A hybrid classifier based on nonlinear-PCA and deep belief networks with applications in dysphagia diagnosis

Chong Su\textsuperscript{a*}, Yue Gao\textsuperscript{a}, Yuxiao Xie\textsuperscript{b*}, Yong Xue\textsuperscript{b}, Lijun Ge\textsuperscript{b} and Hongguang Li\textsuperscript{a}

\textsuperscript{a}School of Information Science and Technology, Beijing University of Chemical Technology, Beijing, China; \textsuperscript{b}Department of Rehabilitation, China-Japanese Friendship Hospital, Beijing, China

ABSTRACT

Traditional dysphagia prescreening diagnostic methods require doctors specialists to give patients a total score based on a water swallow test scale. This method is limited by the high dimensionality of the diagnostic elements in the water swallow test scale with heavy workload (Towards each patient, the scale requires the doctors give score for 18 diagnostic elements respectively) as well as the difficulties of extracting and using the diagnostic scale data’s non-linear features and hidden expertise information (Even with the scale scores, specific diagnostic conclusions are still given by expert doctors under the expertise). In this paper, a hybrid classifier model based on Nonlinear-Principal Component Analysis (NPCA) and Deep Belief Networks (DBN) is proposed in order to effectively extract the diagnostic scale data’s nonlinear features and hidden information and to provide the key scale elements’ locating methods towards the diagnostic results (The key scale elements that affect different diagnostic conclusions are given to improve the efficiency and pertinence of diagnosis and reduce the workload of diagnosis). We demonstrate the effectiveness of the proposed method using the frame of ‘information entropy theory’. Real dysphagia diagnosis examples from the China-Japanese Friendship Hospital are used to demonstrate applications of the proposed methods. The examples show satisfactory results compared to the traditional classifier.

KEYWORDS

Dysphagia diagnosis; nonlinear PCA (NPCA); deep belief networks (DBN); entropy

1. Introduction

Dysphagia has been found to be the most common sequela of stroke [1,2]. Video fluoroscopy and swallow examination (VFSS) evaluating swallowing function, fiberoptic endoscopic evaluation of swallowing (FEES) are the ‘Gold standard’ for swallowing function test. However, the ‘gold standard’ test is often difficult to complete because the patient is uncooperative or in a coma. Typically, doctors give prescreening diagnostic results based on patients’ water swallow test scale scores, which consist of examinations of the feeding process and dysfunction assessments including laryngeal functions (sound occasion, sounding quality, tune), respiratory functions (breathing times, the ratio of fast to slow breathing), zygomatic reflexes (normal and abnormal zygomatic movements), facial examinations (if the patient’s face is symmetrical or paralyzed), and tongue function (degree of flexibility) [3,4]. However, there is an increasing need to improve the accuracy of diagnostic tests, due to challenges including difficulties in evaluating dysfunction (for example, doctors cannot execute the sophageal television lens examination for patients with severe stroke) [5]. Figure 1 shows the anatomical structures of dysphagia, and Table 1 shows the 18 diagnosis items of the water swallow test scale proposed by DG Smithard et al. at the South University of Manchester Hospital [2,6–9].

Water swallow test scale diagnostic elements reveal multiple dimensions and nonlinear features; therefore, it is difficult to quantify and utilize qualitative information, like physician expertise, which is an integral aspect of diagnosis [10,11]. Research on intelligent diagnostics for dysphagia has not been widely published; however, artificial intelligence used in other areas of medicine can be applied to the study of intelligent dysphagia diagnosis. Artificial intelligence methods have shown good applicability to pattern recognition issues with uncertainty models [12–14]
and complex data [15,16]. Presently, a number of intelligent approaches including expert systems, artificial neural networks, and genetic algorithms are used in medical diagnosing and treatment [17]. For example, a fuzzy method was built for tremor assessment, prediction, and rehabilitation [18], a reasoning method was employed to construct medical expert systems [19], and a new heart disease diagnosis expert system has been established [20]. Many reviews have studied [20–24] dysphagia diagnostic classification networks. For example, Oral and pharyngeal phases of the swallowing mechanism often leads to aspiration. Two sets of networks were developed, one for the oral called the oral classifier, and the other one for the pharyngeal phase [20]. Prabhu et al. [21] studied the acceleration signal during swallowing was filtered and segmented. The parameters extracted from the signal were used to develop and train two sets of Neural Network models. The first Neural Network model was developed an ANN model to differentiate between the acceleration signals of normal, dysphagic and coughing, and then, designed another ANN model to differentiate between acceleration during swallowing in normal, mild dysphagic, moderate dysphagic and severe dysphagic subjects. Lee et al. [22] studied 100 children with neurologically-based dysphagia using a single-axis accelerometer. And extracted five potentially discriminatory mathematical features from the accelerometry signals based on radial basis function classifiers. Cola et al. [24] investigated the influence of stimuli, such as taste (i.e. natural, cold, sour, and sour-cold) and temperature based on the ANN (Artificial Neural Network) methods. However, The above-mentioned methods are limited to shallow layer learning, which does not take full advantage of the diagnostic experiences of experts, and traditional dysphagia diagnostic classifiers are not able to locate key water swallow test scale item elements.

