IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received 22 September 2022, accepted 15 October 2022, date of publication 19 October 2022, date of current version 3 November 2022.

Digital Object Identifier 10.1109/ACCESS.2022.3215706

RESEARCH ARTICLE

Prediction Methods of Common Cancers in China Using PCA-ANN and DBN-ELM-BP

HUITAO QI[1], SHUANGBO XIE[2], YANLI CHEN[1], CHENGQIAN WANG[2], TINGTING WANG[2], BIN SUN[2], AND MINGXU SUN[2]

[1]Affiliated Central Hospital of Shandong First Medical University, Jinan 250013, China
[2]School of Electrical Engineering, University of Jinan, Jinan 250022, China

Corresponding author: Mingxu Sun (cse_sunmx@ujn.edu.cn)

This work was supported by the Science and Technology Development Projects of Jinan Health Commission under Grant 2020-3-04.

This work involved human subjects or animals in its research. Approval of all ethical and experimental procedures and protocols was granted by the Medical Ethics Committee of Jinan Central Hospital under Application No. 2022-216-01.

ABSTRACT Accurate prediction of cancer cases is crucial for diagnosis of cancer at an early stage because a long-lasting chronic disease is harmful to both physical and mental health. While medical data about healthcare and health obtained from questionnaire, the true positive rate of cancers predicted by traditional methods is low. Machine learning can provide a pattern for classification for types of cancer (mainly including lung cancer, liver cancer, upper gastrointestinal cancer, lower gastrointestinal cancer and breast cancer) using instances of early questionnaire screening. The screening covered 3411 respondents in this study. Principal component analysis (PCA) is used to generate attributes, coupled with artificial neural network (ANN) technology to conduct cancer prediction by providing 28 attributes into models. While deep belief network (DBN) is used for unsupervised training and extracting relevant attributes. Extreme learning machine (ELM) optimizes DBN and conducts supervised classification. Back propagation (BP) algorithm conducts supervised fine tuning. Finally, PCA-ANN and DBN-ELM-BP common cancers prediction models are established. The training set and testing set of PCA-ANN model gives 35.29% and 37.5% sensitivity, 98.36% and 98.33% specificity, 97.01% and 97.85% accuracy, an area under the receiver operating characteristic curve (AUC) 0.7245 and 0.7221, respectively. While the training set and testing set of DBN-ELM-BP model gives 58.83% and 62.5% sensitivity, 98.31% and 98.52% specificity, 98.03% and 98.24% accuracy, AUC 0.7747 and 0.7238, respectively. The results show that DBN-ELM-BP model can provide a method to predict the possibility of common cancers, which is non-invasive and economical for clinicians to make diagnostic decisions.

INDEX TERMS Cancer prediction, machine learning, common cancers, classification.

I. INTRODUCTION

Cancer has been one of the most severe diseases since it was discovered. Unfortunately, the number of cancer patients is rapidly growing in China. According to global cancer statistics, 18.1 million new cancer patients occurred globally in 2018 and China accounts for 21%. There are 9.6 million people dying from cancer globally every year and China accounts for 23.9%, with higher rates of incidence and mortality than the global average [1]. Therefore, using cancer screening measures to effectively detect cancer and reducing the incidence of cancer is an urgent task to accomplish.

At present, the methods of early prediction and diagnosis of cancer basically rely on imaging, physiology or information related to genes and biomarkers. However, detection based on susceptibility genes can only predict the risk of developing cancer in the lifetime, but cannot dynamically reflect the state of the early presence or absence of tumors in the body. Some more limitations including the method for determining

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

biomarkers is complex and expensive, the nature of many markers is not fully understood, there are hundreds of tumor markers while the sensitivity or specificity of a single marker is often low. The detection of early cancer is important but current solutions are not sufficient. Though the accuracy of pathological diagnosis based on imaging is high, it is only suitable for the diagnosis after the tumor has grown to a certain volume. Also, an interventional surgery is required for the patient, so it is hard to use as an early screening. Thus, with the limitations of these methods, there is an urgent need to establish convenient cancer risk early predicting models that can dynamically detect multiple times, so as to provide early warning of cancer risks for high-risk people as soon as possible.

Multiple risk factors influence incidence of cancer to varying degrees. The risk factors of cancer mainly include personal health condition, family history and personal habits. Considering gastrointestinal cancer (GI), the related risk factors are hypertension, diabetes, Helicobacter pylori, drinking, salt diet, family history and personal history [2]. Recent study shows that the risk of people getting GI cancer is 27% higher when having diabetes at the same time, doubled when having overweight (BMI $\geq$ 25) problem, three times higher when having family history [3]. Smoking history, family history, occupation exposure are related to lung cancer, urban male smokers have 3.78 times higher risk of lung cancer compared to non-smokers [4]. The risk of people getting lung cancer is 1.7 times higher when having a positive lung cancer family history. If their first-degree relatives with the history cases, the risk is added to 2-4 times [5]. The key factors that contribute to liver cancer are Hepatitis B surface antigen, cirrhosis and family history. If their first-degree relatives with history cases, the risk of getting liver cancer is added to 5 times [6]. Dyslipidemia, family history and benign breast tumours are connected with breast cancer. A woman with positive breast cancer family history has 1.07-3.97 times risk of getting it when being compared to a woman without family history [7]. According to research [8], there is an association between cancer and stroke. Since five common cancers in China are multifactorial diseases, we consider getting risk factor information using practically easier means, such as personal health condition, family history and personal habits, which can greatly improve the screening efficiency, reduce the screening cost and make the most use of limited health resources. As a result, a standard predictive model can be established through analyzing the relationships between cancers and myriad risk factors.

Therefore, this study makes use of reliable methods, especially machine learning technology, to predict whether a person has a high probability of getting cancer and compare the performance of the PCA-ANN and DBN-ELM-BP prediction algorithms. The proposed solution shows promising results of cost and time reduction when compared to traditional cancer screening methods. So that assist clinicians in diagnosing patients and achieve the goal of cancer prevention. The contributions of this study are as follows:

- Proposed PCA-ANN model: In this model, PCA is used to reduce the data dimensions and extract important attributes.

Based on the Levenberg-Marquardt (L-M) algorithm, Bayesian regularization algorithm is applied to improve the generalization ability of NN model. This model has good ability in processing nonlinear modeling.

- Proposed DBN-ELM-BP model: The novel proposed DBN-ELM-BP algorithm solves the problem of fine tuning the network weight and deviation, achieves better performance than PCA-ANN, improves the true positive rate of common cancers prediction. In this model, after unsupervised DBN pretraining and supervised ELM fine-tuning, BP is applied to the DBN-ELM algorithm. DBN is used for the unsupervised pretraining phase, solves the problem of setting lots of labeled training data. This model is used to predict the cancer status through the data from screening questionnaire of chronic diseases, which can effectively achieve the purpose of assisting common cancers screening and prevention.
- Proposed ELM and BP to improve DBN algorithm: The ELM classifier is used to calculate the weight between the last hidden layer and the output layer, so that output matrix is received from the last Restricted Boltzmann machine (RBM) of DBN, and the deviation is computed. Finally, the weight matrix and deviation are updated by BP. In this paper, ELM and BP are used for the network of DBN to further improve the precision of prediction.

The remaining portion of this article is arranged as follows. Section 2 depicts the works related to our research. Section 3 proposes the frame of health information preprocessing and common cancers prediction. Section 4 describes the PCA-ANN and DBN-ELM-BP algorithms and performance evaluation. Section 5 presents the results and discuss the inadequacy and future improvements. Finally, Section 6 is the conclusion of this paper.

II. RELATED WORKS

Timely detection of cancer cases is critical for early cancer screening. Accordingly, accurate prediction of cancer cases has become one of the most important tasks. Some relevant AI studies about imaging [9], [10], [11], [12], [13], [14], physiology, interview survey (combined with physiological data) and interview survey (only questionnaire) are shown in Table 1. Cancer-related research has become extensive and effective in recent years, with different research methods selected according to different data types.

Considering the heavier disease burden of five common cancers in China and the need for relatively simple, convenient, and low-cost screening tools. To further study the effectiveness of cancer prediction, PCA-ANN and DBN-ELM-BP algorithms are developed with medical data about health informatics from the screening questionnaire of chronic diseases, the training set of PCA-ANN provides 35.29% sensitivity, 98.36% specificity, 97.01% accuracy, and

TABLE 1. Previous work.

