D-TSVR Recurrence Prediction Driven by Medical Big Data in Cancer

Ai-Min Yang, Yang Han*, Chen-Shuai Liu, Jian-Hui Wu, Dian-Bo Hua

***Abstract*—Secondary use of medical big data is becoming increasingly popular in healthcare services and clinical research in medical industry. Cancer recurrence is a common phenomenon of cancer patients after treatment (recovery period). Studying the time and influencing factors of cancer recurrence can provide effective clinical intervention means, which is the gospel of cancer patients. In this paper, a sample of 50,000 cases of seven cancer patients, including liver cancer, lung cancer, kidney cancer, breast cancer, uterine cancer, stomach cancer, and bowel cancer, was collected. A TSVR algorithm based on DNN (Dependent Nearest Neighbor) weighting is proposed, the eplion-TSVR model is improved by DNN weighted algorithm with local information mining function, and the solution of the improved model is derived. It is proposed to use the cuckoo algorithm to determine the optimal parameters of DNN to determine the optimal DR domain. In this paper, the improved TSVR algorithm is used to establish a cancer recurrence prediction model. The prediction accuracy of the model for various cancers can reach more than 91%, which is significantly higher than that of CNN and e-TSVR models.**

***Index Terms*—Cancer; Cancer Recurrence; SVR; Cuckoo Algorithm; Data Mining**

I. Introduction

With the continuous development of human society and the continuous improvement of productivity, since the 21st century, information technology has become the focus of the development of the times. The application of information technology in various industries has made the information industry develop rapidly. A large number of social resources are invested in the information industry, which provides a rich environment for its technological development. The continuous development of information technology is not only for the emerging industries Development plays an absolute role in promoting. And It brings new development opportunities for traditional industries [1]. Among them, under the combination of traditional medical institutions with information data storage, information transmission management and other technologies, the utilization efficiency of big data in the field of medical and health has been greatly improved. The application of information technology in daily work makes the development of medical industry enter an information mode. Data mining is widely used in the medical industry.

Data mining is a computer technology that clusters, categorizes and predicts large amounts of data. The support vector machine has good robustness and performs better performance [1-2] in linear indivisible problems. Dsouza K J [3] used the SVM algorithm to classify and validate The Breast Cancer Wisconsin Data Sets in 2018, showing better robustness. Hasan MR used Support Vector Machine (SVM), Random Forest Tree (RFT) and Naïve Bayes Classifier (NBC) algorithms to build a machine-learning model based on cancer cell gestures, all of which were more than 82% accurate and had good predictive performance [4]. Yan K [5] proposed An Extended Genetic Algorithm Based Gene Selection Framework for Cancer Diagnosis, which has the highest classification accuracy compared with existing methods. The cuckoo algorithm performs better than the genetic algorithm in some respects. Therefore, this paper uses the cuckoo algorithm to improve the support vector regression, which is more appropriate for the model parameters.

In 1971, the United States first proposed the concept of tumor rehabilitation in the National Cancer Program [6]. Cromes defines it as "helping cancer patients to maximize their physical, social, psychological and occupational functions under the conditions of the cancer disease itself and the limitations of cancer treatment" [7]. Restoring patients' psychological, physiological and physical functions is the main purpose of tumor rehabilitation, and all the means can be taken are the content of rehabilitation. Depending on the nature of the tumor and the stage of the patient's disease, the goal of tumor rehabilitation is also focused on, including physical function recovery, psychology, nutrition, exercise, cancer pain recovery, and other aspects.

In the past few years, a lot of efforts have been made to detect tumor cells and early cancer diagnosis. Early detection and treatment of cancer has a high survival rate and greatly improves the quality of life, but the influencing factors of postoperative recurrence of cancer patients are very complex, so it is very challenging to predict the recurrence time. The postoperative recurrence time of patients affects the patients'

This work was supported in part by the Hebei Province Natural Science Fund for Excellent Young Scholars E2018209248, in part by the National Natural Science Foundation of China under Grant 51674121, in part by the innovation projects of Ph.D. degree students in Hebe CXZZBS2020134.

* Corresponding author is addressed to Yang Han (hanyang@ncst.edu.cn)

Ai-Min Yang is with the College of Science, North China University of Science and Technology, Tangshan, Hebei, China (aimin@ncst.edu.cn).

Yang Han is with the College of Metallurgy and Energy, North China University of Science and Technology, Tangshan, Hebei, China. (hanyang@ncst.edu.cn).

Chen-Shuai Liu is with the College of Science, North China University of Science and Technology, Tangshan, Hebei, China (liucmys@163.com).

Jian-Hui Wu is with the School of Public Health, North China University of Science and Technology, Tangshan, Hebei, China. (wujianhui555@163.com).

Dian-Bo Hua is with Beijing Street Laboratory, Beijing 100000, Chin. (78225474@qq.com).

quality of life and living standards, and the prognostic intervention of patients to study the postoperative relapse time of patients has a crucial impact. Personalized intervention for cancer patients with short recurrence time, prolonging the survival time and quality of life of patients, has made certain contributions to the field of cancer rehabilitation.

Considering that the condition of cancer patients is closely related to the recurrence of cancer, if the condition of cancer patients can be comprehensively evaluated, it is possible to predict the recurrence of cancer. This paper mainly does two aspects of work. On the one hand, the patients were evaluated from seven aspects: basic indicators, immune indicators, tumor indicators, nutrition indicators, psychological indicators, microenvironment indicators, sports and learning work. On the other hand, an improved TSVR model is proposed to predict tumor recurrence time, which has a good prediction effect.

In this paper, we collected the data of 50000 patients with seven kinds of cancer, including liver cancer, lung cancer, kidney cancer, breast cancer, uterine cancer, gastric cancer and colorectal cancer. Firstly, the health status of each patient was evaluated, and the scores of each index were obtained. An improved TSVR model was established to predict the recurrence of various cancers. At the same time, compared with CNN and e-TSVR prediction model, the conclusion is that the accuracy of this model is high.

II. Related Work

A. Cancer Recurrence

In recent years, cancer has become the leading cause of premature death, which seriously threatens human health and has received widespread concern from the world's medical institutions. Jose et al. explored the expression of apoptotic markers in radical cystectomy and bilateral lymphadenectomy and their relationship with tumor prognosis, and found that apoptotic markers have synergistic effect on the progression of bladder cancer by means of multivariate variance [8]. Bartels p H [9] et al. Analyzed the nuclear image of recurrent or non recurrent tumor cases, and predicted the change of cancer recurrence risk through discriminant analysis and unsupervised learning algorithm, with good prediction effect. The TIES.IO Cancer Data Analysis Laboratory, in conjunction with this research team, evaluated and tracked the rehabilitation process of 50000 of cancer patients in China, mined and analyzed their clinical and rehabilitation data, and then established a model for predicting cancer recurrence suitable for Chinese patients. By collecting historical recovery data for cancer patients such as liver cancer, lung cancer, kidney cancer, breast cancer, uterine cancer, stomach cancer, and bowel cancer, there was no significant difference between the recovery of cancer patients and their physical condition, daily activity, dietary intake, environmental conditions and psychological factors, and gender. The rehabilitation nutrition program for cancer patients mainly involves 10 indications: basic indication, tumor indication, immune indication, basic nutrition indication, nutrition contrast indication, safe intake indication, nutrition indication (balanced nutrition indication), microenvironment indication, psychological indication, aerobic activity (and advanced operation) indication [10]. By mining the historical large data sample set of cancer patients, the sample set of the processed cancer recurrence data can be used to predict the recurrence time of cancer patients through machine learning, as shown in **Fig.1**.