Common classification methods such as clustering, support vector machine (SVM), and neural networks are shallow. The traditional clustering methods for high dimensional data have more difficulty finding the cluster centers than low dimensional data clustering, because the data distribution is more sparse, and the distance between each of the data points might be quite large [25]. SVM is a supervised learning algorithm with certain advantages in terms of small sample classification; however, this method is very reliant on the sample data scale and is very complex in terms of algorithm classifying. Additionally, the SVM’s kernel function selection determines the classification accuracy and convergence speed [26]. Due to the lack of pre-training, the neural network cannot mine implicit information [27]. The deep learning technique consists of two phases: the unsupervised pre-training and the supervised back-propagation, which are beneficial for mining the latent relationship [28]. Deep learning methods involve three main models: automatic encoders, convolution neural networks, and deep belief networks [29]. Recently, deep learning techniques have attracted increasing attention in the field of machine learning, inspiring successful application in multiple fields. The deep belief network differs from other deep learning models. It is hierarchically built by stacking a series of restricted Boltzmann machines (RBMs). The supervised back-propagation algorithm can be used to obtain weights after the unsupervised pre-training phase [30]. Li applied a variety of well-trained deep belief network (DBN) models in order to predict the risk of bone disease development [31]. Shang proposed a kind of soft sensor based on deep belief networks in a crude distillation unit (CDU) [32]. In addition, deep learning methods are most widely used in the field of voice and image recognition with the improved DBN towards the different objectives. For example, a CTC-DBN (connectionist temporal classification-DBN) [33] was proposed to recognize the key speech information on the telephone where the new DBN is applicable to the real-time dynamic classification. And the CRF-DBN (Conditional random field-DBN) [34] is designed to deal with the information recognition of the context. Through literature researches, we found that all the improved DBNs are designed for the specific objective models with out applications in the diagnosis of dysphiga. Therefore, it is conceivable that the deep belief network is suitable for constructing a model of dysphagia diagnosis classifications.
Additionally, given current dysphagia diagnostic classifications, it remains difficult to locate key scale item elements in the presence of high dimensional and nonlinear features associated with the water swallow test scale scores. As the most popular approaches for obtaining variable correlations, the variance filter and principal component analysis (PCA) are able to reduce data dimensions [35,36]. However, this demonstrates the limitations of obtaining nonlinear data information. In contrast, nonlinear PCA approaches have been successfully applied in the field of nonlinear feature extractions. Li used the input-training network to build a nonlinear PCA framework [37]. Niu combined the Radial Basis Function (RBF) with PCA to reduce modeling errors [38]. Subsequently, it was found that nonlinear PCA plays an important role in fault detection for process industries [39–42]. It is conceivable that nonlinear PCA methods are eligible for both dimension reductions and nonlinear feature extractions in order to address dysphagia diagnostic data, which could further contribute to the location of key water swallow test scale item elements associated with diagnostic results.

In this work, we first built a diagnostic data dimension reduction structure that combines nonlinear PCA (NPCA) and IT-NN (Input training neural network). We then create a novel classifier based on deep brief networks to improve the effectiveness of dysphagia diagnostic results. We can use NPCA to locate the key water swallow test scale item elements associated with diagnostic results.

The remainder of this article is organized as follows. Section 2 presents the modeling steps for the novel hybrid classifier based on NPCA and DBN. In Section 3, we introduce the dysphagia diagnosis method based on the classifier along with an approach to locate the key item elements in the water swallow test scale. Section 4 provides a dysphagia diagnosis case study to demonstrate the effectiveness of the proposed approaches. Section 5 provides a conclusion and suggests future areas of research.

2. The hybrid classifier

Because the CNN (Convolutional Neural Network) is rarely used to build the model of process data, DBN is more effective to recognize the information in the process data like medical diagnosis problem without strong dynamic and real-time demand there is no dynamic demand strong.

Conventional DBN methods demand large amounts of training data in order to achieve excellent modeling performances. However, in the medical diagnosis field, it is difficult to obtain a large number of diagnostic cases. In the following section, we detail the modeling methods of a novel hybrid classifier based on NPCA and DBN. First, NPCA is used to reduce data dimensions and extract the nonlinear features of the data. Then, the reduced-dimensional data are used to generate the inputs and outputs of the DBN model. Lastly, we introduce a method to locate the key input variables in terms of influencing the model output.

2.1. Data dimension reductions

Linear PCA identifies linear correlations between process variables, whereas the objective of nonlinear PCA is to extract both the latent linear and nonlinear correlations, as described below:

\[
X' = F(T_N) + E
\]

where, \(X\) is the original sampling data, \(T_N = [t_1, t_2, ..., t_k]\) is the matrix of nonlinear principal component scores, \(E\) is the matrix of residuals, \(X'\) is the estimated value of \(X\), and \(F(t)\) is the nonlinear function expressed by an input-training neural network (IT-NN) [37]. Here, IT-NN is used to reconstruct low-dimensional diagnostic data. Figure 2 shows the structure of the data dimension reduction.

Specifically, to train the network, the following goal function must be optimized:

\[
J = \frac{1}{2} \sum_{i=1}^{m} \sum_{j=1}^{n} (X_{ij} - X'_{ij})^2
\]

where, gradient optimization algorithms can be used to achieve the minimum value of \(J\).

2.2. Classifier modeling

DBN techniques [39] are employed to construct the hybrid classifier model, whose inputs are specified by the principal component score vector \(T_N\).

![Figure 2. Structure of data dimension reductions.](image)
The modeling procedure to establish the hybrid classifier is described in the following section.

**1) Basics of RBM**

Based on traditional Boltzmann machines, the restricted Boltzmann machine (RBM) is extended with stratified limits, as well as a binary chart that consists of visual layers and hidden layers. Outputs of the hidden layer correspond to the internal features of data, which can maximize the probability of a specific kind of distribution.