Methods	Type of data	Results	Advantages	Limitations
CNN [15]	Imageology, digital mammographic examination images	Sensitivity 86% Specificity 79%	Well performances Shorter reading times	False-positive assessments Dataset not from screening practice Not consider regional differences in screening practice
CNN combined with humans [16]	Imageology, dermoscopic images	Sensitivity 86.1% Specificity 81.5%	Well performances Better classify the static lesions of skin cancer	Only biopsy-validated images, has some bias May provides answers for physicians in this anonymous survey
CNN [17]	Imageology, endoscopic images of esophageal cancer	Accuracy 98%	Can detect lesions less than 10 mm in size Check stored endoscopic images quickly	False positives Only suitable for high-quality endoscopic images
SVM [18]	Physiology	Accuracy 80.28%	Inexpensive Noninvasive	False positives Low positive predictive value (PPV)
CNN [19]	IEEE International Symposium on Biomedical Imaging (ISBI) 2018 Lung Nodule Malignancy Prediction dataset	AUC 0.913	Well performance Reduce false positives High precision	High intra-class variance Nodule detection unbalanced
ANN [20]	Wisconsin Breast Cancer Dataset	Accuracy 97%	High accuracy Noninvasive	False positives Moderately expensive
Expectation-maximization and ANN [21]	National Health Interview Survey (NHIS) and the Prostate, Lung, Colorectal, Ovarian Cancer Screening (PLCO) datasets	Sensitivity 63% Specificity 82%	Noninvasive Inexpensive	False positives Low PPV
ANN [22]	NHIS and PLCO datasets	Sensitivity 69.4% Specificity 81.3% AUC 0.80	Noninvasive Inexpensive	False positives Low PPV Not suitable for general screening of older woman
ANN [23]	NHIS dataset	Sensitivity 75.3% Specificity 80.6% AUC 0.86	Noninvasive Inexpensive Easy to implement	Less Sensitive than low-dose computed tomography Low PPV
ANN [24]	NHIS and PLCO datasets	Sensitivity 80.7% AUC 0.85	High sensitivity Suitable for pancreatic cancer early prediction	False positives Low PPV
ANN [25]	NHIS dataset	Sensitivity 86.2% Specificity 62.7%	High sensitivity Noninvasive Inexpensive	False positives Low PPV
ANN [26]	NHIS dataset	Sensitivity 23.2% Specificity 89.4% AUC 0.72	Noninvasive Inexpensive	Low sensitivity Low PPV Uncertain the stage of prostate cancer cases
ANN [27]	NHIS dataset	Sensitivity 57% Specificity 89%	Inexpensive Can stratify risk	Low PPV Assumes integrity of data Cannot be used for screening
ANN [28]	NHIS and PLCO datasets	Sensitivity 62.1% Specificity 47.4% AUC 0.56	Noninvasive Inexpensive	Lower performance than Gail model False positives Low PPV

AUC 0.7245, the testing set of PCA-ANN provides 37.50% sensitivity, 98.33% specificity, 97.85% accuracy, and AUC 0.7221; while the training set of DBN-ELM-BP gives 58.83% sensitivity, 98.31% specificity, 98.03% accuracy, and AUC measuring 0.7747 respectively, the testing set of DBN-ELM-BP gives 62.50% sensitivity, 98.52% specificity, 98.24% accuracy, and AUC measuring 0.7238 respectively. Two prediction models are shown to have high specificity and low sensitivity which can be used as an effective measure to aid clinician in diagnosis by providing concurrent objective analysis results during estimation of cancer.

III. MATERIALS AND METHODS

This study collects the data of 3411 participants obtained from the screening questionnaire of chronic diseases in a city in 2019, which contained 28 attributes. PCA is used to

reduce the data dimensions of these characteristic attributes. By using essential features related to cancer, PCA-ANN and DBN-ELM-BP algorithms can classify patients as non-cancer (low risk) or cancer (high risk) to establish common cancers prediction models to improve the efficiency and accuracy of early cancer screening. Figure 1 shows the whole scheme of cancer prediction model. First, fill out the questionnaire as required; then assess the high-risk groups through Harvard cancer risk index, next; preprocess the questionnaire data, including data desensitization, filling in missing values, removing abnormal values, train PCA-ANN and DBN-ELM-BP models, and finally input the test data to get the prediction results of common cancers.

A. DATA COLLECTION

Based on the existing screening questionnaire of chronic diseases [29] and these design methods [30], [31], [32], [33], [34] including monofactor analysis, logistic regression, Chi-square test, using HTML5, CSS, node.js and other technologies to design a common cancers screening questionnaire and database that limit subjective answers according to the response time of patients' reading questions: the questionnaire options need to exceed the response time of the corresponding questions before they can be answered. At the same time, several lie detection questions are added to identify invalid samples in the follow-up processing in order to quickly check whether the samples really answer the questionnaire.

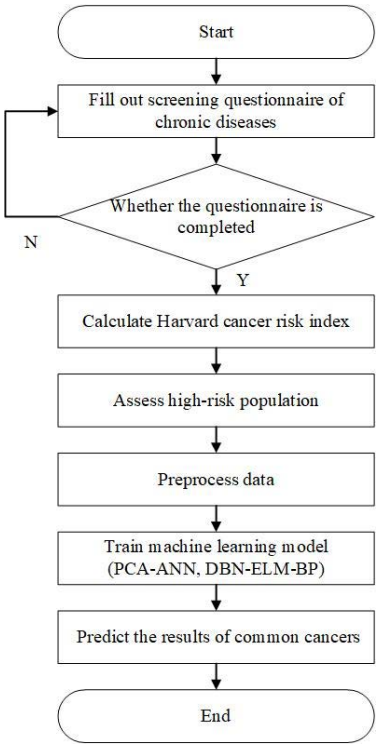


FIGURE 1. Block diagram of common cancers prediction methods.

Cluster sampling is applied to screen suitable population in 25 communities, and the assessment table is filled in. The inclusion criteria: residents of the district; fully capacity; voluntarily signed the informed consent. After the clinical examination such as computed tomography, abdominal color Doppler ultrasound, electronic gastroscopy, electronic bowel endoscopy, breast color Doppler ultrasound can confirm whether a participant has cancer or not. Meanwhile, patients' data mainly come from personal health information, the risk factors of upper GI cancer, lower GI cancer, lung cancer, liver cancer, breast cancer and other information. This questionnaire includes 28-dimension items, mainly included personal information, Privacy information and information of common cancers. The response time of different questions is shown in Table 2.

TABLE 2. The response time for different types of questions.

Type	Content of question	Response time (mean value /s)
Personal information	Height	3.47
	Weight	12.84
	Smoking	11.72
	Phone number	13.76
Privacy information	-	18.19
Common cancers	-	10.02

B. DATASET

The dataset collected in this work includes all the data collected from an entire year's screening questionnaire of chronic diseases in a city to maximize the generalization of the results. The probability of cancer is predicted based on the risk factors associated with cancer. The method used is supervised binary classification. The dataset contains 28 attributes with 3386 (99.27%) negative cases and 25 (0.73%) positive cases among 3411 participants. A positive example is a patient with cancer, denoted by 1 while negative cases represent regular participants, indicated by 0. Table 3 gives a detailed description of all attributes in an entire year's screening questionnaire of chronic diseases.

There are two constructed datasets:

(1) DS1 (training set) = 2371 non-cancer participants+ 17 common cancer cases.

(2) DS2 (testing set) = 1015 non-cancer participants+ 8 common cancer cases.

The questionnaire data are randomly separated as 70% and 30% for training and testing respectively. Table 4 gives the partitioning of datasets. The results are measured by sensitivity, specificity, and accuracy for those datasets.

The Harvard cancer risk working group established the Harvard cancer risk index model [35], [36]. Based on data from lifestyle and physical examination to predict high-risk groups of cancer. The Harvard cancer risk index formula is as follows:

$$RR = \frac{\prod_{i=1}^{N} RI_i}{\prod_{j=1}^{N} (P_j \times RC_j + 1 - P_j)} \quad (1)$$

The formula (1) divides the same-sex age into a group, where *RI* and *RC* are the relevant risk of a disease compared with the general population in the same group, *RI* represents the related risk of a factor existing in the identical group, and *RC* is the assignment approved by the expert group for the related risk of a factor, P_j is the proportion of people in identical group who have a certain risk factor, and N is the amount of risk factors. Among them, *RC* value is the most critical factor in the whole Harvard cancer risk index model and benefits the evaluation effect of the risk assessment model. Of the total 3411 patients in this study, 3376 patients were assessed as high-risk group and 35 patients were assessed as healthy group.