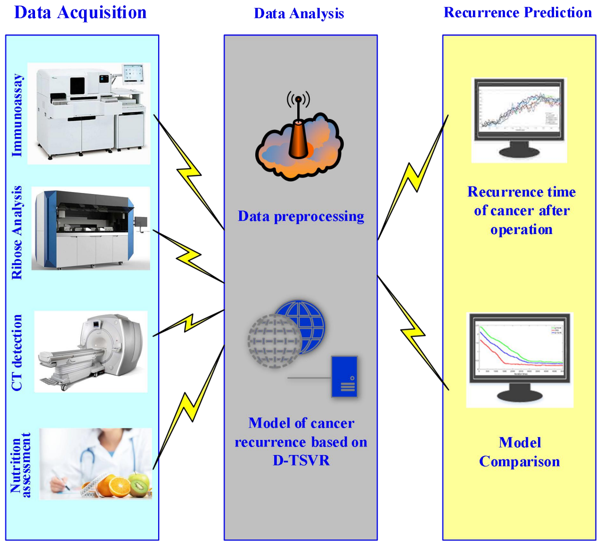


Fig. 1 Cancer recurrence model based on Data Mining

B. Intelligent Algorithm

As early as 2004, Stanford University's medical and mathematical interdisciplinary research on intelligent assessment of cancer patient rehabilitation has been established, and it has shown its huge application prospects in clinical applications targeting hundreds of thousands of American cancer patients. According to the historical data of cancer rehabilitation, based on the data-driven modeling method. Cross-integrated the historical data of cancer rehabilitation with intelligent algorithms, and established a cancer recurrence prediction model [11]. Richard et al. evaluate the risk of recurrence of breast cancer patients and divide them into three groups: low risk, moderate risk and high risk. The CNN is composed of four convolutional layers and the largest pooling layer to make predictions [12]. The rate reaches 90%, indicating that the current deep CNN architecture is trained to predict cancer recurrence. Kumar N and others proposed a prediction model of prostate cancer recurrence based on two independent convolutional neural networks in 2017, and the prediction effect is better [13]. Yunfei H et al. established a 15-miRNA-based SVM classifier to predict the recurrence of osteosarcoma, and its prediction accuracy reached more than 85% [14]. This model provides a potential tool for the prediction of osteosarcoma. Jinting et al. used the gene expression data of the retrieved ovarian cancer (OC) samples to identify the target genes and established a support vector machine (SVM) classifier to predict the recurrence of ovarian cancer, with a prediction accuracy rate of more than 90% [15]. Common cancer prediction model algorithms mainly include convolutional neural network (CNN) and support vector machine (SVM). CNN simulates biological neural network structurally, which has strong learning ability. However, CNN modeling needs a large number of samples. The pooling layer will lose a large number of valuable information, and ignore the local and overall relevance and other shortcomings [16]. SVM algorithm can solve the problem of high dimension with perfect theoretical basis, but only when the number of samples is small,

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TII.2020.3011675, IEEE Transactions on Industrial Informatics

the prediction effect is good, the learning efficiency is relatively low, and the prediction performance is relatively poor [17].

In view of the shortcomings of the above algorithm, this paper proposes a DNN-weighted TSVR algorithm based on DNN, improves the eplion-TSVR model with a DNN-weighted algorithm with local information mining function, and deduces the solution of the improved model, aiming against the DNN algorithm DR domain problem. It is proposed to use the cuckoo algorithm to determine the optimal parameters of DNN and then to determine the optimal DR domain, and finally achieve accurate prediction of cancer recurrence (entropy state) is achieved.

III. Cancer Recurrence Prediction Model Based on Improved TSVR Algorithm

With the development of computer technology, the research of clinical diagnosis and treatment data can be further developed. In recent years, several complex models represented by neural networks, support vector machines and combinatorial lifting algorithms have been widely used. Researchers can find sophisticated methods to predict the risk of disease. Support vector regression machine (SVR) is a branch of support vector machine in the field of regression, twin support regression vector machine (TSVR) is a larger model for regression model improvement, and it greatly reduces the computational complexity of the model by changing the constraints.

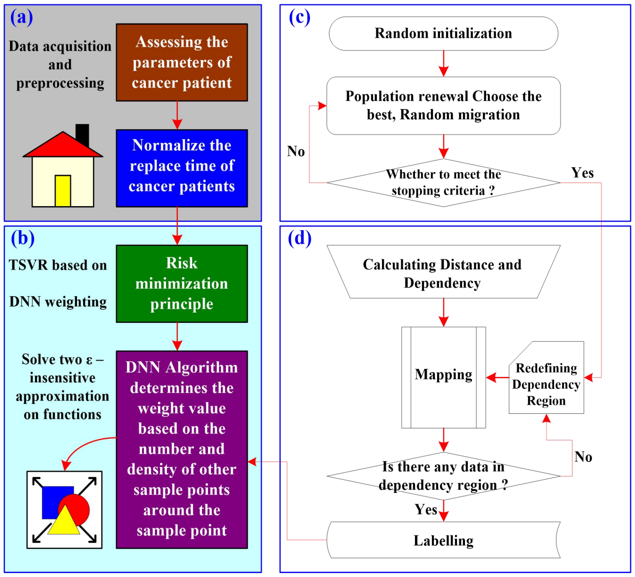

(a)
Data acquisition and preprocessing
Assessing the parameters of cancer patient
Normalize the replace time of cancer patients
(b)
TSVR based on DNN weighting
Risk minimization principle
Solve two ε – insensitive approximation on functions
DNN Algorithm determines the weight value based on the number and density of other sample points around the sample point
(c)
Random initialization
Population renewal Choose the best, Random migration
No
Whether to meet the stopping criteria ?
Yes
(d)
Calculating Distance and Dependency
Mapping
Redefining Dependency Region
No
Is there any data in dependency region ?
Yes
Labelling

Fig.2 TSVR algorithm based on DNN weighting

The addition of the regular item based on the TSVR model improves the generalization performance of the model. The e-TSVR model, by Yuan Shao-Hai et al., introduced the principle of structural risk minimization based on TSVR. It used two small QPP questions to determine a pair of ε insensitive approximation functions [18]. Because the introduction of the principle of structural minimization makes the pair problem positive, it improves the regression prediction performance. Comparing the classic SVR model, the computational complexity is reduced because the problem of two small-scale QPP is solved.

In this paper, a TSVR algorithm based on DNN algorithm weighting is proposed to predict the recurrence time of cancer patients, **Fig.2** is a diagram of the algorithm model. Part (a) represents the evaluation of cancer patients and collects data. The TSVR algorithm, part (c) represents the cuckoo algorithm used to solve the optimal DR domain, and part (d) represents the principle of the DNN algorithm.