First, we introduce basic energy-based models, in which each configuration of variables is assigned to a scalar value of energy. Let \( v \) denote the input vector of the RBM (visible layers), then the probability of distribution through an energy function can be defined as:

\[
P(v) = \frac{\exp(-\text{Energy}(v))}{Z}
\]

where, \( Z = \sum \exp(\text{Energy}(v)) \) is the normalizing factor. A lower energy is preferred to maximize the probability \( P(v) \). By introducing a vector of the hidden layer \( h \), the RBM framework is formulated as follows [42]:

\[
P(v, h) = \frac{\exp(-E(v, h))}{Z} = \frac{1}{Z} \prod_j e^{w_{ij}v_i} \prod_i e^{b_i} \prod_j e^{c_jh_j}
\]

The RBM energy function can be simplified as a second-order polynomial [42]

\[
\text{Energy}(v, h) = -b^Tv - c^Th - h^TWv
\]

where \( v \) and \( h \) are both vectors of binary values (0 or 1) and \( W, b, \) and \( c \) are parameters of the energy function. Because of this, they are referred to as binary RBMs, which are most frequently used in practical applications.

However, binary RBMs can only process discrete inputs. In order to tackle continuous-valued inputs, binary RBMs can be extended to Gaussian RBMs as follows [40]:

\[
\text{Energy}(v,h) = \sum_i \frac{(v_i - \alpha_i)^2}{2\sigma^2} - b^Tv - h^TWv
\]

where \( \alpha_i \) and \( \sigma_i \) are the mean and the standard deviation of the Gaussian distribution for the visible unit \( i \). Here, the input layer \( v \) is continuous-valued and the hidden layer \( h \) is binary. In Eq. (4), \( \theta = \{W, b, c\} \), \( \alpha_i \), and \( i \) are parameters to be estimated in the training procedure.

Therefore, it is reasonable to set the training objective of RBM as maximizing \( P(v) \) which is the probability of the model simply based on the training input data. The optimization is typically performed by a state-of-the-art gradient descent. Using Eq. (2) and the total probability formula, the gradient of the log-likelihood function at a single data point \( v \) is calculated as follows:

\[
\frac{\partial \log P(v)}{\partial \theta} = \frac{\partial \log \sum P(v, h)}{\partial \theta}
\]

\[
\sum e^{-\text{energy}(v, h)}(\partial[-\text{energy}(v, h)/\partial \theta])
\]

\[
\frac{\partial \text{energy}(v, h)}{\partial \theta} = \sum \sum e^{-\text{energy}(v, h)}(\partial[-\text{energy}(v, h)/\partial \theta])
\]

\[
\frac{\partial P(v)}{\partial \theta} = \sum P(h|v) \frac{\partial [-\text{energy}(v, h)]}{\partial \theta}
\]

\[
- \sum \sum P(v, h) \frac{\partial [-\text{energy}(v, h)]}{\partial \theta}
\]

where, \( \theta = \{W, b, c\} \). Two terms in Eq. (6) are interpreted such that the positive term represents the conditional expectation of \( \partial[-\text{Energy}(v, h)/\partial \theta] \) the given visible units \( v \), and the negative term represents the expectation of \( \partial[-\text{Energy}(v, h)]/\partial \theta \) the joint distribution \( P(v, h) \) with parameter \( \theta \). Note that \( \partial[-\text{Energy}(v, h)]/\partial \theta \) is easy to compute, which makes the positive term available with \( P(h|v) \). However, for RBMs with numerous units, computations of the negative term through sampling become very difficult [42].

**2) The improved DBN structures**

As mentioned in Section 2, we can stack a series of RBMs to construct a DBN. The hidden layer in an RBM can suggest underlying features behind the input data. The hidden layers of high-level RBMs represent high-level features because high-level RBMs are established based on low-level features from the hidden layer of low-level RBMs. Therefore, DBN can be regarded as a latent variable model with the inner layer representing low-level features and output amounting to high-level features. And the nodes are full linked. The framework of the novel DBN is shown in Figure 3 [42].

The traditional DBN [32] proposed by Hinton consists of restricted Boltzmann machines, where each restricted Boltzmann machine is designed as a node of the network connecting between layers of nodes, as well, the nodes in the layer is not connected [32]. From Figure 3, we can see that the novel DBN is different from the traditional DBN.
The DBN output is characterized as follows:

\[
Y = S(T_N) + E
\]  

(8)

where, \(S(\cdot)\) is a nonlinear function to be approximated by the deep learning neural network, \(E\) refers to the error.

(3) Location of contributions

For linear PCA, the \(Q\) and \(T_2\) statistics must be modified to calculate the contributions of the original process variables. While in nonlinear PCA, the \(T_2\) statistic is changed due to the nonlinear information. Because the first-order partial derivative of the output results in the change of variables, we can obtain the following formula:

\[
K = \frac{\partial Y}{\partial X_i}
\]  

(9)

where, \(Y\) indicates the output of the DBN, \(X_i\) indicates the input variables, and \(i\) is the number of variables. The relationship between \(Y\) and \(X\) is represented by Eqs. (8) and (1). Therefore, we obtain the following relationship:

\[
K_{|_{x_i}} = \frac{\partial Y}{\partial X_i} = \frac{\partial Y}{\partial T_N} \frac{\partial T_N}{\partial X_i} = F'(X_i) K(10)
\]

where, \(\frac{\partial Y}{\partial X_i}\) is the first-order partial derivative function between output \(Y\) and nonlinear principle components \(T_N\), and the relationship between \(Y\) and \(T_N\) is shown in Eq. (8) in the classification stage, \(\frac{\partial Y}{\partial X_i}\) is the first-order partial derivative function between \(T_N\) and \(X_i\), and the relationship is represented in Eq. (1) in the dimension reduction stage, the nonlinear functions \(H(\cdot)\) and \(F(\cdot)\) are modified by the neural networks.

Thereafter, the contribution \(K_{|_{x_i}}\) is adopted to represent the contribution in order to locate the key input variables that influence the DBN’s outputs.