The formula (1) divides the same-sex age into a group, where *RI* and *RC* are the relevant risk of a disease compared with the general population in the same group, *RI* represents the related risk of a factor existing in the identical group, and *RC* is the assignment approved by the expert group for the related risk of a factor, P_j is the proportion of people in identical group who have a certain risk factor, and N is the amount of risk factors. Among them, *RC* value is the most critical factor in the whole Harvard cancer risk index model and benefits the evaluation effect of the risk assessment model. Of the total 3411 patients in this study, 3376 patients were assessed as high-risk group and 35 patients were assessed as healthy group.

C. DATA PREPROCESSING

In an actual dataset, there may be missing values or noisy data, which may affect the accuracy and calculation speed of classification. It is crucial to preprocess the data to obtain high accuracy. Data cleaning, transformation, and simplification are conducted during data preprocessing.

Data desensitization is performed to filter out personal privacy information such as name, ID number, contacts and address. Data cleaning includes filling in missing values and deleting noise data. Using the mean of the attribute to impute missing values about ''smoking history'' and ''HBsAg''. The noise data contain outliers and will be deleted. Some abnormal data values of ''age'' and ''weight'' are obviously high (more than 100 times higher than the normal value). At the same time, in our dataset, ''height'' and ''weight'' shouldn't be zero. All zero values have been taken place of the median value of the attributes. In addition, breast cancer related attributes (''family history of malignant breast tumor'' and ''history of benign breast tumor'') in male patient instances remain null.

After data preprocessing, missing data were filled out, and then normalized according to the reference range of each inspection data to eliminate the effect of its size difference for the following data analysis. Normalize the data and convert the value scope of the data into a unified region [0, 1]. The function expression is:

$$Y_{norm} = \frac{Y_i - Y_{\min}}{Y_{\max} - Y_{\min}} \tag{2}$$

where Y_{norm} is the normalized data, Y_i is the i-th column of the original data, and Y_{max}, Y_{min} is the maximum value and minimum value of the original dataset respectively.

IV. PREDICTION MODELS

In this study, two artificial intelligence models are used to predict the probability of common cancers.

A. PCA-ANN

Too many variables in properties will add a lot of computational loads. PCA is a technology applied to data dimension reduction, and it is a method to extract important features from processed data of artificial intelligence models. PCA can create new variables that are independent of each other. The original information can be retained by these new variables to the greatest extent. PCA is used to analyze the correlation between attributes and extract important attributes from a whole dataset. Family history, personal history of upper GI cancer; family and personal history of lower GI cancer; family history of lung cancer and smoking are the important attributes in this study. The PCA algorithm is described below.

Calculate the covariance matrix C of the scalar matrix X:

$$C = \frac{XX^T}{n-1} \tag{3}$$

To calculate the characteristic equation of covariance matrix C:

$$\left|C - \lambda I_p\right| = 0 \tag{4}$$

$$\frac{\sum_{j=1}^{k} \lambda_j}{\sum_{j=1}^{n} \lambda_j} \geq 95\% \tag{5}$$

The eigenvalues are arranged from large to small, and the corresponding eigenvectors of the first K largest eigenvalues are taken to obtain an eigenvector matrix V of K rows and P columns.

Matrix Y has n rows and K columns is calculated. This matrix X is the variable matrix after dimensionality reduction, containing data of K principal components.

$$Y = (VX^T)^T = ZX^T \tag{6}$$

ANN is a mathematical representation of information processing behaviour in the biological nervous system, which refers to various logical structures of the human brain neural network. As a simulation of the human brain neuron network, one artificial neural network contains many nodes. Each node with a specific output function is connected to other nodes. After passing through the node, the data enter the next node through the activation function. The connection between different nodes has different influence degrees. The influence degree of the connection between two nodes on the output result can be determined by assigned weights. The neural network can change its internal structure according to different situations which provides good adaptability. This is characterized by highly nonlinear interaction between input

TABLE 3. Description of parameters in the dataset.

Parameter	Type	Range	Details of parameter
Gender	Binary	0 or 1	0 is a woman and 1 is a man.
Age	Continuous	22-73	Age range is 22-73
Height	Continuous	143-190	Height range is 143-190
Weight	Continuous	40-180	Weight range is 40-180(Kg)
Helicobacter pylori	Binary	0 or 1	0 is negative and 1 is positive
Heavy Drinking	Binary	0 or 1	Drinking over 3 times a week, and men drink more than 25g of alcohol a day while women drink more than 12.5g
Salt Diet	Binary	0 or 1	Daily salt intake > 6g is high salt
Family history of upper GI tumors	Binary	0 or 1	0 is not have and 1 is have
Pickled and Dried Food	Binary	0 or 1	Intake more than 3 times a week is excessive intake
Personal history of upper GI diseases	Binary	0 or 1	Atrophic gastritis, gastric ulcer or polyp, hypertrophic gastritis and Postoperative remnant stomach
First-degree relatives with Gastric cancer	Binary	0 or 1	0 is no and 1 is yes
Family history of lower GI tumors	Binary	0 or 1	0 is no and 1 is yes
Personal history of lower GI diseases	Binary	0 or 1	Intestinal polyps, Chronic colitis, chronic diarrhea; chronic constipation
Smoking history	Binary	0 or 1	Smoking ≥ 20 packs/year, ever smoking ≥ 20 packs/year and quit smoking < 15 years
History of high-risk occupation exposure	Binary	0 or 1	Asbestos, beryllium, uranium, radon and other exposures
History of pulmonary tuberculosis	Binary	0 or 1	0 is no and 1 is yes
Family history of lung cancer	Binary	0 or 1	0 is no and 1 is yes
HBsAg	Binary	0 or 1	0 is negative and 1 is positive
Other high-risk factors of liver cancer	Binary	0 or 1	Heavy drinking, nonalcoholic fatty liver, smoking
Cirrhosis	Binary	0 or 1	0 is no and 1 is yes
Family history of liver cancer	Binary	0 or 1	0 is no and 1 is yes
Family history of malignant breast tumor	Binary	0 or 1	0 is no and 1 is yes
History of benign breast tumor	Binary	0 or 1	0 is no and 1 is yes
Overweight or obese	Binary	0 or 1	0 is BMI < 26 and 1 is BMI ≥ 26
Hypertension	Binary	0 or 1	Blood pressure ≥ 140/90mmHg
Dyslipidemia	Binary	0 or 1	Triglyceride ≥ 2.26mmol/L, or total Cholesterol ≥ 6.22 mmol/L
Diabetes	Binary	0 or 1	Fasting blood glucose of venous blood ≥ 7.0 mmol/L, blood glucose of 2 hours postprandial ≥ 11.1 mmol/L
History of stroke	Binary	0 or 1	1 for have while 0 for the opposite

data and prediction target. Thus, this study uses an ANN model to classify cases as non-cancer or cancer. For optimal training of our neural networks, the training, and testing set ratios are 70%:30%, respectively. The number of neurons in two hidden and output layer is 20 and 1, respectively. There are 28 neurons in the input layer.

TABLE 4. The partitioning of datasets.

Sample	Total	Training	Testing
Total number	3411	2388	1023
Normal	3386	2371	1015
Common cancer	25	17	8

Using weight ω_1, ω_2, ω_3 and deviation θ_1, θ_2, θ_3, The artificial neural network propagates forward from input X to output Y by the logical activation function $f = f(z)$ with independent variable z ($z = \theta_i + \omega_i X$), where:

$$f(z) = 1/(e^{-z} + 1) \tag{7}$$

$$Y = \sum_{i=1} \omega_i f(z) + \theta_i \tag{8}$$

Two hidden layers are used to find the best training method based on the training algorithm. The training algorithm chosen is the Bayesian regularization algorithm, which takes longer but is better at solving complex problems. According to Levenberg-Marquardt (L-M) algorithm, the optimal combination of weights and deviations is determined to generate a well-distributed network. L-M algorithm updates the input data n-dimensional vector X as follows:

$$X_{k+1} = X_k - [J_k^T J_k + \lambda_k I]^{-1} J_k^T \varepsilon_k \tag{9}$$

In equation (9), ε is the error vector, and J is n dimensional Jacobian matrix. λ is the iteratively modified parameter during the operation of the algorithm, and I is an identity matrix, k is the dimension of the input data.