A. DNN Algorithm and Ample Weighting Based on DNN

Ömer Faruk Ertuğrul et al. [19] proposed the Dependent nearest neighbor (DNN). The target sample in the DNN is mapped to the center point of the two-dimensional vector plane. All other data in the sample set will be mapped to this plane. The mapping relationship is determined by the connection function p, and the distance and similarity between the target sample and other samples need to be calculated Sex. The mapping function is.

$$p_i(x,y)=p_i(d_i\cos(\theta_i),d_i\sin(\theta_i)) \quad (1)$$

Where x and y represent the coordinate axis of a two-dimensional vector plane, respectively. And represents the position of the sample on the plane. The more similar the sample in the training set to the target sample is, the closer it is to the center, and it is distributed positively towards the axis. Such a mapping allows the model to have better accuracy and less computation time [20].

The DNN algorithm needs to calculate two indexes of similarity and dependency. And on here, Euclidean distance is used as the distance calculation method to measure the similarity of samples. The dependence is given by calculating the projection angle between the samples. The magnitude of the projection angle is related to the correlation between the two vectors, which also be called the dependence. The two samples are said to be interdependent when the two samples have the same angles, and the smaller the angle θ, the dependency between the samples p and q the higher.

This study proposes to improve the DNN algorithm and introduce a new weighting method. Suppose there are N sample points $x_1, x_2, \cdots, x_N$ in the sample space X. The DNN method is used to describe the local structure of the sample space. The DNN weight matrix is defined as:

$$W_{i,j}=\begin{cases}1, & x_i,x_j\in(\mathrm{D}-\varepsilon,\mathrm{D}+\varepsilon)\\ 0, & \text{other}\end{cases} \quad (2)$$

$$D=diag(d_1,d_2,\cdots,d_N),$$

$$d_i=\sum_{j=1}^{N}W_{i.j}(i=1,2,\cdots,N) \quad (3)$$

Among them, $W_{i,j}$ is a symmetric matrix, and find the sum of its elements in each row or column as d_i. The value d_i obtained by $W_{i,j}$ represents the D nearest neighbor number of the i sample point. By introducing a weighted diagonal matrix gives the weight coefficient of each sample point in the sample space. The weight value represents how dense the sample is located. In order to obtain the optimal weight matrix, this study uses cuckoo algorithm to find parameters to determine the optimal dependency region (DR) domain.

B. Linear e-TSVR model

E -TSVR determines the final regression function by finding two insensitive approximation functions. Empirical risk is measured by the following formula.

$$R_{emp}^{\varepsilon_1}[f_1] = \sum_{i=1}^{l} \max\left\{0, (y_i - f_1(x_i))^2\right\} + c_1 \sum_{i=1}^{l} \max\{0, -(y_i - f_1(x_i) + \varepsilon_1)\} \tag{4}$$

$$R_{emp}^{\varepsilon_2}[f_2] = \sum_{i=1}^{l} \max\left\{0, (y_i - f_2(x_i))^2\right\} + c_2 \sum_{i=1}^{l} \max\{0, -(y_i - f_2(x_i) + \varepsilon_2)\} \tag{5}$$

Among them, parameters c_1, $c_2 > 0$, and the empirical risk function consists of two parts. The second term of formula 4 represents the marginal ε insensitive loss function of f_1. The same as the second term of formula 5 represents the marginal ε insensitive loss function of f_2. By introducing regular terms, and relaxation variables ξ and η, the original problem formula is expressed as follows:

$$\begin{gathered} \min_{w_1, b_1, \xi} \frac{1}{2} c_1 \left(w_1^T w_1 + b_1^2\right) + \frac{1}{2}\left(Y - (Aw_1 + eb_1)\right) \\ \left(Y - (Aw_1 + eb_1)\right) + c_1 e^T \xi \\ s.t.\ Y - (Aw_1 + eb_1) \ge -\varepsilon e - \xi, \xi \ge 0 \end{gathered} \tag{6}$$

$$\begin{gathered} \min_{w_2, b_2, \xi} \frac{1}{2} c_2 \left(w_2^T w_2 + b_2^2\right) + \frac{1}{2}\left(Y - (Aw_2 + eb_2)\right) \\ \left(Y - (Aw_2 + eb_2)\right) + c_2 e^T \eta \\ s.t.\ Y - (Aw_2 + eb_2) \ge -\varepsilon e - \eta, \eta \ge 0 \end{gathered} \tag{7}$$

Where c_1, c_2, ε_1 and ε_2 are parameter greater than zero, e is the unit vector. The geometric visual representation of the model is shown in **Fig. 3**. **Fig. 3** shows the data fitting conditions of the two approximate functions of the dual support vector regression machine proposed in [15] and [18] respectively.

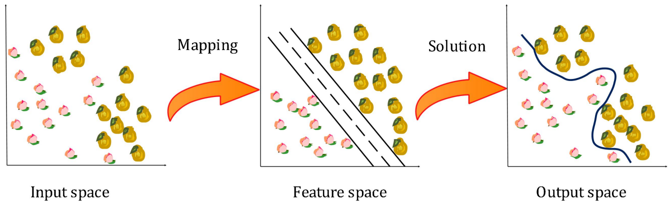


Fig.3 Data fitting

In order to solve model formula (6) and model formula (7), we need to solve their dual problems. According to Lagrange function and KKT condition, this problem can be solved.

C. Non-linear e-TSVR model

The above model is a linear model, and the probability of a non-linear relationship in the complex cancer patient evaluation data is obviously greater than the probability of a linear relationship, so the non-linear model is described below. The non-linear model is all the same with the linear model except introducing the kernel function. Here is a rough introduction to the non-linear model is introduced considering the length of the problem.

Nonlinear two ε-insensitive approximation functions are:

$$\begin{gathered} f_1(x) = K\left(x^T, A^T\right)w_1 + b_1 \\ f_2(x) = K\left(x^T, A^T\right)w_2 + b_2 \end{gathered} \tag{8}$$

Where K is the selected kernel function. And the non-linear original problem is derived as:

$$\begin{gathered} \min_{w_1, b_1, \xi} \frac{1}{2} c_3 \left(w_1^T w_1 + b_1^2\right) + c_1 e^T \xi + \\ \frac{1}{2}\left(Y - \left(K\left(A, A^T\right)w_1 + eb_1\right)\right)^T \cdot \\ \left(Y - \left(K\left(A, A^T\right)w_1 + eb_1\right)\right) \\ s.t.\ Y - \left(K\left(A, A^T\right)w_1 + eb_1\right) \ge -\varepsilon e - \xi, \xi \ge 0 \end{gathered} \tag{9}$$

$$\begin{gathered} \min_{w_2, b_2, \xi} \frac{1}{2} c_4 \left(w_2^T w_2 + b_2^2\right) + c_2 e^T \eta + \\ \frac{1}{2}\left(Y - \left(K\left(A, A^T\right)w_2 + eb_2\right)\right)^T \cdot \\ \left(Y - \left(K\left(A, A^T\right)w_2 + eb_2\right)\right) \\ s.t.\ Y - \left(K\left(A, A^T\right)w_2 + eb_2\right) \ge -\varepsilon e - \eta, \eta \ge 0 \end{gathered} \tag{10}$$

