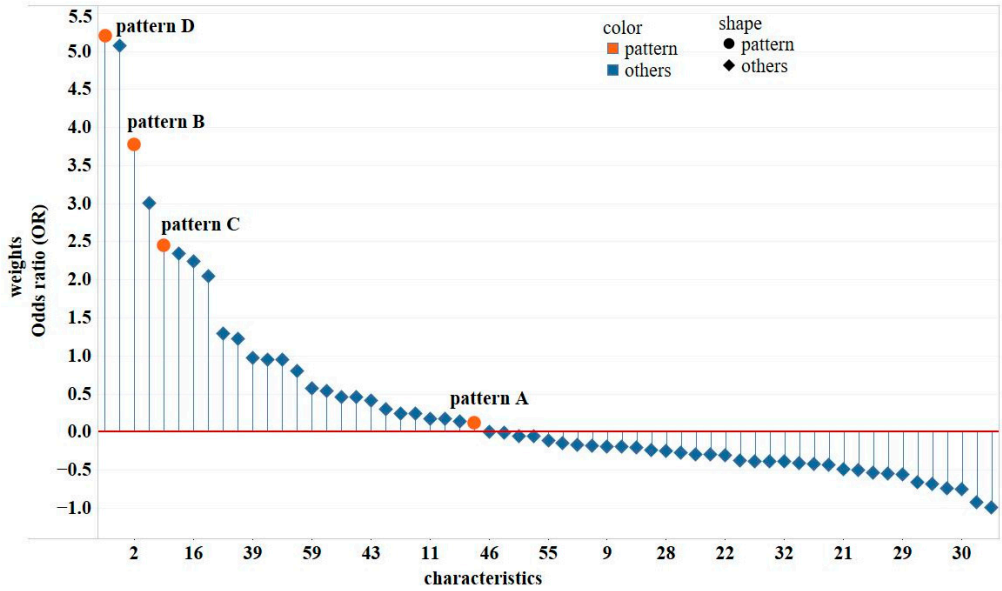

Figure 10. Important factors for risk prediction.

3.5.3. Relative Importance Analysis of Characteristics for Risk Prediction

We take the final weight value of each characteristic in the weight network constructed in Section 2.3.3 as the relative importance for risk prediction. Higher values indicate more importance in risk prediction. According to the final weight values, we selected the top 15 characteristics as important factors for the risk prediction. We analyzed the importance from two aspects: one is comparing the performance between different hidden layers in DNN and the other one is between DNN and the other four methods: LASSO, RF, RNN, SVM.

As shown in Figure 11, the row indicates the prediction method, the line indicates the characteristic number, and the deep red color indicates that the characteristic on the line was selected as an important factor by the method on the row. B, C, and D indicate concurrent medical-use patterns, respectively. It is seen that characteristic 17 (opioid use disorder), characteristic 45 (sleep disorder), and concurrent medical-use pattern D were the most important for risk prediction, with the proportion reaches 5/5. There are 10 identical important factors selected by all hidden layers in DNN. Compared with the second hidden layer, the importance factors selected by the third layer are more similar to the final selection by DNN.

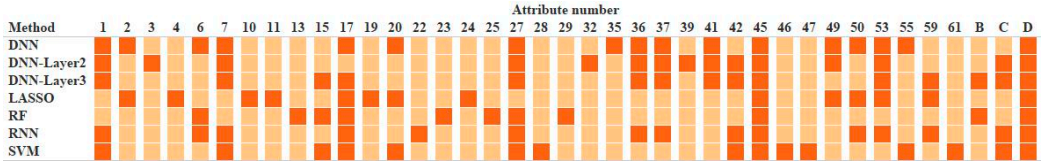

Figure 11. Important factors for risk prediction.