3. Complexity analysis of improved DBN

A complexity analysis is shown below, in order to demonstrate the advantages of the improved DBN in terms of network structure and algorithm complexity:

(1) The structural complexity of the improved DBN

In 1958, Shannon proposed the concept of information entropy, which describes the uncertainty of information sources. When researchers seek to improve the generalization ability, they use an approach to control the ‘effective complexity’ of the network by regulariza-

\[
H = -\sum_{j=1}^{Q_1} d_j \log d_j + \frac{1}{P} \sum_{i=1}^{P} \sum_{j=1}^{Q_1} d_j \log d_j
\]  

(12)

where, \(P\) is the number of input samples, and \(Q\) is the nodes’ number of hidden layers, and \(d_j\) corresponds to the normalized output from the \(i\)th samples to the first \(j\)th hidden units.

**Theorem 1:** Assume that there are two DBNs (net1 and net2) that have the same information feature. If the two DBNs have the same ‘effective complexity’ \((H_{\text{net1}} = H_{\text{net2}})\) when their network input layer nodes’ number meet \(P_{\text{net1}} < P_{\text{net2}}\), then the total hidden layer nodes’ number meet \(Q_{\text{net1}} < Q_{\text{net2}}\).

**Proof:**

Assume the mutual information entropy function of the first DBN network (net1) is minimized, and the first RBMs’ layer in net1 fully reflects the mutual information of the input layer and the hidden layer. Based on the principle of information entropy, we can obtain [24]:

\[
-\sum_{j=1}^{Q_1} d_j \log d_j = -\sum_{j=1}^{Q_1} \frac{1}{Q_1} \log \frac{1}{Q_1} = \log Q_1
\]  

(13)

where, \(Q_1\) represents the number of nodes in the first layer of the hidden layer. Based on equations (12) and (13), we have:

\[
H = \log Q_1 + \frac{1}{P} \sum_{i=1}^{P} \sum_{j=1}^{Q_1} d_j \log d_j
\]  

(14)

According to the principle of DBN, the improved DBN (net2) proposed in this paper can make each RBM completely reconstruct the input variables,
therefore, the net2 should minimize all of the mutual information entropy. The improved mutual information entropy $H'$ is described as:

$$H' = \log Q'_1 + \frac{1}{P} \sum_{l=1}^{P} \sum_{j=1}^{Q'_l} d_{lj} \log d_{lj}$$ (15)

Further, net2’s NPCA has completely extracted the sample features; therefore, net2’s first layer of RBMs can map the hidden layer to the original sample information. Then, we have:

$$\sum_{l=1}^{P} \sum_{j=1}^{Q'_l} d_{lj} \log d_{lj} = \sum_{l=1}^{P} \sum_{j=1}^{Q'_l} d_{lj} \log d_{lj}$$ (16)

Thereafter, if we use two DBNs (net1 and net2) of the same network structure to reconstruct the same sample data, the ‘effective complexity’ should be equal ($H = H'$). If $P' \leq P$, then $Q'_1 \leq Q_1$. Similarly, the number of RBM nodes in the subsequent hidden layer also follows this rule. We have:

$$Q_{\text{net2}} = Q'_1 + Q'_2 + \cdots + Q'_n \leq Q_{\text{net1}} = Q_1 + Q_2 + \cdots + Q_n$$

($n$ is the total number of hidden layers of the network).

Therefore, the total number of the improved DBN’s node meet $S_{\text{net1}} < S_{\text{net2}}$.

(2) The time and space complexity analysis of the improved DBN

As a computer science technique, analysis of an algorithm’s complexity typically assesses the costs associated with time and space complexity [44]. To address this, we briefly present the complexity analysis of the effective learning algorithms as follows.

**Theorem 2:** Assume there is a DBN network (net1) containing $n$ hidden layers where the total node number of the hidden layers is $[h_1, h_2 \ldots h_n]$. Therefore, there must be a class of DBN (net2) under the nonlinear feature extraction mechanism based on the ITNN neural network (assuming that the nonlinear feature extraction mechanism can effectively ensure the equivalent of samples’ mutual information entropy). When the total node number of the improved DBN (net2) hidden layers is reduced to $[h'_1, h'_2 \ldots h'_n]$, the total node number of ITNN neural network’s hidden layer nodes is $h_0$, which satisfies $h_0 < \beta$. We have: the time complexity of the improved DBN is less than the original DBN, which is described as:

$$\beta = \prod h_n - \prod h'_n$$

**Proof:**

1. The pseudo-code of the traditional DBN algorithm is as follows:

   ```
   for (1: h_1 (the node number of the original DBN’s first hidden layer))
   for (1: h_2 (the node number of the original DBN’s second hidden layer))
   ... for (1: h_n (the node number of the original DBN’s n-th hidden layer))
   Calculate the objective function of the network and estimate if the objective function meets the requirements;
   end
   ...
   end
   ```

   Therefore, the time complexity of DBN is $O(DBN) = \prod h_n$.

2. The improved NPCA-DBN algorithm pseudo-code is as follows:

   ```
   for (1: h_0 (the node number of NPCA’s hidden layer))
   Calculate the objective function of the network and estimate if the objective function meets the requirements;
   end
   for (1: h'_1 (the node number of improved DBN’s first hidden layer))
   for (1: h'_2 (the node number of improved DBN’s second hidden layer))
   ... for (1: h'_n (the node number of improved DBN’s n-th hidden layer))
   Calculate the objective function of the network and estimate if the objective function meets the requirements;
   end
   ...
   end
   ```

   Therefore, the time complexity of the improved DBN is $O(NPCA - DBN) = h_0 + \prod h'_n$.