Where c_1, c_2, c_3, and c_4 are parameter greater than zero. According to the Lagrange function and the KKT condition, the dual problem of the problem can be solved, and the (w_1, b_1) and (w_2, b_2) can be obtained, so that two approximate functions can be obtained, so the final regression function can be obtained.

$$\begin{gathered} f(x) = \frac{1}{2}\left(f_1(x) + f_2(x)\right) = \\ \frac{1}{2}\left(w_1 + w_2\right)^T x + \frac{1}{2}\left(b_1 + b_2\right) \end{gathered} \tag{11}$$

D. Improved DNN weighted e-TSVR model

The DNN weighting method determines the weight value according to the number and density of other sample points around the sample points. In order to predict the accuracy, it is obvious that the weight value should be added to the sample data that exceeds the error tolerance. Here, the nonlinear model is given directly, and the specific solution steps are given. The specific model is shown equation (12) and equation (13).

$$\begin{gathered} \min_{w_1, b_1, \xi} \frac{1}{2} c_3 \left(w_1^T w_1 + b_1^2\right) + c_1 De^T \xi + \\ \frac{1}{2}\left(Y - \left(K\left(A, A^T\right)w_1 + eb_1\right)\right)^T \cdot \\ \left(Y - \left(K\left(A, A^T\right)w_1 + eb_1\right)\right) \\ s.t.\ Y - \left(K\left(A, A^T\right)w_1 + eb_1\right) \ge -\varepsilon_1 e - \xi, \xi \ge 0 \end{gathered} \tag{12}$$

$$\begin{gathered} \min_{w_2, b_2, \eta} \frac{1}{2} c_4 \left(w_2^T w_2 + b_2^2\right) + c_2 De^T \eta + \\ \frac{1}{2}\left(Y - \left(K\left(A, A^T\right)w_2 + eb_2\right)\right)^T \cdot \\ \left(Y - \left(K\left(A, A^T\right)w_2 + eb_2\right)\right) \\ s.t.\ Y - \left(K\left(A, A^T\right)w_2 + eb_2\right) \ge -\varepsilon_2 e - \eta, \eta \ge 0 \end{gathered} \tag{13}$$

As with the above model, we take problem as an example to give a specific solution process, its Lagrange equation is:

$$\begin{gathered} L(w_1, b_1) = \frac{1}{2}\left(Y - \left(K\left(A, A^T\right)w_1 + eb_1\right)\right)^T \cdot \\ \left(Y - \left(K\left(A, A^T\right)w_1 + eb_1\right)\right) + c_1 De^T \xi + \\ \frac{1}{2}c_3\left(\left\|w_1\right\|^2 + b_1^2\right) - \alpha^T - \beta^T \xi \cdot \\ \left(Y - \left(K\left(A, A^T\right)w_1 + eb_1\right) + \varepsilon_1 e + \xi\right) \end{gathered} \tag{14}$$

Among them, the Lagrange multiplier vector is α = (α_1, α_1, … , α_l), β = (β_1, β_1, … , β_l). KKT conditions are as follows:

$$\begin{gathered} -K\left(A, A^T\right)^T\left(Y - K\left(A, A^T\right)w_1 - eb_1\right) + c_3 w_1 + \\ K\left(A, A^T\right)^T \alpha = 0 \end{gathered} \tag{15}$$

$$\begin{gathered} -e^T\left(Y - K\left(A, A^T\right)w_1 - eb_1\right) + c_3 b_1 + \\ e^T \alpha = 0 \end{gathered} \tag{16}$$

$$c_1 De - \beta - \alpha = 0 \tag{17}$$

$$Y - \left(K\left(A, A^T\right)w_1 + eb_1\right) \geq -\varepsilon_1 e - \xi, \xi \geq 0 \tag{18}$$

$$\alpha^T\left(Y - \left(K\left(A, A^T\right)w_1 + eb_1 + \varepsilon_1 e + \xi\right)\right) = 0 \tag{19}$$

$$\beta^T \xi = 0 \tag{20}$$

$$\alpha \geq 0, \beta \geq 0 \tag{21}$$

Based on $\beta \geqq$ 0 and According to formula (15), we can get:

$$0 \leq \alpha \leq c_1 De \tag{22}$$

According to the above solution process, the final dual problem can be solved as:

$$\begin{gathered} \max_{\alpha} - \frac{1}{2}\alpha G\left(G^T G + c_3 I\right)^{-1} G^T \alpha^T + Y^T G \cdot \\ \left(G^T G + c_3 I\right)^{-1} G^T \alpha - \left(e^T \varepsilon_1 + Y^T\right)\alpha \\ s.t. 0 \leq \alpha \leq c_1 De \end{gathered} \tag{23}$$

E. Cuckoo Search Algorithm

The cuckoo search algorithm is based on cuckoos constantly searching, comparing, and finally choosing a nest suitable for their nestlings [20-21]. The characteristics of Levi's flight are mainly small steps, but there are relatively large displacements. This characteristic makes the individual not stay in one place for repeated searches [22-23]. This constantly optimizing life habits and Levi's flight characteristics have resulted in cuckoo search algorithms.

Fig. 4 shows the random search path of Levi's flight. As can be seen from the figure, in two-dimensional space, Levi's flight presents frequent short distance and occasional long distance laws. The valley bird algorithm combines the Levi's flight, optimal selection and random migration to make the population seek Youshi can jump out of the local optimum. Therefore, the global search ability of the cuckoo algorithm is strong.

In cuckoo search algorithm, a nest position corresponds to a nestling egg，One solution in solution space corresponds to one function. It means to choose the best solution during the solution process, so that the fitness value of the function is the best.

The candidate solution of the function in the search space is simulated with the bird's nest position, and the mathematical formula is：

$$s_i^t = Lb + (Ub - Lb) \cdot \mathrm{rand}(size(Lb)) \tag{24}$$

Among them, t represents the number of iterations, *Ub*, *Lb* represents the range of search space.

Iteratively update the bird's nest，Iterations to generate new bird's nest simulations as function candidate solutions， The resulting formula is：

$$s_i^{t+1} = s_i^t + \alpha \oplus levy(\lambda) \tag{25}$$

Among them, α is a real number greater than 0 and is a parameter for adjusting the distance，$s^t{}_i$ represents the current solution, $s^{t+1}{}_i$ represents the solution of the next generation after one iteration.