Based on the L-M algorithm, Bayesian regularization algorithm is applied to improve the generalization ability of the NN model.

The mean squared error function is:

$$E_d = \frac{1}{n} \sum_{i=1}^{n} (t_i - a_i)^2 \tag{10}$$

where: n is the total number of samples, t_i is target output of group i of this sample, a_i is the actual output of group i of the sample.

The objective function of neural network under regularization is:

$$F = \alpha E_\omega + \beta E_d \tag{11}$$

$$E_\omega = \frac{1}{m} \sum_{j=1}^{m} \omega_j \tag{12}$$

E_ω is the mean square deviation of all weights of the network, α and β are the regularization coefficients, m is the sum weights of the network, ω_j is the weight of the network.

When both α and β are small, the output of the training effect can be guaranteed to be smooth, and there will be no under-fitting or over-fitting. The Bayesian regularization method can adaptively adjust the size of the parameters α and β during the training process of the network, it can also effectively control the complexity of the network under the premise of ensuring the minimum sum of squares of the network error, thereby significantly improving the generalization ability of the network.

B. DBN-ELM-BP

1) DBN

Deep belief network is a probabilistic generative model which has several hidden layers. By training the network and updating the weights among neurons, the entire network is able to recover input data by the utmost probability. The deep belief network is composed of two networks. At the bottom of the network, a number of RBMs are superimposed, and the RBM of each layer adjusts the parameter θ, so that the features of the sample data are propagated to the hidden layer by the utmost probability, and then the extracted feature values of the hidden layer data are taken as the input of the next RBM. The data is extracted layer by layer by multiple RBMs, and finally a high-level data feature value is obtained. On top of the DBN is a BP neural network entity classifier. At this time, a label set is attached to the top layer. At the same time, the high-level eigenvalues extracted by multiple RBMs are regarded as the input of the BPNN for supervised learning. At this time, the network will adjust the parameter values between various networks according to the BP algorithm, and finally build the deep belief network. Figure 2 is the structure of DBN.

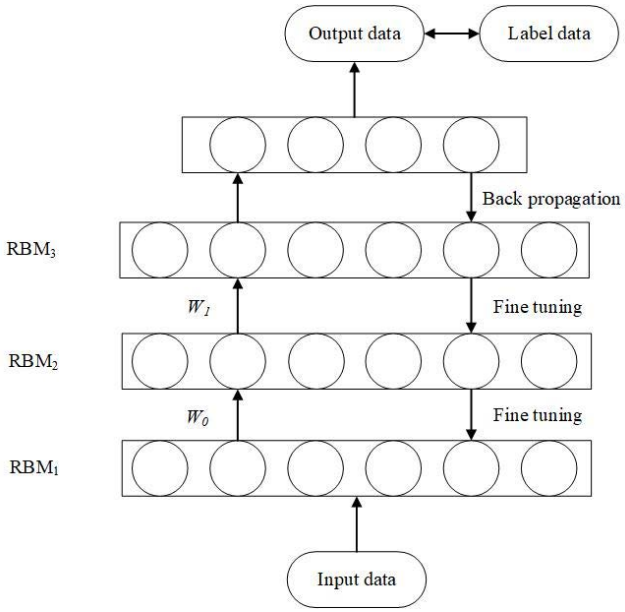


FIGURE 2. The structure of DBN.

DBN is made up of a stack of RBMs which are learned layer by layer through unsupervised greedy methods. Given the initial weights, train the first RBM and use its output as the input to next layer. The output of the lower RBM is always used as the input of the upper RBM. Under these circumstances, suppose that the neurons of the visible layer v and the hidden layer h of RBM are binary random, only take 0 or 1. Given (v, h), the energy function is defined:

$$E(v, h \mid \theta) = -\sum_{i=1}^{m} \sum_{j=1}^{n} v_i \omega_{ij} h_j - \sum_{i=1}^{m} a_i v_i - \sum_{j=1}^{n} b_j h_j \tag{13}$$

In equation (13), a_j and b_j are the deviation parameters of the visible unit i and the hidden unit j respectively; ω_{ij} is the connection weight; $\{a_j, \omega_{ij}, b_j\}$ are the parameters θ of RBM model. We can obtain the joint probability distribution function of the visible units and hidden units by the energy function.

$$P(v, h \mid \theta) = \frac{1}{Z_\theta} e^{-E(v,h \mid \theta)} \tag{14}$$

$$Z_\theta = \sum_v \sum_h e^{-E(v,h \mid \theta)} \tag{15}$$

where Z_θ is the normalization factor.

RBM has two-layer neural network, namely the visual layer v and hidden layer h. The superior unsupervised learning ability of RBM has attracted close attention since Hinton proposed contrast divergence, a fast-learning algorithm for RBM in 2002. Figure 3 is the structure of RBM.

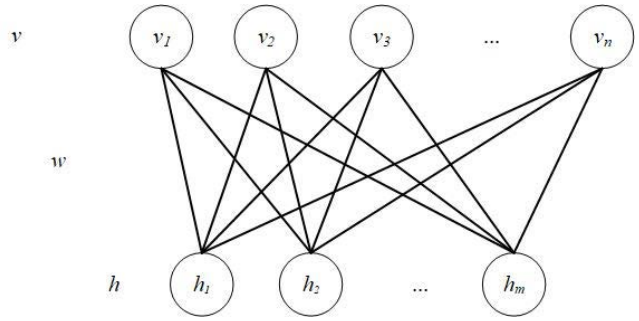


FIGURE 3. The structure of RBM.

The conditional probability is derived from the principle of Bayesian formula:

$$p\left(h_j = 1 \mid v\right) = \sigma(b_j + \sum_i \omega_{ij} v_i) \tag{16}$$

$$p\left(v_i = 1 \mid h\right) = \sigma(a_i + \sum_j \omega_{ij} h_j) \tag{17}$$

where, $\sigma(x)$ is logistic function.

$$\sigma(x) = \frac{1}{1 + e^{-x}} \tag{18}$$

To compute the update equation of each parameter, Hinton proposed contrast divergence (CD) algorithm [37]. Above all, initialize the visible layer on the basis of the first input sample; After that, calculate the conditional probability of hidden neurons based on values of visible layer and the formula of conditional probability; At last, Gibbs sampling is used to extract a sample by calculated probability. Using the sample to reconstruct the visible layer. The rules of updating related parameters through repeating this process.

$$W_{ij} = E(\langle V_i h_j \rangle_{\text{data}} - \langle V_i h_j \rangle_{\text{recon}}) \tag{19}$$

$$a_i = E(\langle V_i \rangle_{\text{data}} - \langle V_i \rangle_{\text{recon}}) \tag{20}$$

$$b_j = E(\langle h_j \rangle_{\text{data}} - \langle h_j \rangle_{\text{recon}}) \tag{21}$$

where, ε is the learning rate, $\langle \cdot \rangle_{d\text{ata}}$ and $\langle \cdot \rangle_{\text{recon}}$ represent the mathematical expectations of the data itself and after model reconstruction, respectively. With the criterion, proper weights are obtained, and the same method applied until the whole weights of RBM renewed.