   Additionally, because the improved DBN reduces the number of hidden layer nodes in the original DBN, it can effectively reduce space complexity.
4. Dysphagia diagnosis based on the improved DBN

Based on the proposed hybrid classifier, we establish the dysphagia diagnostic method. Then, according to the dysphagia diagnostic results, the key water swallow test scale item elements are determined, which can help make the treatment process more effective.

The digital description method of water swallow test scale elements is carried out before the dysphagia diagnostic approaches are presented based on the proposed hybrid classifier model. Then a detailed description is given of the key elements of the water swallow test scale locating process based diagnostic results. The flow chart of the dysphagia diagnostic process is shown in Figure 4.

4.1. Digitalization of the water swallow test scale item elements

The digitalizing description of 15 diagnostic results (mouth period (I), tongue period (II), throat period (III), pharynx period (IV), and syndromes with a combination of the four conditions above [42]) is shown in Table 2.

4.2. Diagnostic classifications

Unsupervised learning, based on the proposed classifier, is used to learn the diagnostic data data features. Considering 150 historic diagnostic data points as \( X' \), the principal component scores vector \( T_N \) (Eq. (1)) is specified as the input of the DBN model, which is trained to obtain the diagnostic classification model.

According to Table 2, if the DBN's output is 0.8, then we can conclude that the patient is suffering from the a mix of Mouth, Tongue, and Throat periods (I/II/IV). If the DBN's output is 0.139, the patient may be experiencing a Tongue period (as the output is close to II (0.133)).

4.3. Key scale element locations

Locating the key water swallow test scale elements is important for ensuring accurate treatment. According to Eq. (10), we can obtain the key water swallow test scale elements' contribution to the diagnostic results, which can help doctors formulate treatment plans.

Consider that the historic dysphagia diagnosis cases \( X = [X_1, X_2, X_3, ..., X_{18}] \) are specified as the inputs of the DBN. The DBN output \( Y \) corresponds to the classification result. The contribution of the key water swallow test scale element is termed as \( K_{|x_i} \) in Eq. (10). In this paper, we define the contribution score limits as 80%. The input variable whose contribution is over the

![Figure 4. The flow chart of the dysphagia diagnostic process.](image)

Details of the key steps for dysphagia diagnosis are described as follows.

| Table 1. Diagnostic items for water swallow test scales. |
|---------------------------------------------------------|
| Water swallow test scales                               |
| 1 The level of consciousness                            |
| 2 The control of head and torso                         |
| 3 The pattern of breathing                              |
| 4 The lip closure                                        |
| 5 The movement of the soft palate                       |
| 6 The judgment of laryngeal function                    |
| 7 The pharyngeal reflex                                 |
| 8 The independent cough                                 |
| 9 Outflow in the first feed water stage (5 ml) 3 times   |
| 10 Laryngeal movement in the first feed water stage     |
| 11 Repeat swallowing in the first feed water stage      |
| 12 Cough in the first feed water stage                  |
| 13 Stridor in the first feed water stage                |
| 14 Outflow in the second feed water stage (60 ml)       |
| 15 Cough in the second feed water stage                 |
| 16 Stridor in the second feed water stage               |
| 17 The judgment of laryngeal function in the second feed water stage |
| 18 Regurgitation phenomena                              |
limit is recognized as the key element of the diagnosis.

5. Case studies

In this section, 180 cases collected by the Department of Physical Rehabilitation in China-Japanese Friendship Hospital were employed to verify the proposed methods. 150 of the cases were used to train the dysphagia diagnostic classifier, while the other 30 cases were used as test data. The boxplots of the DBN input variables and the outputs that were created for the training and generalizing case data are listed in Figure 4. The training data and the testing data share the same general characteristics, which enhances the availability of the data. We summarize the key water swallow test scale diagnostic item elements’ contributions for all types of dysphagia diagnostic results. Finally, the validation of the diagnostic model and key scale elements’ location results are verified by experts using statistical approaches.

The abscissa in the Figure 5 represents the boxplots’ statistical distribution of eighteen diagnostic scale items and the diagnostic results in the training group (150 patients) and the testing group (30 patients) [45]. In addition, the ordinates corresponds the normalized results of the eighteen diagnostic scale items and the diagnostic results. We can see that the distribution of the testing group data is approximately the same as that of the training samples. Therefore, it justifies that the sample used for survey is representative.

The eighteen patients’ water swallow test scale item elements can be converted to fall between 0 and 1, as seen in Table 3.

5.1. Dysphagia diagnosis

The proposed NPCA-DBN hybrid classifier was used to diagnose the dysphagia cases. The NPCA based data dimension reducer converted the 18 original water swallow test scale diagnostic item elements into 10 principal components before DBN is trained to obtain the dysphagia diagnostic model. Figures 5 and 6 show comparisons in terms of training performance and ability to generalize between the proposed hybrid classifier and the traditional DBN model.

| The digitalized diagnostic results |  |
|-----------------------------------|---|
| I (0.067)                         | II (0.133)                     |
| I/II (0.333)                      | III (0.200)                    |
| I/II/III (0.733)                  | IV (0.267)                     |
| I/I/IV (0.800)                    | I/III (0.467)                  |
| I/II/IV/II (0.933)               | II/III (0.533)                 |
| I/II/III/IV (1.000)              | II/IV (0.600)                  |
|                                  | III/IV (0.667)                 |

Table 2. The digitalized diagnostic results.
Table 3. Digital diagnostic elements in water swallow test scales.