In the initial stage of algorithm search, the larger step size parameter enables individuals in the population to explore in the feasible region at a greater distance, and strengthens the global search ability of the algorithm; in the later stage of algorithm search, it is necessary to reduce the step length to strengthen the local search ability. In the improved algorithm, logarithmic adaptive control step length is adopted in this chapter, and its calculation formula is as follows:

$$\alpha = \frac{\alpha_0}{\ln(t)} + (1 - \alpha_0) r_i \tag{26}$$

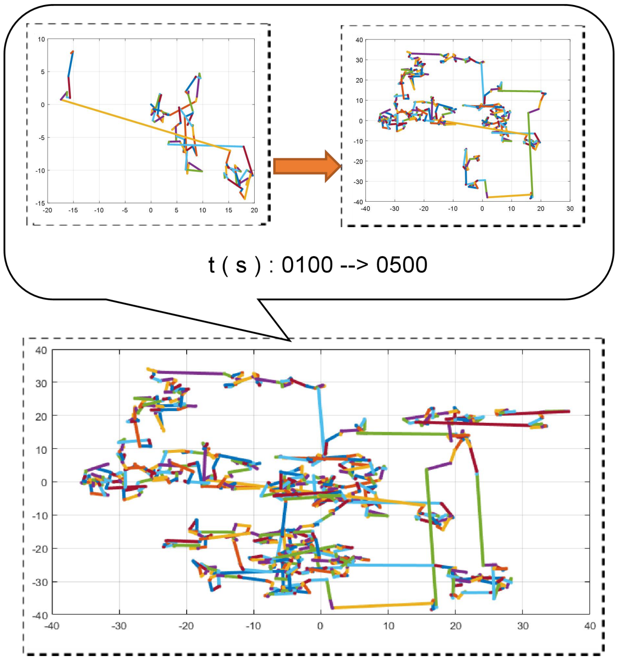


Fig.4 Levy's flight random path

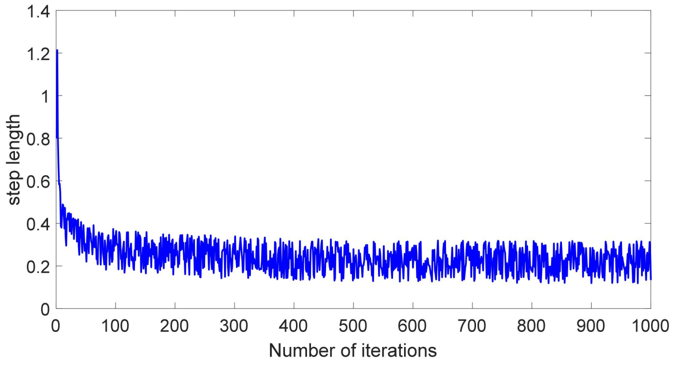


Fig.5 Adaptive iteration step

Where, α_0 is the initial step, t is the current iteration number, and r_i is the random number subject to uniform distribution between [0,1]. Take α_0 = 0.8 for example, the maximum number of iterations is T = 1000, and the step change is shown in **Fig.5**. It can be seen from **Fig.5** that the improved adaptive step size gradually reduces from 1.2 at the beginning to about 0.2, so that the adaptive step size can increase the later search ability.

The cuckoo search algorithm is shown in **Table 1**.

Gain new bird's nest location with Levi's flight. Levi's flight strategy is a funny way to search for space，Its formula is：

$$Levy: u = t^{(-\lambda)}, (1 \le \lambda \le 3) \tag{27}$$

The Levy distribution step size is defined by the algorithm stride l. The model is represented as：

$$l = u / |v|^{\frac{1}{\beta}} \tag{28}$$

Among them， u and v obey normal distribution, among them：

$$u \sim N\left(0, \partial_u^2\right), v \sim N\left(0, \partial_v^2\right)$$
$$\partial_u = \left\{ \frac{\Gamma(1+\beta)\sin(\pi\beta/2)}{\Gamma[(1+\beta)/2]\beta \cdot 2^{(\beta-1)/2}} \right\}^{\frac{1}{\beta}}, \partial_v = 1 \tag{29}$$

Table 1 Cuckoo search algorithm

```
Algorithm 1: Cuckoo search algorithm
Initialization random parameter, population number n = 20,
maximum evolution algebra $t$ = 1000, adaptive step
parameter $\alpha_0$ , discovery probability $p_a$.
$s_i^1 = Lb + (Ub - Lb) \cdot \mathrm{rand}(\mathrm{size}(Lb))$
While $t < T$
    $\alpha = \alpha_0 / \ln(t) + (1 - \alpha_0) \cdot r_i$
    $l = u / \left( |v|^{\frac{1}{\beta}} \right)$
    $s_i^{t+1} = s_i^t + \alpha \oplus l$
    If $f\left(s_i^{t+1}\right) < f\left(s_i^t\right)$
        $s_i^{t+1} = s_i^t$
    end
    $s_i^{t'} = s_i^t + r_1 \oplus H(r_2 - p_a) \oplus \left(s_j^t - s_k^t\right)$
    $t = t + 1$
    end
```

In random migration, the probability 1-p_a of an individual is selected for replacement. The replacement formula is:

$$s_i^{t'} = s_i^t + r_1 \oplus H(r_2 - p_a) \oplus \left(s_j^t - s_k^t\right) \tag{30}$$

In the formula, r_1, r_2 > 0 is generally a random number vector subject to [0,1] uniform distribution, s^t_j and s^t_k are randomly selected two individuals in the population, and $H(\bullet)$ is a step function.

F. Overview of Cancer Prediction Model Based on D-TSVR Algorithm

Step1: collect sample data from patients with cancer recurrence, due to the wide range of recurrence time, some interference with the modeling of cancer recurrence time, do the following for this.

$$t_i^{'} = \frac{t_i - \min(t_i)}{\max(t_i) - \min(t_i)} \tag{31}$$

Step2: Dividing the time to relapse of treated cancer patients into two parts. Training and testing samples, Training samples for improved support vector machine learning, establish a cancer recurrence prediction model. Test samples used to analyze the generalization performance of the cancer recurrence time model.

Step3: Input training samples into D-TSVR for learning.

Step4: Using cuckoo algorithm optimization algorithm to solve the optimal DR domain of DNN algorithm, establish a cancer recurrence prediction mode.

Step5: Prediction of cancer recurrence time using a well established cancer recurrence time prediction model, and analyze the generalization performance of cancer recurrence prediction model.

The D-TSVR algorithm is shown in **Table 2**.

Table 2 D-TSVR algorithm

```
Algorithm 2: D-TSVR algorithm
    Initialize：
     $W_{i,j} = 1$
     $d_i = \sum_{j=1}^{N} W_{i,j}$
     $D$ = diag($d_1$,$d_2$,...$d_N$)
     For $i$=1,2， ...， $m$
        Use Levy flight to get a new $D$ '
        The KKT condition is used to solve the dual problem
  of D-TSVR
        If  New(MAPA)>Old(MAPA)
            Update  $D = D'$
        End
        $D_{best}$=$D$
    End
    Output $D_{best}$=$D$
    Get the optimal D-TSVR model
```

In this paper, a D-TSVR algorithm is proposed, which is DNN weighted TSVR algorithm. For the solution of Dr domain, this paper uses adaptive step size improved cuckoo algorithm to search. The improved cuckoo algorithm has a more stable search ability in the later stage of the search, and the convergence of the algorithm is faster.