4. Conclusions

In this paper, a mode called MSNN based on multi-stage neural network was proposed to deal with the application of multi-dimensional time series data. We validated it in the scene of risk of adverse event prediction for concurrent medical use. The LSTM-DAE neural network is adopted to perform feature learning of multiple trajectories of medical use, and combined with bi-kmeans clustering method to produce synthetic factor for multiple trajectories of medical use. The DNN is used to produce two kinds of feature vectors for risk prediction and risk-related factor analysis, respectively. Compared with other methods, the MSNN method performs better on effectively extracting patients' features

in electronic medical data, and predicting the risk of concurrent medical use. Our study contributes to the application of multi-dimensional time series data, especially how to use multiple trajectories of medical use in risk prediction for concurrent medical use. To our best knowledge, we are the first to convert the multiple trajectories of medical use into a synthetic factor and merge the synthetic factor with other characteristics for the risk of concurrent medical-use prediction based on a multi-stage neural network. Our findings shed direct implications to health practitioners. It may provide important implications and guidance in clinical practice. At present, this paper only analyzes the concurrent medical use of two medications. In future research, both the number of medications and the reverse events will be added to cover more clinical scenarios.

Author Contributions: Conceptualization, T.W.; methodology, T.W. and Y.C.; validation, T.W.; formal analysis, N.W.; investigation, T.W. and J.L.; resources, Y.C.; data curation, T.W.; writing—original draft preparation, T.W.; writing—review and editing, Juan.Liu.; visualization, N.W.; supervision, T.W.; project administration, Juan.Liu.; funding acquisition, Y.C. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by [Cooperative Innovation Task of Chinese Academy of Agricultural Sciences] grant number [No. CAAS-ASTIP-2016-AII] and [Special Fund of National Science and Technology Library] grant number [No. 2022XM2805] and [Beijing Innovative Team Project of 2022 Modern Agricultural Technology System] grant number [No. BAIC10-2022-E10] and [China Scholarship Council Project] grant number [No. 201803250020].

Data Availability Statement: The data is unavailable due to privacy or ethical restrictions.

Conflicts of Interest: The authors declare no conflict of interest.

Abbreviations

The following abbreviations are used in this manuscript:

MDFL	multi-stage deep learning based feature learning method
LSTM	Long Short-Term Memory
AE	Auto-encoder
PCA	Principal Components Analysis
DNN	Deep neural network
Bi-kmeans	Bisecting K-means
MME	morphine milligram equivalent
CDC	the Centers for Disease Control and Prevention
SC	silhouette coefficient
CDF	cumulative density function
OPI	opioid
BZD	benzodiazepine
ICPR	International Conference Pattern Recognition
SVM	support vector machine
SAE	stacked auto-encoder
DBN	dynamic bayesian network
FCN	fully convolutional neural network
ANN	artificial neural network
GSN	generative stochastic network
DST-NN	deep spatio-temporal neural network
GCN	Graph convolutional network

References

1. Gao, X.T.; Sthephen, L.; Wong, Y.T. Automatic Feature Learning to Grade Nuclear Cataracts Based on Deep Learning. *IEEE Trans. Bio-Med. Eng.* **2015**, *62*, 2693–2701. [CrossRef] [PubMed]
2. Ciresan, D.C.; Giusti, A.; Gambardella, L.M.; Schmidhuber, J. Mitosis Detection in Breast Cancer Histology Images with Deep Neural Networks. *Med. Image Comput. Comput. Assist. Interv.* **2013**, *16*, 411–418. [PubMed]
3. Esteva, A.; Kuprel, B.; Novoa, R.A.; Ko, J.; Swetter, S.M.; Blau, H.M.; Thrun, S. Dermatologist-level classification of skin cancer with deep neural networks. *Nature* **2017**, *542*, 115–118. [CrossRef] [PubMed]