| Num | Items                        | Grading | Normalized diagnostic description                                      |
|-----|------------------------------|---------|------------------------------------------------------------------------|
| 1   | The level of consciousness   | 1/2/3/4 | Awake = 0.25, drowsy but can awaken = 0.5, response but no open and language = 0.75, response to pain = 1 |
| 2   | Control of head and torso    | 1/2/3/4 | Can sit normally = 0.25, Cannot sit = 0.5, Can only control the head = 0.75, Cannot control the head = 1 |
| 3   | Breathing pattern            | 1/2     | Normal = 0.5, abnormal = 1                                              |
| 4   | Lip closure                  | 1/2     | Normal = 0.5, abnormal = 1                                              |
| 5   | Movement of the soft palate  | 1/2/3   | Normal = 0.33, Weak = 0.66, Not present = 1                              |
| 6   | Judgment of laryngeal function| 1/2    | Present = 0.5, Not present = 1                                           |
| 7   | Pharyngeal reflex            | 1/2/3   | Normal = 0.33, Weak = 0.66, Not present = 1                              |
| 8   | Independent cough            | 1/2     | No or once = 0.5, More than once = 1                                    |
| 9   | Outflow in the first feed water stage (5ml) 3 times | 1/2 | No or once = 0.5, More than once = 1                                    |
| 10  | Laryngeal movement in the first feed water stage | 1/2 | No or once = 0.5, More than once = 1                                    |
| 11  | Repeat swallowing in the first feed water stage | 1/2 | No or once = 0.5, More than once = 1                                    |
| 12  | Cough in the first feed water stage | 1/2 | No = 0.5, Cough = 1                                                    |
| 13  | Stridor in the first feed water stage | 1/2 | Normal = 0.33, Weaken or voice hoarse = 0.66, Can’t pronounce = 1      |
| 14  | Outflow in the second feed water stage (60ml) | 1/2 | Can = 0.5, Cannot = 1                                                  |
| 15  | Cough in the second feed water stage | 1/2 | Can = 0.5, Cannot = 1                                                  |
| 16  | Stridor in the second feed water stage | 1/2 | Can = 0.5, Cannot = 1                                                  |
| 17  | The judgment of laryngeal function in the second feed water stage | 1/2/3 | Normal = 0.33, Weaken or voice hoarse = 0.66, Can’t pronounce = 1 |
| 18  | Regurgitation phenomena      | 1/2     | No = 0.33, Potencial = 0.67, Regurgitation = 1                          |

Figure 6. Comparisons in terms of training performance between the proposed hybrid classifier and the traditional DBN mode ((a): the training performance of the hybrid classifier. (b): the training performance of the traditional classifier).

Table 4. Comparisons in terms of ability to generalize between the proposed hybrid classifier and the traditional DBN model.

| Case number | Terms      | Diagnostic results |
|-------------|------------|--------------------|
| 1-10        | Experts    | 0.533 0.667 0.200 0.467 0.800 1.000 0.533 0.933 0.667 0.867 |
|             | DBN        | 0.509 0.675 0.201 0.466 0.883 0.931 0.548 0.879 0.611 0.858 |
|             | NPCA-DBN   | 0.555 0.665 0.202 0.477 0.808 0.995 0.521 0.202 0.656 0.871 |
| 11-20       | Experts    | 0.133 0.400 0.267 0.400 0.867 0.200 0.200 0.467 0.467 0.467 |
|             | DBN        | 0.139 0.485 0.226 0.430 0.858 0.196 0.200 0.493 0.466 0.466 |
|             | NPCA-DBN   | 0.135 0.656 0.267 0.414 0.871 0.204 0.189 0.446 0.460 0.489 |
| 21-30       | Experts    | 0.467 0.007 0.933 0.467 0.600 0.933 0.267 0.933 1.000 0.200 |
|             | DBN        | 0.466 0.081 0.964 0.493 0.579 0.879 0.291 0.925 0.982 0.200 |
|             | NPCA-DBN   | 0.460 0.030 0.944 0.446 0.603 0.808 0.267 0.929 0.987 0.189 |
From the **Table 4**, we can see that there are three cases (10th, 15th, 23rd) with big deviation of the diagnostic results by the improved DBN proposed in this paper, as well, there are six cases (9th, 10th, 12th, 15th, 22th, 28th) with the error diagnostic results. Furthermore, the accuracy of the proposed DBN is higher than that of the traditional methods.

**Table 4** presents the diagnostic accuracies of the proposed NPCA-DBN classifier and the traditional classifiers.

From **Table 5**, we found that the diagnostic accuracy of the proposed NPCA-DBN classifier is higher than the other the traditional classifiers. Firstly, PLS is suitable to deal with the linear classifying problems. Then, SVM showed better performance than the traditional DBN when the sample data is small. However, the structure of SVM is more complex when the sample data is larger, as well, the computational complexity increases with the number of training samples becomes exponentially.

From **Table 6** we can see that the maximum relative errors of the proposed classifier and the traditional DBN are 10.621 and 3.241, respectively.

### 5.2. Locations of key elements

Based on Eq. (10), contributions of the 18 original water swallow test scale diagnostic item elements are used to represent the diagnostic results. **Figure 7** shows that the contributions of the 4th, 6th, and 10th water swallow test scale items (lip closure, judgment of laryngeal function, and the laryngeal movement in the first feed water stage) in the 5th generalization case (whose diagnostic results are mouth period, tongue period, and pharynx period) are the highest. Therefore, we can conclude that the laryngeal movement in the first feed water stage is the key factor of the diagnostic results. **Figure 8** shows that contribution of the 4th and 7th water swallow test scale item (lip closure and the pharyngeal reflex) in the 25th generalization sample (whose diagnostic result is tongue period and pharynx period) is the highest. Similarly, we can conclude that the ‘pharyngeal reflex’ is the key factor of the diagnostic results.