IV. Experimental Verification

Recurrence after cancer treatment is an important factor affecting the long-term survival of cancer patients, Study the factors that affect recurrent cancer, and carry out certain clinical intervention strategies, can effectively improve the survival rate of cancer patients.

Relationship between the source of relapsed cancer cell clones and time to relapse: 6 recurrences with unicentricity. The recurrence time is 3 to 13 months from the first operation. The average is (6.5 ± 3.25) months (X ± SD), however 9 cases of recurrence with polycentricity, relapse time is 7 to 54 months, (33.8 ± 17.8) months on average. After statistical processing, there are very significant differences [18-20].

Within 2 years after the first operation, occurred 9 relapses. Among them multicenter recurrence in 3 cases (33％), 6 cases recurred after 2 years, polycentric recurrence (100％), Through checking, prompts recurrence within 2 years after surgery can be monocentric or polycentric;

The recurrences after 2 years were all polycentric recurrences, namely the second primary cancer [24-25]. Single-center recurrence within 2 years after operation, and multiple-center recurrence after 2 years.

A total of 150 cases of recurrence of liver cancer, 150 cases of recurrence of lung cancer, 200 cases of recurrence of kidney cancer, 200 cases of recurrence of breast cancer, 150 cases of recurrence of uterine cancer, 150 cases of recurrence of gastric cancer, 150 cases of recurrence of colon cancer are collected in this article. According to the recurrence of cancer patients after surgery, a recurrence prediction model based on various symptoms of cancer patients is established in this paper.

A. Cancer assessment and data preprocessing

We have selected a total of six symptoms including immune, tumor, microenvironment, psychological, nutrition, aerobic and advanced tasks as predictors of recurrence of cancer patients. **Table 3** is the evaluation criteria of the six indicators. Includes related terms, weights, and evaluation levels for each indication.

Table 3 Cancer Patient Index Evaluation Criteria

Finger syndrome	Related terms and Weights
Basic index	Systemic disease, 2; Family cancer history, 2; Non-essential dependency, 3; age, 1; fat, 1; Habitually high risk, 1; Regional high risk,1.
Immune index	CD3+CD4+CD8+/CD45+, 4; CD3+CD4+/CD45+, 8; CD4+/CD8+, 10; CD3+CD16+CD56+/CD45+, 6; CD3-CD56+,5; CD4+CD25+, 1; Exercise ECG (X±SD), 2; Sports Leather (X±SD), 2.
Tumor index	Size, 10; Placeholder, 10; Violate the relationship, 10; Angiogenesis, 10; Pathological typing, 3; CTC value, 9; Differentiation, 10; Mutation target, 1.
Nutrition index	Total nutrition, 6; Balanced nutrition, 3; Nutrition safety assessment, 5; Cancer cell proliferation, 10; Immune cell proliferation, 10; Angiogenesis, 8; Amino acid evaluation, 5; Proteomics evaluation, 10.
Psychological indication	Life event scale, 1; Cornell Medical Index, 2; Self-rating anxiety scale, 5; Self-rating depression scale,5; Baker Anxiety Scale, 5; Baker Depression Questionnaire,5; Pittsburgh sleep Quality index, 4; Texas Social Behavior Questionnaire, 3; Family function assessment, 1; Exercise ECG (X±SD), 2; Sports Leather (X±SD), 2.
Microenvironment index	O2, 3; PH value, 4; Interstitial pressure, 2; Inflammatory response, 7; Vascular permeability, 6; CTC value, 9; Proteomic analysis, 8.
Exercise and advanced work	Aerobic exercise, 4; Advanced social work, 3; Texas Social Behavior Questionnaire, 3.

Table 3 shows the evaluation criteria for each indication. Among them the evaluation criteria for tumor index are different for different cancers, and are omitted in the article. Surveillance of five other indications. Weighting is a weight given based on the doctor's experience. Convert this weight to a specific value. Level A is best with 1 point and Level D is 9 points with worst. The formula for calculating the comprehensive score of each indication is as follows.

$$I = \sum_{i=1}^{n} x_i w_i \Big/ \sum_{i=1}^{n} w_i \tag{32}$$

Among them x_i is the value of the *i-th* item, w_i is weight for item i. **Fig. 6** is a sample chart of the recurrence time of various cancer patients. we regard the recurrence time of 60 months as complete recovery. It can be seen from the figure that the sample recurrence time distribution is relatively uniform, which is more suitable for cancer recurrence analysis.

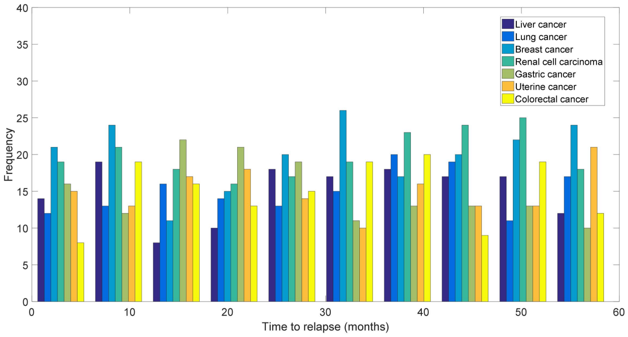


Fig.6 Distribution of recurrence time of various cancer samples

B. Result Analysis

We give the performance evaluation index as shown below.

MSE, which is used to calculate the fitting ability of the model.

$$MSE = \frac{1}{N}\sum_{i=1}^{N}\left(p_i - q_i\right)^2 \tag{33}$$

Mean absolute percentage accuracy (MAPA), the index is used to calculate the fitting ability and generalization ability of the model. The larger the index is, the stronger the fitting ability and generalization ability are.

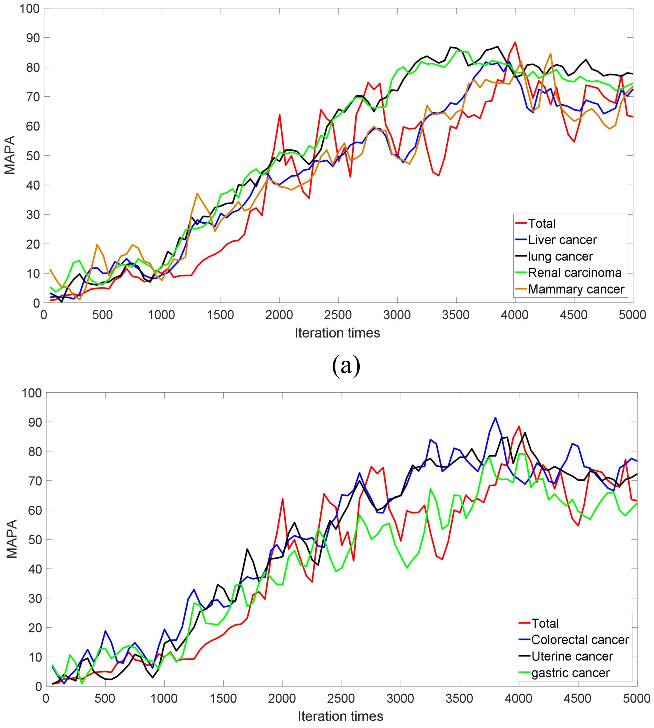
$$MAPA = 1 - \frac{1}{N}\sum_{i=1}^{N}\left|\frac{p_i - q_i}{q_i}\right| \times 100\% \tag{34}$$

Among them, p_i represents the predicted recurrence time of the model, q_i represents the actual recurrence time of cancer patients.