4. Suk, H.-I.; Lee, S.-W.; Shen, D. Hierarchical feature representation and multimodal fusion with deep learning for AD/MCI diagnosis. *Neuroimage* **2014**, *101*, 569–582. [CrossRef]
5. Ithapu, V.K.; Singh, V.; Okonkwo, O.C.; Chappell, R.J.; Dowling, N.M.; Johnson, S.C. Imaging-based enrichment criteria using deep learning algorithms for efficient clinical trials in mild cognitive impairment. *J. Alzheimer's Assoc.* **2015**, *11*, 1489–1499. [CrossRef]
6. Suk, H.I.; Lee, S.W.; Shen, D. Deep sparse multi-task learning for feature selection in Alzheimer's disease diagnosis. *Brain Struct. Funct.* **2016**, *221*, 2569–2587. [CrossRef]
7. Suk, H.I.; Lee, S.W.; Shen, D. Latent feature representation with stacked auto-encoder for AD/MCI diagnosis. *Brain Struct. Funct.* **2013**, *220*, 841–859. [CrossRef]
8. Wee, C.-Y.; Liu, C.; Lee, A.; Poh, J.S.; Ji, H.; Qiu, A. Cortical graph neural network for AD and MCI diagnosis and transfer learning across populations. *NeuroImage Clin.* **2019**, *23*, 101929. [CrossRef]
9. Agah, A. *Medical Applications of Artificial Intelligence*; Taylor and Francis: Boca Raton, FL, USA, 2013. [CrossRef]
10. Xu, J.; Lei, X.; Hang, R.L.; Gilmore, H.; Wu, J.; Tang, J.; MAdabhushi, A. Stacked Sparse Autoencoder (SSAE) for Nuclei Detection on Breast Cancer Histopathology Images. *IEEE Trans. Med. Imaging* **2016**, *35*, 19–30. [CrossRef]
11. Fakoor, R.; Ladhak, F.; Nazi, A.; Huber, M. Using Deep Learning to Enhance Cancer Diagnosis and Classification. Available online: https://www.researchgate.net/publication/281857285_Using_deep_learning_to_enhance_cancer_diagnosis_and_classification (accessed on 1 January 2023).
12. Zuluaga-Gomez, J.; Al-Masry, Z.; Benaggoune, K.; Meraghni, S.; Zerhouni, N. A CNN-based methodology for breast cancer diagnosis using thermal images. *Comput. Methods Biomech. Biomed. Eng. Imaging Vis.* **2021**, *9*, 131–145. [CrossRef]
13. Shamshirband, S.; Fathi, M.; Dehzangi, A.; Chronopoulos, A.T.; Alinejad-Rokny, H. A review on deep learning approaches in healthcare systems: Taxonomies, challenges, and open issues. *J. Biomed. Inform.* **2021**, *113*, 103627. [CrossRef] [PubMed]
14. Liu, G.P.; Yan, J.J.; Wang, Y.Q.; Zheng, W.; Zhong, T.; Lu, X.; Qian, P. Deep Learning Based Syndrome Diagnosis of Chronic Gastritis. *Comput. Math. Methods Med.* **2014**, *2014*, 938350. [CrossRef] [PubMed]
15. Ikenoyama, Y.; Hirasawa, T.; Ishioka, M.; Namikawa, K.; Yoshimizu, S.; Horiuchi, Y.; Ishiyama, A.; Yoshio, T.; Tsuchida, T.; Takeuchi, Y.; et al. Detecting early gastric cancer: Comparison between the diagnostic ability of convolutional neural networks and endoscopists. *Dig. Endosc.* **2021**, *33*, 141–150. [CrossRef] [PubMed]
16. Helbich, M.; Yao, Y.; Liu, Y.; Zhang, J.; Liu, P.; Wang, R. Using deep learning to examine street view green and blue spaces and their associations with geriatric depression in Beijing, China. *Environ. Int.* **2019**, *126*, 107–117. [CrossRef]