2) ELM

Based on feedforward neural network (FNN), network structure of ELM includes input, hidden and output layer. Input weights and offsets be initialized randomly, then the corresponding output weights are obtained [38]. Here, assuming that there are M samples (X_j, t_j), this neural network representation of L hidden layer nodes is as follows:

$$\sum_{i=1}^{L} \beta_i g\left(W_i \cdot OX_j + b_i\right) = y_j, \quad j = 1, \ldots, M \tag{22}$$

where $g(x)$ is activation function of hidden layer, W_i is input weights while β_i is output weights, and b_i is offset of ith hidden layer. Obtain the minimum output error is the goal of single hidden layer neural network, namely:

$$\sum_{i=1}^{L} \beta_i g\left(W_i \cdot OX_j + b_i\right) = t_j, \quad j = 1, \ldots, M \tag{23}$$

O is output of the hidden layer node while H is the desired output:

$$O\beta = H \tag{24}$$

When the ELM is applied to train the single hidden layer neural network, parameters W_i and β_i are random. Once these two parameters are determined, the output matrix T is solely obtained, and training of whole network can be converted to solve the linear system to obtain output weight β, as follows:

$$\hat{\beta} = O^T H \tag{25}$$

3) DBN-ELM

Since Hinton et al. proposed the deep belief network in 2006, the algorithm has been successfully used in classification, regression, dimensionality reduction and other tasks [39]. Aiming at the problems of slow convergence and falling into local minimum caused by random initialization of parameters, the concept of unsupervised pretraining is proposed for deep belief network, and the problem of setting labeled training samples is solved at the same time. When unsupervised pretraining is used and there are training samples, both training errors and generalization errors will be significantly reduced [40]. In this paper, pretraining initialization parameters and BP for fine-tuning the network to further improve the accuracy and efficiency of classification. On the whole, DBN-ELM consists of DBN for feature extraction and ELM as classifier. It unites the ability of DBN feature extraction and ELM for fast learning speed and generalization, so as to improve the prediction performance.

Suppose DBN include n hidden layers, n-1 layers are initialized through greedy training, offsets and weights from n-1 to n hidden layer and from n to next layer are certained by ELM. m and l is the number of neurons in nth and n-1th hidden layer, respectively.

$$\sum_{i=1}^{m} \beta_i g\left(W_i \cdot O_{n-1} + b_i\right) = y_j, \quad j = 1, \ldots, l \tag{26}$$

The best prediction results can be obtained with the minimum output error:

$$\sum_{j=1}^{m} \left\| y_j - t_j \right\| = 0 \tag{27}$$

In addition, a special β_i is obtained, which makes the following formula true, namely:

$$\sum_{i=1}^{m} \beta_i O_{n,j} = t_j, \quad j = 1, \ldots, l \tag{28}$$

The above formula can be converted to:

$$O_n \beta = H \tag{29}$$

where, O_n represents the output from n-1th to nth layer, which can be expressed in the following form:

$$O_n \left(W_{1,\ldots,W_m}, b_1, \ldots, b_m, O_{n-1,1}, \ldots, O_{n-1,i} \right) = \begin{bmatrix} g\left(W_1 \cdot O_{n-1,l} + b_1\right) & \cdots & g\left(W_m \cdot O_{n-1,l} + b_m\right) \\ \vdots & \vdots & \vdots \\ g\left(W_1 \cdot O_{n-1,l} + b_i\right) & \cdots & g\left(W_m \cdot O_{n-1,l} + b_m\right) \end{bmatrix} \tag{30}$$

$$\beta = \begin{bmatrix} \beta_1^T \\ \vdots \\ \beta_m^T \end{bmatrix}, \quad H = \begin{bmatrix} H_1^T \\ \vdots \\ H_m^T \end{bmatrix} \tag{31}$$

At this time, the network is trained to obtain a special $\hat{W}_i$, $\hat{b}_i$, $\hat{\beta}$, so that:

$$\left\| O_m \left(\hat{W}_i, \hat{b}_i \right) \hat{\beta} - H \right\| = \left\| O_n \left(\hat{W}_i, \hat{b}_i \right) \hat{\beta} - H \right\| \tag{32}$$

Parameters W_i and β_i are selected by random determination in the last layer. When these two model parameters are determined, the output H is solely obtained. The solution of output weight of DBN is as follows:

$$\hat{\beta} = O_m^T H \tag{33}$$

4) DBN-ELM-BP

In this model, after unsupervised pretraining and ELM fine-tuning, BP is applied to fine-tune the DBN-ELM network by using labeled data. The unsupervised pretraining phase of DBN is performed first. Then, calculate the weights between the last hidden layer and the output layer by using ELM. Matrix H is considered identical to the weight matrix acquired through the last RBM, the matrix b is obtained. At last, the error is obtained, the weight matrix is updated by BP, and the whole network is trained. Figure 4 graphically illustrates the training process of the DBN-ELM-BP.

BP algorithm can divide learning process of network into two stages: First of all, signal is propagated forward, the input is processed through input, hidden and output layer, and actual output value of every unit is calculated; secondly, error back propagation: if there is an error between the expected and actual output, error signal will go back along the former route, and the error will be separated to all units of each layer to correct the neuron weights of each layer. These two processes are repeated until the network error meets the

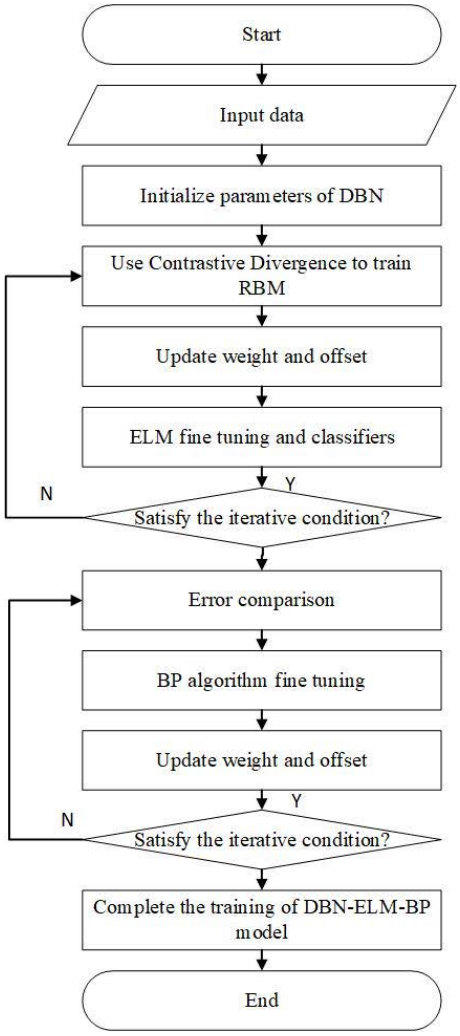


FIGURE 4. The prediction model of DBN-ELM-BP.

requirements. Assuming d_k and y_k are the expected and actual outputs of the kth neuron of output layer, the residual term of the node is:

$$\delta_k = y_k(1 - y_k)(d_k - y_k) \tag{34}$$

Error back propagation for layer l neurons δ_k^l is expressed as follows:

$$\delta_k^l = y_k^l(1 - y_k^l) \sum_{j=1}^{m} \omega_{ij}^l \delta_j^{l+1} \tag{35}$$

According to the gradient descent method, the weights and offsets of the network at the fine-tuning stage are updated according to equation (36) (37), where: ε Is the learning rate of the fine-tuning phase.

$$\omega_{ij}^l = \omega_{ij}^l + \varepsilon y_i^l \delta_i^{l+1} \tag{36}$$

$$b_j^l = b_j^l + \varepsilon \delta_j^{l+1} \tag{37}$$

C. PERFORMANCE EVALUATION

Sensitivity and specificity as the indexes to assess the performance of classification tests. Sensitivity weighs the ratio

of actual positives that are truly detected, while specificity weighs the ratio of actual negatives.

$$Sensitivity = \frac{TP}{TP + FN} \tag{38}$$

$$Specificity = \frac{TN}{TN + FP} \tag{39}$$

TP is for true positive counts, FP for false positive, FN for false negative, and TN for true negative. Accuracy can evaluate the overall effectiveness of the algorithm.

Several metrics are applied to measure the performance of classifiers of supervised AI algorithms. The measure of classification quality is based on a confusion matrix, which records correct and incorrect identification of each category.

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \tag{40}$$

The lateral axis of the ROC is 1−Specificity, and the vertical axis is Sensitivity. The closer the AUC is to 1, the better the classifier is.

V. EXPERIMENTAL RESULTS AND DISCUSSION

This work applies different classification algorithms to the dataset. Those algorithms show slightly different results from task to task due to the different measurement of each algorithm. Then, the results are evaluated with the metrics sensitivity, specificity and AUC. The sensitivity, specificity, accuracy and AUC of training set of the PCA-ANN model are 35.29%, 98.36%, 97.01% and 0.7245, respectively. The sensitivity, specificity, accuracy and AUC of testing set of the PCA-ANN model are 37.50%, 98.33%, 97.85% and 0.7221, respectively. The performance of PCA-ANN prediction model is shown in Table 5, and the ROC curve obtained by PCA-ANN is shown in Figure 5.