We build a cancer recurrence prediction model based on the improved TSVR algorithm, use the DNN algorithm with local information mining function to weight the TSVR algorithm, and use the cuckoo algorithm to solve the optimal Dr domain of the DNN algorithm. The model shows good performance, and **Fig. 7** shows the iterative map of the MAPA of cancer recurrence prediction model training.

As can be seen from **Fig. 7**, with the increase of training times, the accuracy rate of the model for prediction of different cancer recurrence tends to be stable, reaching more than 90%, and the prediction results are relatively accurate.

Fig. 8 is the error iteration diagram of D-TSVR, CNN [23] and e-TSVR algorithm for cancer recurrence prediction. We know that the training errors of the three algorithms decrease with the increase of training times. The training errors of e-TSVR are finally stable at about 0.44, the CNN algorithm is finally stable at about 0.37, and the D-TSVR algorithm is finally stable at about 0.24.



(a)

(b)

Fig.7 Model training iteration

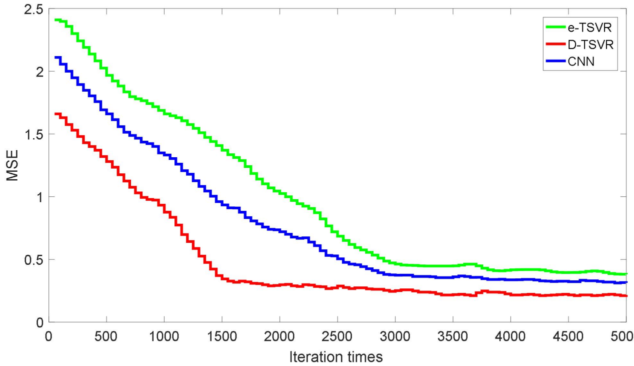


Fig.8 Algorithm training error iteration diagram

We compared different prediction models. Among D-TSVR, e-TSVR and CNN, D-TSVR has the best prediction performance. The CNN network is provided with 1 input layer, 3 convolutional layers, 3 pooling layers, a fully connected layer and an output layer. The size of the convolution kernel set in the 1st to 3rd convolutional layers of the CNN network is 2 × 2. The number of convolution kernels is 20, the three pooling are all used for maximum sampling, and the kernel size is 2 × 2 The activation function adopts the Relu function. The number of fully connected layer neurons is 25. The weight *w* and the bias *b* are initialized randomly. **Fig.8** shows the comparison of MSE of three different algorithms for predicting the time of cancer recurrence. As can be seen from the figure, the D-TSVR algorithm has a faster convergence rate and has already converged at 1500 iterations. The other two the convergence speed of this algorithm is slow. If the prediction accuracy is uniform, the time complexity of the D-TSVR algorithm is the least of the three algorithms.

Table 4 is the average result of each algorithm running 20 times. It can be seen from **Table 4** that the prediction accuracy of the improved TSVR algorithm is significantly higher than that of the other three algorithms, and the prediction accuracy remains above 91%. The prediction accuracy of CNN algorithm and e-TSVR algorithm is basically the same.

Table 4 MAPA of Cancer Recurrence Prediction by Different Algorithms (%)

Material	e-TSVR	CNN	D-TSVR
Liver cancer samples	76.25	85.32	93.52
Lung cancer samples	78.24	86.59	94.52
Breast cancer samples	76.49	84.38	93.68
Renal cell carcinoma samples	75.98	89.26	96.24
Gastric cancer samples	79.64	87.67	91.25
Uterine cancer samples	80.15	83.68	95.68
Colorectal cancer samples	76.29	89.46	92.39
All samples	73.26	85.29	91.26

We performed paired T-tests on the MAPAs simulated by e-TSVR, CNN, and D-TSVR for 20 times. Table 4 shows the results of the T-test. As can be seen from **Table 5**, the P values of the T-test in all cases Less than 0.05, that is, in a statistical sense, the mean of the three algorithms is different, and the performance of the D-TSVR algorithm is optimal.

Table 5 T test of MAPA with different algorithms

	e-TSVR_CNN		e-TSVR_D-TSVR		CNN_D-TSVR	
	T-Value	P-Value	T-Value	P-Value	T-Value	P-Value
Liver cancer samples	-22.183	0.000	-74.421	0.000	-30.986	0.000
Lung cancer samples	-30.360	0.000	-58.297	0.000	-34.470	0.000
Breast cancer samples	-29.388	0.000	-63.028	0.000	-46.391	0.000
Renal cell carcinoma samples	-58.583	0.000	-89.891	0.000	-25.997	0.000
Gastric cancer samples	-28.251	0.000	-36.475	0.000	-14.231	0.000
Uterine cancer samples	-12.144	0.000	-69.405	0.000	-61.089	0.000
Colorectal cancer samples	-58.061	0.000	-75.949	0.000	-13.163	0.000
All samples	-50.951	0.000	-83.595	0.000	-30.984	0.000

V. Conclusion

In this paper, we collected 50000 of cancer recurrence data, including liver cancer, lung cancer, renal cancer, breast cancer, uterine cancer, gastric cancer and colorectal cancer. First of all, we get the scores of each index by the physical health evaluation (medical evaluation) of each patient, establish a cancer recurrence prediction model based on D-TSVR algorithm, and predict the recurrence time of various cancers. Then compared with e-TSVR and CNN algorithm, it is found that the improved TSVR prediction accuracy is significantly higher than other models. The prediction accuracy rate is above 91%. It shows that the cancer recurrence prediction model proposed in this paper is more suitable for predicting the recurrence time of cancer patients, and improving the survival rate of patients by corresponding clinical interventions.

In future work, we will continue to improve the algorithm to further improve the accuracy of the model prediction, and then through nutrition intervention, change the patient's nutritional indicators and other indicators to further improve the patient's survival time and quality of life.