17. Bertini, F.; Allevi, D.; Lutero, G.; Calzà, L.; Montesi, D. An automatic Alzheimer's disease classifier based on spontaneous spoken English. *Comput. Speech Lang.* **2022**, *72*, 101298. [CrossRef]
18. Kumar, D.; Wong, A.; Clausi, D.A. Lung Nodule Classification Using Deep Features in CT Images. *Wirel. Pers. Commun.* **2015**, *116*, 655–690.
19. Ma, J.; Sheridan, R.P.; Liaw, A.; Dahl, G.E.; Svetnik, V. Deep Neural Nets as a Method for Quantitative Structure–Activity Relationships. *J. Chem. Inf. Model.* **2015**, *55*, 263–274. [CrossRef]
20. Bhinder, B.; Gilvary, C.; Madhukar, N.S.; Elemento, O. Artificial Intelligence in Cancer Research and Precision Medicine. *Cancer Discov.* **2021**, *11*, 900–915. [CrossRef]
21. Prasoon, A.; Petersen, K.; Igel, C.; Lauze, F.; Dam, E.; Nielsen, M. Deep Feature Learning for Knee Cartilage Segmentation Using a Triplanar Convolutional Neural Network. *Med. Image Comput. Comput. Assist. Interv.* **2013**, *16*, 246–253.
22. Song, Y.; Zhang, L.; Chen, S.; Ni, D.; Li, B.; Zhou, Y.; Lei, B.; Wang, T. A deep learning based framework for accurate segmentation of cervical cytoplasm and nuclei. In Proceedings of the 36th Annual International Conference of the IEEE Engineering in Medicine and Biology Society, Chocago, IL, USA, 26–30 August 2014; p. 2903.
23. Majumdar, A. Real-Time Dynamic MRI Reconstruction using Stacked Denoising Autoencoder. Available online: https://arxiv.org/ftp/arxiv/papers/1503/1503.06383.pdf (accessed on 1 January 2023).
24. Chaudhari, A.S.; Sandino, C.M.; Cole, E.K.; Larson, D.B.; Gold, G.E.; Vasanawala, S.S.; Lungren, M.; Hargreaves, B.A.; Longlotz, C.P. Prospective deployment of deep learning in MRI: A framework for important considerations, challenges, and recommendations for best practices. *J. Magn. Reson. Imaging* **2021**, *54*, 357–371. [CrossRef]
25. Zhang, W.; Li, R.; Deng, H.; Wang, L.; Lin, W.; Ji, S.; Shen, D. Deep convolutional neural networks for multi-modality isointense infant brain image segmentation. *Neuroimage* **2015**, *108*, 214–224. [CrossRef] [PubMed]
26. Zeng, T.; Li, R.; Mukkamala, R.; Ye, J.; Ji, S. Deep convolutional neural networks for annotating gene expression patterns in the mouse brain. *BMC Bioinform.* **2015**, *16*, 147. [CrossRef] [PubMed]
27. Wang, D.; Shang, Y. Modeling Physiological Data with Deep Belief Networks. *Int. J. Inf. Educ. Technol.* **2013**, *3*, 505–511. [PubMed]
28. Spencer, M.; Eickholt, J.; Cheng, J. A Deep Learning Network Approach to ab initio Protein Secondary Structure Prediction. *IEEE/ACM Trans. Comput. Biol. Bioinform.* **2015**, *12*, 103–112. [CrossRef] [PubMed]
29. Heffernan, R.; Paliwal, K.; Lyons, J.; Dehzangi, A.; Sharma, A.; Wang, J.; Sattar, A.; Yang, Y.; Zhou, Y. Improving prediction of secondary structure, local backbone angles and solvent accessible surface area of proteins by iterative deep learning. *Sci. Rep.* **2015**, *5*, 11476. [CrossRef]
30. Zhou, J.; Troyanskaya, O.G. Deep Supervised and Convolutional Generative Stochastic Network for Protein Secondary Structure Prediction. *Comput. Sci.* **2014**, *2014*, 745–753.