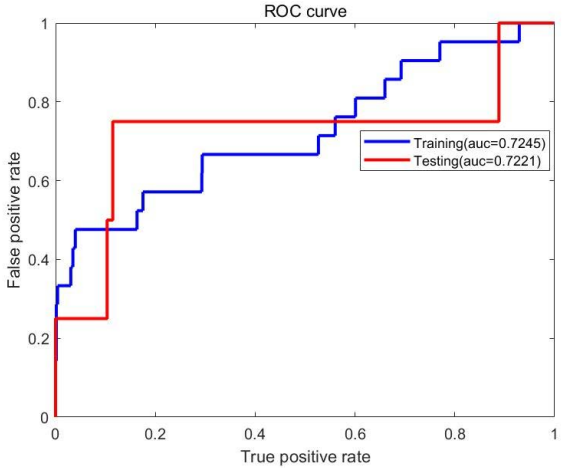


FIGURE 5. ROC curve of PCA-ANN.

After the PCA-ANN algorithm, the results of the DBN-ELM-BP model are better. The output value is range from 0 to 1, and the nearer the outcomes is to 1, the higher the possibility of cancer. The sensitivity, specificity, accuracy and AUC of training set of DBN-ELM-BP model are 58.83%, 98.31%, 98.03% and 0.7747, respectively. The sensitivity, specificity, accuracy and AUC of testing set of DBN-ELM-BP model are 62.50%, 98.52%, 98.24% and 0.7238, respectively. The performance of DBN-ELM-BP prediction model is shown in Table 6. The ROC curve obtained by DBN-ELM-BP is shown in Figure 6. ANN is often used as a good universal approximator in prediction and classification because of its strong ability in modeling nonlinear. However, shallow ANN may cause some problems because of its limited modeling and representation ability in complex problems such as common cancers prediction. In this case, DBN-ELM-BP achieves better performance than ANN. The sensitivity of the training and testing set exceeds 20% of that of ANN by using ELM and BP to fine tune the network weight and deviation.

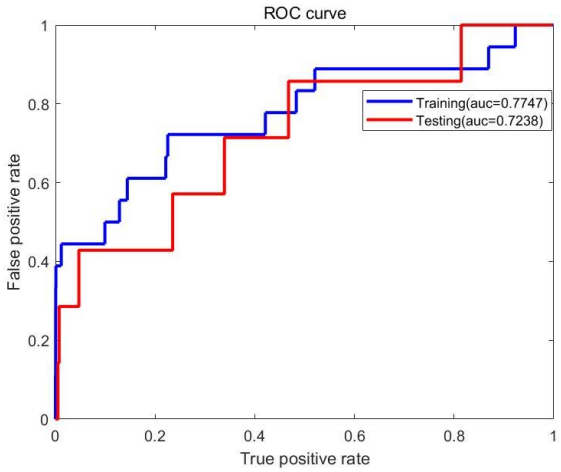


FIGURE 6. ROC curve of DBN-ELM-BP.

An important reason for high cancer mortality is that many cancer patients are diagnosed at an advanced stage and missed the time window for optimal treatment and intervention. Early screening of cancer is currently an important means to deal with cancer problems, which mainly depends on imaging methods. Final diagnosis of suspicious nodules and polyps needs to rely on biopsy or even pathological examination. These methods do not detect lesion until the disease has developed to a certain extent, which have great limitations in practice. At the same time, questionnaire samples are easy to obtain and facilitate multiple inspections. AI technology has achieved comparable or even better results in cancer screening than manual checks using imaging and physiological methods in refs [21], [22], [23], [24], [25], [26], [27], and [28]. With questionnaire screening, prediction models of common cancers based on AI technology can be established to buy time for cancer prevention and early intervention.

This study has advantages in several aspects. First, both models are trained and validated using a large dataset collected in 2019 with 3411 participants. The performances of models tested using the training and testing sets are close, indicating that the internal verification effect of models is good, and it shows a good generalization ability for other

TABLE 5. The performance of PCA-ANN.

PCA-ANN		Predicted cancer	Predicted noncancer	Sensitivity	Specificity	Accuracy
Training	Actual cancer	6	11	35.29%	98.36%	97.01%
	Actual noncancer	39	2332			
Testing	Actual cancer	3	5	37.50%	98.33%	97.85 %
	Actual noncancer	17	998			

TABLE 6. The performance of DBN-ELM-BP.

DBN-ELM-BP		Predicted cancer	Predicted noncancer	Sensitivity	Specificity	Accuracy
Training	Actual cancer	10	7	58.83%	98.31%	98.03%
	Actual noncancer	40	2331			
Testing	Actual cancer	5	3	62.50%	98.52%	98.24%
	Actual noncancer	15	1000			

people in the population. Second, the two models in this study achieved comparable or even better sensitivity and AUC compared with existing methods using questionnaire or survey data. Compared with the results of a previous work [26], the sensitivity and specificity of both models proposed in this study are better. The sensitivity of the DBN-ELM-BP model is higher than some previous work [26], [27] and comparable to others [21], [22], [28]. Furthermore, both models are used here to predict cancer probability of patients based only on the screening questionnaire. Although two prediction models give lower sensitivity compared to some previous work [15], [16], the cost of collecting similar data for filling out the screening questionnaire of chronic diseases is also very low. As the risk of common cancers can be predicted through the questionnaire, it is suitable for large-scale common cancers screening. It will further benefit individual participants as an effective but inexpensive way to carry out early screening for common cancers.

There are some limitations of this study, though. First, the classification problem is tested using imbalanced data: the sample size is large but the total number of positive cases is very small (25 cancer patients). The percentage of negative instances of the training set and testing set is close to 100% while the percentage of positive ones is around 15% due to the low prevalence of common cancers in the population. This is also the reason why we use sensitivity and AUC to measure the performance of predictive models. A small number of positive cases may lead to biased training. Thus, learning of these cases' information by two models will affect the performance of predictions. This is why the sensitivity of the two prediction models is not high. Second, all cases occurred during the screening period, so there is no follow-up data to verify the effect of the model. Therefore, it is necessary to observe external validity and be cautious in deduction.

VI. CONCLUSION

This study has developed PCA-ANN and DBN-ELM-BP prediction models. The training set of PCA-ANN provides 35.29% sensitivity, 98.36% specificity, 97.01% accuracy, and AUC 0.7245, the testing set of PCA-ANN provides 37.50% sensitivity, 98.33% specificity, 97.85% accuracy, and AUC 0.7221; while the training set of DBN-ELM-BP gives 58.83% sensitivity, 98.31% specificity, 98.03% accuracy, and AUC measuring 0.7747 respectively, the testing set of DBN-ELM-BP gives 62.50% sensitivity, 98.52% specificity, 98.24% accuracy, and AUC measuring 0.7238 respectively. Based on the information in the screening questionnaire of chronic diseases, this study can predict and classify the cancers probability of patients. The methods in this study are easy to implement, non-invasive and economical. At the same time, they have considerable specificity and accuracy compared with other methods which usually need patients' biopsy, imaging and genomic data. They have high diagnostic value and are suitable for the primary screening of five common cancers among people aged 20∼80 in China. For better clinical applications, our algorithms need more data and tests to further improve the performance of common cancers prediction models.

Moving forward, the machine learning prediction methods developed will help doctors determine which patients have a high risk of cancers and then transfer them to the clinic for pathological examination. Also, the authors will attempt to find and handle answers that are casual or not under conscientious consideration in the questionnaire to improve the sensitivity of the prediction models because those data can obscure facts to produce noise or errors which adversely affect experimental results. With the screening questionnaires of patients, the model could easily be implemented on a mobile device. These methods will allow clinicians to immediately have

access to patient condition during data entry stage, therefore, help to achieve common cancers screening and prevention.

ACKNOWLEDGMENT

(Huitao Qi and Yanli Chen are co-first authors.)