Using the statistical results of the diagnostic cases, key element locations towards different diagnosis classification results are listed in **Table 7**.

| Case number | Terms | Diagnostic results's MRE |
|-------------|-------|--------------------------|
| 1–10        | dbn   | 0.044 0.012 0.006 0.002 0.104 0.069 0.027 0.058 0.085 0.010 |
|             | NPCA-DBN | 0.041 0.665 0.202 0.477 0.808 0.995 0.521 0.202 0.656 0.871 |
| 11–20       | dbn   | 0.047 0.214 0.154 0.075 0.010 0.022 0.001 0.055 0.001 0.002 |
|             | NPCA-DBN | 0.012 0.641 0.001 0.035 0.004 0.018 0.054 0.045 0.014 0.016 |
| 21–30       | dbn   | 0.001 10.621 0.033 0.055 0.036 0.058 0.089 0.008 0.018 0.001 |
|             | NPCA-DBN | 0.014 3.241 0.011 0.045 0.005 0.134 0.002 0.004 0.013 0.054 |
generate diagnosis results objectively, leading to more effective treatments.

Additionally, the Cohen’s Kappa [46] index is used to measure [47,48] the agreement between the proposed method and the group of physicians. The index (Kappa = 0.867 and P < 0.05) indicates that the proposed method has a high consistency with the doctors’ diagnosis.

6. Conclusion

In order to extract nonlinear features and qualitative information (expert doctors’ expertise) from high dimensional diagnostic scale data and reduce doctors’ heavy workload during the diagnostic process, this study proposes a hybrid classifier based on nonlinear-principal component analysis (NPCA) and deep belief networks (DBN) in contrast to traditional approaches. The hybrid classifier can extract the diagnostic scale data’s nonlinear features and hidden information and provide the key scale elements’ locating methods towards the diagnostic results, along with improve the efficiency and pertinence of diagnosis.

In addition, based on the ‘information entropy theory’, we prove that the novel NPCA-DBN classifier takes advantage of modeling accuracy and convergence speed, and overcomes the shortcomings of small sample sizes. Cases of dysphagia diagnoses collected from a hospital are used to test the applications of the methods, yielding satisfactory results compared with conventional classifiers. Based on comparative diagnostic test with the 180 real dysphagia diagnosis examples from the China-Japanese Friendship Hospital, we found that the diagnostic accuracy of the proposed NPCA-DBN classifier is higher than the other the traditional classifiers (i.e. DBN, PLS and SVM).

This research is still in the beginning stages, and there is need for further investigation on key issues such as diagnostic modeling methods with uncertain information (e.g. water swallow test scale item values that are unavailable in the presence of serious physical conditions of dysphagia patients), or with small diagnosis samples.

Acknowledgements

This study is supported by the Open Research Project under Grant from SKLMCCS [20150103], the Beijing outstanding talent training Project [2015000020124G041] and the National Natural Science Foundation of China [61603023].

Disclosure statement

No potential conflict of interest was reported by the authors.

Funding

This study is supported by the Open Research Project under Grant from SKLMCCS [20150103], the Beijing outstanding talent training Project [2015000020124G041] and the National Natural Science Foundation of China [61603023].