31. Lena, P.D.; Nagata, K.; Baldi, P. Deep architectures for protein contact map prediction. *Bioinformatics* **2012**, *28*, 2449–2457. [CrossRef]
32. Yuvaraj, N.; Srihari, K.; Chandragandhi, S.; Raja, R.A.; Dhiman, G.; Kaur, A. Analysis of protein-ligand interactions of SARS-Cov-2 against selective drug using deep neural networks. *Big Data Min. Anal.* **2021**, *4*, 76–83. [CrossRef]
33. Quang, D.; Chen, Y.; Xie, X. DANN: A deep learning approach for annotating the pathogenicity of genetic variants. *Bioinformatics* **2014**, *31*, 761–763. [CrossRef]
34. Li, J.; Pu, Y.; Tang, J.; Zou, Q.; Guo, F. DeepATT: A hybrid category attention neural network for identifying functional effects of DNA sequences. *Briefings Bioinform.* **2021**, *22*, 159. [CrossRef]
35. Leung, M.K.K.; Xiong, H.Y.; Lee, L.J.; Frey, B.J. Deep learning of the tissue-regulated splicing code. *Bioinformatics* **2014**, *30*, 121–129. [CrossRef]
36. Xiong, H.Y.; Alipanahi, B.; Lee, L.J.; Merico, D.; Yuen, R.; Hua, Y.; Hughes, T.; Morris, Q.; Barash, Y.; Jojic, N.; et al. The human splicing code reveals new insights into the genetic determinants of disease. *Science* **2015**, *347*, 1254806. [CrossRef] [PubMed]
37. Bhattacharjee, S.; Ghosh, A.; Saha, B.; Saha, S. Machine Learning in Genomics. In *Machine Learning and Systems Biology in Genomics and Health*; Springer: Berlin/Heidelberg, Germany, 2022; pp. 69–90.
38. Almagro Armenteros, J.J.; Sønderby, C.K.; Sønderby, S.K.; Nielsen, H.; Winther, O. DeepLoc: Prediction of protein subcellular localization using deep learning. *Bioinformatics* **2017**, *33*, 3387–3395. [CrossRef] [PubMed]
39. Zou, Z.; Tian, S.; Gao, X.; Li, Y. mlDEEPre: Multi-functional enzyme function prediction with hierarchical multi-label deep learning. *Front. Genet.* **2019**, *9*, 714. [CrossRef] [PubMed]
40. Kulmanov, M.; Khan, M.A.; Hoehndorf, R. DeepGO: Predicting protein functions from sequence and interactions using a deep ontology-aware classifier. *Bioinformatics* **2018**, *34*, 660–668. [CrossRef] [PubMed]
41. Ma, J.; Yu, M.K.; Fong, S.; Ono, K.; Sage, E.; Demchak, B.; Sharan, R.; Ideker, T. Using deep learning to model the hierarchical structure and function of a cell. *Nat. Methods* **2018**, *15*, 290–298. [CrossRef]
42. Zitnik, M.; Agrawal, M.; Leskovec, J. Modeling polypharmacy side effects with graph convolutional networks. *Bioinformatics* **2018**, *34*, i457–i466. [CrossRef]
43. Zong, N.; Kim, H.; Ngo, V.; Harismendy, O. Deep mining heterogeneous networks of biomedical linked data to predict novel drug–target associations. *Bioinformatics* **2017**, *33*, 2337–2344. [CrossRef]
44. Wan, F.; Hong, L.; Xiao, A.; Jiang, T.; Zeng, J. NeoDTI: Neural integration of neighbor information from a heterogeneous network for discovering new drug–target interactions. *Bioinformatics* **2019**, *35*, 104–111. [CrossRef]
45. Rumelhart, D.E.; Hinton, G.E.; Williams, R.J. Learning Internal Representations by Error Propagation. *Read. Cogn. Sci.* **1988**, *323*, 399–421.
46. Hinton, G.E.; Salakhutdinov, R.R. Reducing the Dimensionality of Data with Neural Networks. *Science* **2006**, *313*, 504–507. [CrossRef] [PubMed]
47. Tibshirani, R.; Walther, G.; Hastie, T. Estimating the number of clusters in a data set via the gap statistic. *J. R. Stat. Soc. Ser. B* **2001**, *63*, 411–423. [CrossRef]