AUTHOR'S CONTRIBUTIONS

Conception and design: Huitao Qi, Shuangbo Xie, Yanli Chen, Mingxu Sun. Collection and assembly of data: Shuangbo Xie, Yanli Chen, Chengqian Wang, Tingting Wang, Mingxu Sun. Data analysis and interpretation: Huitao Qi, Shuangbo Xie, Chengqian Wang, Tingting Wang, Bin Sun, Mingxu Sun. Final approval of manuscript: All authors. Accountable for all aspects of the work: All authors.

CONFLICTS OF INTERESTS

The authors declare that they have no conflicts of interest to report regarding the present study.

REFERENCES

[1] F. Bray, J. Ferlay, I. Soerjomataram, R. L. Siegel, L. A. Torre, and A. Jemal, "Global cancer statistics 2018: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries," *CA, Cancer J. Clin.*, vol. 68, no. 6, pp. 394–424, 2018, doi: 10.3322/caac.21492.

[2] M. Madmoli, M. Yarbig, N. Sedighi, and P. Darabiyan, "Communication between body mass index and the risk of obesity-related cancer: A 5-year study on patients with cancer," *Med. Sci.*, vol. 23, no. 9, pp. 69–74, 2019.

[3] J. W. Chung, J. J. Park, Y. J. Lim, and J. Lee, "Gastrointestinal cancer risk in patients with a family history of gastrointestinal cancer," *Korean J. Gastroenterol*, vol. 71, no. 6, pp. 338–348, 2018, doi: 10.4166/kjg.2018.71.6.338.

[4] D. Yang, Y. Liu, C. Bai, X. Wang, and C. A. Powell, "Epidemiology of lung cancer and lung cancer screening programs in China and the United States," *Cancer Lett.*, vol. 468, pp. 82–87, Jan. 2020, doi: 10.1016/j.canlet.2019.10.009.

[5] K. C. Thandra, A. Barsouk, K. Saginala, J. S. Aluru, and A. Barsouk, "Epidemiology of lung cancer," *Współczesna Onkologia*, vol. 25, no. 1, pp. 45–52, 2021, doi: 10.5114/wo.2021.103829.

[6] X. Liu, A. Baecker, M. Wu, and J. Y. Zhou, "Family history of liver cancer may modify the association between HBV infection and liver cancer in a Chinese population," *Liver Int.*, vol. 39, no. 8, pp. 1490–1503, Aug. 2019, doi: 10.1111/liv.14182.

[7] H. J. Youn and W. Han, "A review of the epidemiology of breast cancer in Asia: Focus on risk factors," *Asian Pacific J. Cancer Prevention*, vol. 21, no. 4, pp. 867–880, Apr. 2020, doi: 10.31557/APJCP.2020.21.4.867.

[8] O. Y. Bang, J.-W. Chung, M. J. Lee, W.-K. Seo, G.-M. Kim, and M.-J. Ahn, "Cancer-related stroke: An emerging subtype of ischemic stroke with unique pathomechanisms," *J. Stroke*, vol. 22, no. 1, pp. 1–10, Jan. 2020, doi: 10.5853/jos.2019.02278.

[9] X. Xu, Z. Fang, J. Zhang, Q. He, D. Yu, L. Qi, and W. Dou, "Edge content caching with deep spatiotemporal residual network for IoV in smart city," *ACM Trans. Sensor Netw.*, vol. 17, no. 3, pp. 1–33, Jun. 2021, doi: 10.1145/3447032.

[10] X. Xu, H. Li, W. Xu, Z. Liu, L. Yao, and F. Dai, "Artificial intelligence for edge service optimization in Internet of vehicles: A survey," *Tsinghua Sci. Technol.*, vol. 27, no. 2, pp. 270–287, Apr. 2022, doi: 10.26599/TST.2020.9010025.

[11] X. Xu, H. Tian, X. Zhang, L. Qi, Q. He, and W. Dou, "DisCOV: Distributed COVID-19 detection on X-ray images with edge-cloud collaboration," *IEEE Trans. Services Comput.*, vol. 15, no. 3, pp. 1206–1219, May 2022, doi: 10.1109/TSC.2022.3142265.

[12] X. Xu, Z. Fang, L. Qi, X. Zhang, Q. He, and X. Zhou, "TripRes: Traffic flow prediction driven resource reservation for multimedia IoV with edge computing," *ACM Trans. Multimedia Comput., Commun., Appl.*, vol. 17, no. 2, pp. 1–21, May 2021, doi: 10.1145/3401979.

[13] B. Shen, X. Xu, L. Qi, X. Zhang, and G. Srivastava, "Dynamic server placement in edge computing toward internet of vehicles," *Comput. Commun.*, vol. 178, pp. 114–123, Oct. 2021, doi: 10.1016/j.comcom.2021.07.021.

[14] Z. C. Hu, X. L. Xu, Y. L. Zhang, and H. S. Tang, "Cloud-edge cooperation for meteorological radar big data: A review of data quality control," *Complex Intell. Syst.*, vol. 8, pp. 3789–3803, Nov. 2021, doi: 10.1007/s40747-021-00581-w.

[15] A. R. Ruiz, E. Krupinski, J. J. Mordang, and K. Schilling, "Detection of breast cancer with mammography: Effect of an artificial intelligence support system," *Int. J. Med. Radiol.*, vol. 42, no. 2, pp. 235–244, 2019, doi: 10.1148/radiol.2018181371.

[16] A. Hekler, J. S. Utikal, A. H. Enk, A. Hauschild, M. Weichenthal, R. C. Maron, C. Berking, S. Haferkamp, J. Klode, D. Schadendorf, B. Schilling, T. Holland-Letz, B. Izar, C. Von Kalle, S. Fröhling, and T. J. Brinker, "Superior skin cancer classification by the combination of human and artificial intelligence," *Eur. J. Cancer*, vol. 120, pp. 114–121, Oct. 2019, doi: 10.1016/j.ejca.2019.07.019.

[17] Y. Horie, T. Yoshio, K. Aoyama, S. Yoshimizu, Y. Horiuchi, A. Ishiyama, T. Hirasawa, T. Tsuchida, T. Ozawa, S. Ishihara, Y. Kumagai, M. Fujishiro, I. Maetani, J. Fujisaki, and T. Tada, "Diagnostic outcomes of esophageal cancer by artificial intelligence using convolutional neural networks," *Gastrointestinal Endoscopy*, vol. 89, no. 1, pp. 25–32, 2019, doi: 10.1016/j.gie.2018.07.037.

[18] L. Á. Menéndez, F. J. de Cos Juez, F. S. Lasheras, and J. A. Á. Riesgo, "Artificial neural networks applied to cancer detection in a breast screening programme," *Math. Comput. Model.*, vol. 52, nos. 7–8, pp. 983–991, Oct. 2010, doi: 10.1016/j.mcm.2010.03.019.

[19] G. Perez and P. Arbelaez, "Automated lung cancer diagnosis using three-dimensional convolutional neural networks," *Med. Biol. Eng. Comput.*, vol. 58, no. 8, pp. 1803–1815, Aug. 2020, doi: 10.1007/s11517-020-02197-7.

[20] S. T. Chandrasekaran, R. Hua, I. Banerjee, and A. Sanyal, "A fully-integrated analog machine learning classifier for breast cancer classification," *Electronics*, vol. 9, no. 3, p. 515, Mar. 2020, doi: 10.3390/electronics9030515.

[21] B. J. Nartowt, G. R. Hart, W. Muhammad, Y. Liang, G. F. Stark, and J. Deng, "Robust machine learning for colorectal cancer risk prediction and stratification," *Frontiers Big Data*, vol. 3, pp. 6–18, Mar. 2020, doi: 10.3389/fdata.2020.00006.

[22] G. R. Hart, B. J. Nartowt, W. Muhammad, Y. Liang, G. S. Huang, and J. Deng, "Stratifying ovarian cancer risk using personal health data," *Frontiers Big Data*, vol. 2, pp. 3–12, Jul. 2019, doi: 10.3389/fdata.2019.00024.

[23] G. R. Hart, D. A. Roffman, R. Decker, and J. Deng, "A multi-parameterized artificial neural network for lung cancer risk prediction," *PLoS ONE*, vol. 13, no. 10, Oct. 2018, Art. no. e0205264, doi: 10.1371/journal.pone.0205264.

[24] W. Muhammad, G. R. Hart, B. Nartowt, J. J. Farrell, K. Johung, Y. Liang, and J. Deng, "Pancreatic cancer prediction through an artificial neural network," *Frontiers Artif. Intell.*, vol. 2, pp. 2–11, May 2019, doi: 10.3389/frai.2019.00002.