References

[1] Logemann JA. Evaluation and treatment of swallowing disorders. 2nd ed. Austin: Pro-Ed:1998.
[2] Smithard DG, O’Neill PA, England RE, et al. The natural history of dysphagia following a stroke. Dysphagia. 1997;12:188–193.
[3] Lawrence ES, Coshall C, Dundas R, et al. Estimates of the prevalence of acute stroke impairments and disability in a multiethnic population. Stroke. 2001;32:1279–1284.
[4] Hassan HE, Aboloyoun AI. The value of bedside tests in dysphagia evaluation. Egypt J Ear Nose Throat Allied Sci. 2014; 15:197–203.
[5] Ellis AL, Hannibal RR. Nursing swallow screens: why is testing water only not enough? J Neurosci Nurs. 2013;45:244–253.
[6] Smithard DG, O’Neill PA, Park C, et al. Complications and outcome after acute stroke does dysphagia matter? Stroke. 1996;27:1200–1204.
[7] Smithard DG, O’Neill PA, Martin DF, et al. Aspiration following stroke: is it related to the side of the stroke? Clin Rehabil. 1997;11:73–76.
[8] Ramsey DJ, Smithard DG. Assessment and management of dysphagia. Hosp Med. 2004;65:274–279.
[9] Osawa A, Maeshima S, Tanahashi N. Water-swallowing test: screening for aspiration in stroke patients. Cerebrovasc Dis. 2013;35:276–281.
[10] MartinoMaki R, Diamant EN. Identification of dysphagia using the Toronto Bedside Swallowing Screening Test (TOR-BSSST): are 10 teaspoons of water necessary? Int J Speech Lang Pathol. 2014;6:193–198.
[11] Suiter DM, Leder SB. Clinical utility of the 3-ounce water swallow test. Dysphagia. 2008;23:244–250.
[12] Mallela UK, Upadhya A. Buckling load prediction of laminated composite stiffened panels subjected to in-plane shear using artificial neural networks. Thin-Wall Struct. 2016;102:158–164.
[13] Stojanovic B, Milivojevic M, Milivojevic N, et al. A self-tuning system for dam behavior modeling based on evolving artificial neural networks. Adv Eng Soft. 2016;97:85–95.
[14] Wahab OA, Mourad A, Otrok H, et al. CEAP: SVM-based intelligent detection model for clustered vehicular ad hoc networks. Expert Syst Appl. 2016;15:40–54.
[15] Yan Y, Zheng WK, Bao J, et al. An enhanced Mary SVM algorithm for multi-category classification and its application. Neurocomputing. 2016;187:119–125.
[16] Fu JH, Lee SL. A multi-class SVM classification system based on learning methods from indistinguishable Chinese official documents. Expert Syst Appl. 2012;39:3127–3134.
[17] Yardimci A. Soft computing in medicine. Applied Soft Comput. 2009;9:1029–1043.
[18] Teodorescu IA, Chelaru M, Kandela A, et al. Fuzzy methods in tremor assessment, prediction, and rehabilitation. Artif Intell Med. 2001;21:107–130.
[19] Montani S. Exploring new roles for case-based reasoning in heterogeneous AI systems for medical decision support. Artif Intell. 2008; 28:275–285.
[20] Pralreddy S, Reddy NP, Green P, et al. Neural networks in computer-aided diagnosis classification of dysphagic patients. Paper presented at 14th Annual International Conference of the IEEE Engineering in Medicine and Biology Society; 1992(4). p. 1517–1518.
[21] Prabhu DNF, Reddy NP, Canilang EP. Neural networks for recognition of acceleration patterns during swallowing and coughing. Proceedings of the 16th Annual International Conference of the IEEE, 1994. p. 1105–1106.
[22] Reddy NP, Prabhu D, Palreddy, et al. Redundant neural networks for medical diagnosis: diagnosis of dysphagia. Intell Syst through Artif Neural Networks. 1995;5:699–704.
[23] Lee J, Blain S, Casas, et al. A radial basis classifier for the automatic detection of aspiration in children with dysphagia. J NeuroEng Rehabilit. 2006;3:1–3.
[24] Cola PC, Gatto AR, da Silva RG, et al. Classification of normal swallowing and oropharyngeal dysphagia using wavelet. Appl Math Comput. 2009;207:75–82.
[25] Saki F, Kehtarnavaz N. Online frame-based clustering with unknown number of clusters. Pattern Recogn. 2016;57:70–83.
[26] Li H, Chung F-I, Wang S. A SVM-based classification method for homogeneous data. Appl Soft Comput. 2015;36:228–235.
[27] Embrechts MJ, Rossi F, Schleif F-M, et al. Advances in artificial neural networks, machine learning, and computational intelligence. Neurocomputing. 2014;141:1–2.
[28] Zhang YP, Li X, Zhang ZF, et al. Deep learning driven blockwise moving object detection with binary scene modeling. Neurocomputing. 2015;168:454–463.
[29] Zhou Chen S, Wang QX. Active deep learning method for semi-supervised sentiment classification. Neurocomputing. 2013;120:536–546.
[30] Shen FR, Chao J, Zhao JX. Forecasting exchange rate using deep belief networks and conjugate gradient method. Neurocomputing. 2015;167:243–253.
[31] Li H, Li XY, Ramanathan M, et al. Identifying informative risk elements and predicting bone disease progression via deep belief networks. Methods. 2014;69:257–265.
[32] Shang C, Yang F, Huang DX, et al. Data-driven soft sensor development based on deep learning technique. J Process Control. 2014;24:223–233.
[33] M, Wöllmer F, Eyben Bö, Schuller, et al. Spoken term detection with connectionist temporal classification: a novel hybrid CTC-DBN decoder. Poster presented at Acoustics, speech, and signal processing, ICASSP-88; 2010.
[34] Zhong Pin, Gong gZ. A hybrid DBN and CRF model for Spectral-Spatial classification of hyperspectral images. Stat Optim Inform Comput 2017;5:75–98.
[35] Verbanck M, Josse, et al. Regularised PCA to denoise and fault isolation. Chemometrics Intell Lab Sys. 2009;5:699–704.
[36] ZhouChen S, Wang QX. Active deep learning technique. J Process Control. 2015;69:257–265.
[37] J, Song ZZ. Dynamic mixture probabilistic PCA classifier modeling and application for fault classification. J Chemometrics. 2015;29:361–370.
[38] Li EG, Yu JS. An input-training neural network based nonlinear principal component analysis approach for fault diagnosis. Control Decision. 2003;18:229–231.
[39] Niu Z, Liu ZJ, Niu yg. RBF networks-based nonlinear principal component analysis for process fault detection. Machine Learn Cybernet. 2005;8:4784–4788.
[40] Hessam M, Tarek E, Nejib S. A nonlinear principal component analysis approach for turbulent combustion composition space. Int J Hydrogen Energy. 2014;34:4622–4633.
[41] Mansouri M, Nounou M, Nounou H, et al. Kernel PCA-based GLRT for nonlinear fault detection of chemical processes. J Loss Prevent Process Ind. 2016;40:334–347.

[42] Tan S, Mavrovouniotis ML. Reducing data dimensionality through optimizing neural network inputs. Aiche J. 1995;41:1471–1480.

[43] Liu J, Yu J, Yan B. A new method to fault nonlinear diagnosis for a class of systems. Chinese J Electron. 2011;2:217–222.

[44] Sanchez D, Melin P, Castillo O. Optimization of modular granular neural networks using a hierarchical genetic algorithm based on the database complexity applied to human recognition. Inform Sci. 2015;309:73–101.

[45] Ni Y-y, Zhang J-x. Modeling and forecasting for multivariate time series using a vector autoregression model. Chinese J Health Stat. 2014;1:53–56.

[46] Liang X, Chen S-w, Zhou P, et al. Diagnosis and differential diagnosis of two dimensional and three dimensional transvaginal Ultrasound. Chinese Ultrasound Med. 2012;28:355–357.

[47] Deng F, Guan S, Yue X, et al. Energy-based sound source localization with low power consumption in wireless sensor networks. IEEE Trans Ind Electron. 2017;64:4894–4902.

[48] Deng F, Guo S, Zhou R, et al Sensor multifault diagnosis with improved support vector machines. IEEE Trans Autom Sci Eng. 2017;14:1053–1063.