References

[1] Shi D, Che J, Yan Y, et al. Expression and clinical value of CD105 in renal cell carcinoma based on data mining in The Cancer Genome Atlas[J]. Experimental and therapeutic medicine, 2019, 17(6): 4499-4505. DOI:10.3892/etm.2019.7493

[2] Jia-Hao Bi, Yi-Fan Tong, Zhe-Wei Qiu, et al. ClickGene: an open cloud-based platform for big pan-cancer data genome-wide association study, visualization and exploration[J]. BioData Mining, 2019, 12(1):12. DOI:10.1186/s13040-019-0202-3

[3] Dsouza K J, Ansari Z. Experimental exploration of support vector machine for cancer cell classification[C]//2017 IEEE International Conference on Cloud Computing in Emerging Markets (CCEM). IEEE, 2017: 29-34. DOI:10.1109/CCEM.2017.15

[4] Hasan M R, Hassan N, Khan R. Classification of cancer cells using computational analysis of dynamic morphology[J]. Comput Methods Programs Biomed, 2018, 156:105-112. DOI:10.1016/j.cmpb.2017.12.003

[5] Yan K, Lu H. An Extended Genetic Algorithm Based Gene Selection Framework for Cancer Diagnosis[C]// 2018 9th International Conference on Information Technology in Medicine and Education (ITME). IEEE Computer Society, 2018. DOI:10.1109/ITME.2018.00021

[6] Jack B. Fu, Diana M. Molinares, Shinichiro Morishita, et al. Retrospective Analysis of Acute Rehabilitation Outcomes of Cancer Inpatients with Leptomeningeal Disease: Leptomeningeal Disease Rehabilitation[J]. PM&R, 2019. DOI:10.1002/pmrj.12207

[7] Fadime Cenik, Bruno Mähr, Stefano Palma, et al. Role of physical medicine for cancer rehabilitation and return to work under the premise of the “Wiedereingliederungsteilzeitgesetz”[J]. Wiener klinische Wochenschrift, 2019, 131(12):1-7. DOI:10.1007/s00508-019-1504-7

[8] Jose, A, Karam, et al. Use of combined apoptosis biomarkers for prediction of bladder cancer recurrence and mortality after radical cystectomy[J]. The Lancet Oncology, 2007. DOI:10.1016/S1470-2045(07)70002-5

[9] Bartels P H, Montironi R. Karyometry-based method for prediction of cancer event recurrence[J]. 2007. DOI: US20070100562 A1

[10] Y Han, Z Han, J Wu, Y Yu, et al. Shuqing Gao, Artificial Intelligence Recommendation System of Cancer Rehabilitation Scheme Based on IoT Technology[J]. IEEE Access, 2020. DOI: 10.1109/ACCESS.2020.2978078

[11] Li A, Wang R, Liu L, et al. BCRAM: A social-network-inspired breast cancer risk assessment model[J]. IEEE Transactions on Industrial Informatics, 2018, 15(1): 366-376. DOI: 10.1109/TII.2018.2825345

[12] Richard, Ha, Peter, et al. Convolutional Neural Network Using a Breast MRI Tumor Dataset Can Predict Oncotype Dx Recurrence Score[J]. Journal of Magnetic Resonance Imaging Jmri, 2018. DOI: 10.1002/jmri.26244

[13] Kumar N, Verma R, Arora A, et al. Convolutional neural networks for prostate cancer recurrence prediction[C]// Medical Imaging 2017: Digital Pathology. International Society for Optics and Photonics, 2017. DOI: 10.1117/12.2255774

[14] Yunfei H, Jun M, An W, et al. A support vector machine and a random forest classifier indicates a 15-miRNA set related to osteosarcoma recurrence[J]. Oncotargets & Therapy, 2018, Volume 11:253-269. DOI: 10.2147/OTT.S148394

[15] Jinting, Zhou, Lin, et al. Establishment of a SVM classifier to predict recurrence of ovarian cancer[J]. Molecular medicine reports, 2018. DOI: 10.3892/mmr.2018.9362

[16] A Yang, Y Zhuansun, Y Shi, et al. IoT System for Pellet Proportioning Based on BAS Intelligent Recommendation Model[J]. IEEE Transactions on Industrial Informatics, 2019. DOI: 10.1109/TII.2019.2960600

[17] Jiang Q, Yan X. Learning deep correlated representations for nonlinear process monitoring[J]. IEEE Transactions on Industrial Informatics, 2018, 15(12): 6200-6209. DOI: 10.1109/TII.2018.2886048

[18] Massrur H R, Niknam T, Mardaneh M, et al. Harmonic elimination in multilevel inverters under unbalanced voltages and switching deviation using a new stochastic strategy[J]. IEEE transactions on industrial informatics, 2016, 12(2): 716-725. DOI: 10.1109/TII.2016.2529589

[19] Peng X. TSVR: An efficient Twin Support Vector Machine for regression[J]. Neural Networks, 2010, 23(3):365-372. DOI:10.1016/j.neunet.2009.07.002

[20] Shao Y H, Zhang C H, Yang Z M, et al. An ε -twin support vector machine for regression[J]. Neural Computing & Applications, 2013, 23(1):175-185. DOI:10.1007/s00521-012-0924-3

[21] Ömer Faruk Ertuğrul, Mehmet Emin Tağluk. A novel version of k nearest neighbor: Dependent nearest neighbor[J]. Applied Soft Computing, 2017, 55:480-490. DOI:10.1016/j.asoc.2017.02.020

[22] Cupit-Link M, Syrjala K L, Hashmi S K. Damocles’ Syndrome Revisited: Update on the Fear of Cancer Recurrence in the Complex World of Today’s Treatments and Survivorship[J]. Hematology/Oncology and Stem Cell Therapy, 2018, 11(3):129-134. DOI:10.1016/j.hemonc.2018.01.005

[23] Sharpe L, Turner J, Fardell J E, et al. Psychological intervention (ConquerFear) for treating fear of cancer recurrence: mediators and moderators of treatment efficacy[J]. Journal of Cancer Survivorship, 2019(8):695-702. DOI:10.1007/s11764-019-00788-4

[24] Zhou Z, Liu X, Hu K, et al. The clinical value of PET and PET/CT in the diagnosis and management of suspected cervical cancer recurrence[J]. Nuclear Medicine Communications, 2018, 39(2):97-102. DOI:10.1097/MNM.0000000000000775

[25] Ruan M, Ren Y, Wu Z, et al. The Survey of CNN-based Cancer Diagnosis System[J]. IOP Conference Series Materials Science and Engineering, 2018, 466. DOI:10.1088/1757-899X/466/1/012095

Ai-Min Yang received his B.Sc. degree in 2002 from Yanshan University, received his M.Sc. degree in 2004 from Yanshan University, received his Ph.D. degree in 2015 from Yanshan University, now he is Professor in North China University of Science and Technology. His main research interests include Numerical computation, Iron and steel big data and Intelligent computation, etc.

Yang Han received the B.S. and M.S. degrees from North China University of Science and Technology in 2010 and 2018, respectively. He is currently a Ph.D. student at North China University of Science and Technology. His main research interests include numerical computation, iron and steel big data, and intelligent computation, etc.

Chen-Shuai Liu was born in Zhengzhou, Henan in 1998. He received the B.S. degree from North China University of Science and Technology in 2020. His interests include numerical calculation, mathematical statistics and high performance computing.

Jian-Hui Wu received the B.Sc. degree from Yanshan University, in 2002, and the M.Sc. Degree from Hebei United University, in 2012. He is currently pursuing the Ph.D. degree with the North China University of Science and Technology, where he is also an Associate Professor. His main research interests include data mining and occupational epidemiology

Dian-Bo Hua, Co-founder of Beijing Stre-Laboratory.The lab focuses on the field of cancer data analysis. (Address: B-602 wanda plaza, tongzhou district, Beijing ;E-mail: 78225474@qq.com). He was born in Xingtai, Hebei, China, in 1981. He received the BS degrees from Yanshan University of Mathematics in 2002. His main research interests include Cancer rehabilitationl and nutritional support big data , etc.