[25] D. Roffman, G. Hart, M. Girardi, C. J. Ko, and J. Deng, "Predicting non-melanoma skin cancer via a multi-parameterized artificial neural network," *Sci. Rep.*, vol. 8, no. 1, pp. 1701–1707, Dec. 2018, doi: 10.1038/s41598-018-19907-9.

[26] D. A. Roffman, G. R. Hart, M. S. Leapman, J. B. Yu, F. L. Guo, I. Ali, and J. Deng, "Development and validation of a multiparameterized artificial neural network for prostate cancer risk prediction and stratification," *JCO Clin. Cancer Informat.*, vol. 2, pp. 1–10, Dec. 2018, doi: 10.1200/CCI.17.00119.

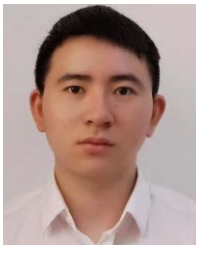
[27] B. J. Nartowt, G. R. Hart, D. A. Roffman, X. Llor, I. Ali, W. Muhammad, Y. Liang, and J. Deng, "Scoring colorectal cancer risk with an artificial neural network based on self-reportable personal health data," *PLoS ONE*, vol. 14, no. 8, Aug. 2019, Art. no. e0221421, doi: 10.1371/journal.pone.0221421.

[28] G. F. Stark, G. R. Hart, B. J. Nartowt, and J. Deng, "Predicting breast cancer risk using personal health data and machine learning models," *PLoS ONE*, vol. 14, no. 12, Dec. 2019, Art. no. e0226765, doi: 10.1371/journal.pone.0226765.

[29] L. F. Xiao, Y. L. Chen, Y. X. Xing, and L. N. Mou, "The analysis and AI prospect based on the clinical screening results of chronic diseases," in *Proc. 11th Int. Conf. Comput. Eng. Netw. (CENet)*, 2021, pp. 563–572, doi: 10.26914/c.cnkihy.2021.045057.

[30] M. Bosch-Baliarda, O. Soler Vilageliu, and P. Orero, "Toward a sign language-friendly questionnaire design," *J. Deaf Stud. Deaf Educ.*, vol. 24, no. 4, pp. 333–345, Oct. 2019, doi: 10.1093/deafed/enz021.
[31] P. R. Regmi, E. Waithaka, A. Paudyal, P. Simkhada, and E. Van Teijlingen, "Guide to the design and application of online questionnaire surveys," *Nepal J. Epidemiol.*, vol. 6, no. 4, pp. 640–644, May 2017, doi: 10.3126/nje.v6i4.17258.
[32] V. Toepoel, B. Vermeeren, and B. Metin, "Smileys, stars, hearts, buttons, tiles or grids: Influence of response format on substantive response, questionnaire experience and response time," *Bull. Sociol. Methodol./Bull. de Méthodolog. Sociologique*, vol. 142, no. 1, pp. 57–74, Apr. 2019, doi: 10.1177/0759106319834665.
[33] S. Yaddanapudi and L. N. Yaddanapudi, "How to design a questionnaire," *Indian J. Anaesthesia*, vol. 63, no. 5, pp. 335–337, 2019, doi: 10.4103/ija.IJA_334_19.
[34] W. G. Madow, H. Nisselson, and I. Olkin, *Imcomplete Data in Sample Surveys*, vol. 5, no. 6. New York, NY, USA, 2012.
[35] D. J. Kim, B. Rockhill, and G. A. Colditz, "Validation of the Harvard cancer risk index: A prediction tool for individual cancer risk," *J. Clin. Epidemiol.*, vol. 57, no. 4, pp. 332–340, 2012, doi: 10.1016/j.jclinepi.2012.08.013.
[36] G. A. Colditz, K. A. Atwood, K. Emmons, and R. R. Monson, "Harvard report on cancer prevention, volume 4: Harvard cancer risk index. Risk index working group, Harvard center for cancer prevention," *Cancer Causes Control, CCC*, vol. 11, no. 16, pp. 477–478, 2000, doi: 10.1023/A:1008984432272.
[37] G. E. Hinton, S. Osindero, and Y.-W. Teh, "A fast learning algorithm for deep belief nets," *Neural Comput.*, vol. 18, no. 7, pp. 1527–1554, Aug. 2006, doi: 10.1162/neco.2006.18.7.1527.
[38] X.-D. Chen, Y. Hai-Yue, J.-S. Wun, C.-H. Wu, C.-H. Wang, and L.-L. Li, "Power load forecasting in energy system based on improved extreme learning machine," *Energy Explor. Exploitation*, vol. 38, no. 4, pp. 1194–1211, Jul. 2020, doi: 10.1177/0144598720903797.
[39] M. A. Salama, A. E. Hassanien, and A. A. Fahmy, "Deep belief network for clustering and classification of a continuous data," in *Proc. 10th IEEE Int. Symp. Signal Process. Inf. Technol.*, Dec. 2010, pp. 473–477, doi: 10.1109/ISSPIT.2010.5711759.
[40] D. Erhan, P. A. Manzagol, Y. Bengio, and S. Bengio, "The difficulty of training deep architectures and the effect of unsupervised pretraining," in *Proc. 12th Int. Conf. Artif. Intell. Statist.*, vol. 5, 2009, pp. 153–160.

HUITAO QI was born in Jinan, Shandong, in 1967. She received the bachelor's degree in acupuncture from the Shandong University of Traditional Chinese Medicine, in 1990. From July 1990 to January 2003, she worked with the Department of Acupuncture and Moxibustion, Jinan Central Hospital in Shandong Province. From February 2003 to August 2003, she studied rehabilitation therapy technology with the China Rehabilitation Research Center. Since September 2003, she has been working with the Department of Rehabilitation Medicine of Jinan Central Hospital.



SHUANGBO XIE was born in Yongzhou, China, in 1998. He received the B.S. degree in electronic engineering and automation from the Civil Aviation University of China, in 2019. He is currently pursuing the master's degree majoring in control science and engineering with the University of Jinan. His research interests include using artificial intelligence technology to predict the risk of cancer and processing in health informatics.

YANLI CHEN received the B.S. degree in nursing. She is currently the Head of the Admission Preparation Center of the Affiliated Central Hospital of Shandong First Medical University. She is also the Deputy Director of the First Rheumatism and Immune Disease Nursing Professional Committee of Shandong Nursing Association; Shandong Anti-Cancer Association Tumor Epidemiology Member of the First Committee of the Academic Club; and a member of the First Committee of the Disease and Health Management Professional Committee of Shandong Hospital Association. She is in charge of nurses. Her research interests include nursing management and blood rheumatism nursing.

CHENGQIAN WANG was born in Dezhou, Shandong, China, in 1996. She received the B.S. degree in computer science and technology from Jining University. She is currently pursuing the M.S. degree with Jinan University. She is the author of two articles and three inventions. Her research interests include deep learning, human action recognition, and multi-sensor fusion.

TINGTING WANG was born in Shandong, China, in 1996. She received the B.S. degree in communication engineering from Ludong University, China, in 2020. She is currently pursuing the master's degree in control science and engineering with the University of Jinan. Her research interests include adaptive control of rehabilitation bicycles and electrical stimulator rehabilitation.

BIN SUN received the B.S. degree from Shandong University, in 2010, and the Ph.D. degree from the Blekinge Institute of Technology, in 2018. He has long been engaged in theoretical research and practice related to big data intelligent analysis and prediction of the Internet of Things.

MINGXU SUN was born in Jinan, China, in 1984. He received the B.S. degree in control engineering from the University of Jinan, in 2007, and the M.S. degree in manufacturing engineering and the Ph.D. degree in medical engineering from the University of Salford, Manchester, U.K., in 2014.

Following periods of postdoctoral work with the Rehabilitation Technologies & Biomedical Engineering Group, University of Salford, he was appointed to an Assistant Professor of rehabilitation technologies with the School of Electrical Engineering, University of Jinan, in 2018. His research interests include development of functional electrical stimulation systems for use in stroke rehabilitation, together with novel approaches of using inertial sensors for their control; and use of inertial sensors to understand the real world use of walking aids.