Submitted: 02 August, 2023 Accepted: 26 September, 2023 Published: 15 May, 2025 DOI:10.22514/ejgo.2025.068

ORIGINAL RESEARCH

European Journal of Gynaecological Oncology

Breast cancer recurrence time prediction based on the MOPSO-RF model

Jian Wang[1,2,3], Hao Li[1,2,3], Shujuan Yuan[3,*], Fengchun Liu[1,2,3,4], Aimin Yang[1,2,3,4], Dianbo Hua[5]

[1]Hebei Key Laboratory of Data Science and Application, North China University of Science and Technology, 063210 Tangshan, Hebei, China
[2]The Key Laboratory of Engineering Computing in Tangshan City, North China University of Science and Technology, 063210 Tangshan, Hebei, China
[3]College of Science, North China University of Science and Technology, 063210 Tangshan, Hebei, China
[4]Tangshan Intelligent Industry and Image Processing Technology Innovation Center, North China University of Science and Technology, 063210 Tangshan, Hebei, China
[5]Beijing Sitairui Cancer Data Analysis Joint Laboratory, 101100 Beijing, China

***Correspondence**
yuanshujuan@ncst.edu.cn
(Shujuan Yuan)

Abstract

Background: Cancer is a complex disease where malignant tumors have high clinical incidence and mortality rates. A significant risk of postoperative recurrence also exists. This work was aimed at diagnosing the time to postoperative cancer recurrence which could help the patients. The post-operative breast cancer recovery data from Beijing Stre Cancer Data Analysis Joint Laboratory was utilized. The study was conducted to determine the weighting of adaptive symptoms regarding the time to breast cancer recurrence by selecting six indicators, *i.e.*, immune, tumour, microenvironmental, psychological, nutritional and aerobic exercise and progressive work, involved in the mechanism of breast cancer recovery. **Methods**: A multi-objective particle swarm optimised random forest (MOPSO-RF) model for the adjuvant cancer diagnosis was introduced to predict the time to breast cancer recurrence. The random forest model was optimised to verify its accuracy. The multi-objective optimisation combined each indicator with time to the breast cancer recurrence to get an objective function for finding the optimal Pareto solution. **Results**: The results exhibited that the constructed model was effective in predicting the time to breast cancer recurrence and thus provided early warning indications in this regard. **Conclusions**: The model achieved 92.17% forecast accuracy compared to the Gradient Boosted Tree (GBDT), Support Vector Machine (SVM), Linear/Logistic Regression (LR), XGBoost and Random Forest (RF) algorithm models.

Keywords

Multi-objective optimization; Particle swarm optimization; Random forest; Recurrence time; Assisted diagnosis; Cancer prediction

1. Introduction

Human health problems have escalated because of the continuous development of internet technologies, the improvement in life standards of human communities, the living conditions, and the transformed ways of human life. An aggressive form of cancer usually develops in breast cells. The breast cancer incidence in women has increased by 0.5% per year between 2014 and 2018. The breast, lung and colorectal cancers account for 51% of all the new diagnoses with nearly one third of breast cancer alone [1]. In 2022, there are about 4.82 million and 2.37 million new cancer cases, and 3.21 million and 0.64 million cancer deaths in China and US, respectively, with breast cancer becoming the most common in both countries [2]. Literature [3] defines machine learning as "a computational method that uses experience to improve performance or make accurate predictions". Early cancer research often combines diagnosis and machines [4]. Machine learning techniques have thus been employed in health care [5] to handle huge data volumes, and to improve the data usage efficiency. This degree of intelligence can give a new impetus to the long-term development of medicine field.

Early identification and corresponding cancer therapy have high survival rates with improved patient's life quality. However, the influential factors of post-surgical recurrence in oncology patients are complicated and it is challenging to quantitatively analyse them [6]. Some work has been done in this regard. Studies have focused on the data from cancer patients condition status, tumour trends such as the prediction of benign or malignant tumours, the prediction of time to recurrence after surgery, and prediction of tumour type [7]. The time to postoperative recurrence influences the patient's health and living standards. The patient prophylaxis interventions have impact on the time to postoperative recovery in patients under study. The patients' survival duration and life quality have been prolonged through personalised interventions in cancer patients with short time to recurrence. These are the significant contributions towards cancer rehabilitation [8]. The study herein investigates the relationship between postoperative relapse time of cancer patients and each indicator. It proposes a model based on the combination of multi-objective particle swarm optimisation and random forest. It collects the clinical

data of patients and combines with the prediction model to predict relapse time of patients recovering from breast cancer. It optimises the objective function composed of relapse time of each indicator of the patient. The correlation coefficient is found out for each indicator and the relapse time. The six indicators regarding patient's postoperative recurrence time are utilized to find the optimal solution of each indicator without affecting the other indicators. The doctor's diagnosis is thus assisted by improving the recovery rate of cancer patients which is important for cancer care, treatment, and diagnosis.

Section 2 of this manuscript presents the current state of cancer recurrence prediction, and identifies factors influencing the postoperative recurrence in cancer patients by assessing the relevant medical knowledge and criteria; Section 3 describes the multi-objective particle swarm optimization algorithm, the random forest algorithm, and the multi-objective particle swarm optimized random forest model; Section 4 constructs a random forest model based on multi-objective particle swarm optimisation by collecting the data from breast cancer patients and comparing it with GBDT, SVM, LR (linear regression), LR (logistic regression), XGBoost and RF (random forest). It is concluded that the prediction accuracy of random forest model based on multi-objective particle swarm optimisation is higher than those of other models. Results suggest that the recurrence time prediction model is suitable for predicting the recurrence time of cancer patients. The objective function is composed of each indicator and optimised for the effect on recurrence time of breast cancer. The effect of each indicator on the postoperative recurrence time of breast cancer patients is found out without being influenced by the other indicators; Section 5 presents the specific effects and roles of the model by summarising full text, multi-objective optimisations for each objective function, and encompassing the implications of multi-objective optimisations for breast cancer patients. The study of cancer auxiliary diagnostic models assists doctors in diagnosing patients, providing nutritional solutions to the patients, and consequently prolonging the patients' life. The study outcomes have guiding significance for cancer handling and its reinforcement.

2. Related work

Cancer has consistently posed a global medical challenge. Achieving a complete cancer cure is difficult. Cancer treatments often focus on alleviating the symptoms to improve patients' life quality and extend lifespan. Cancer patients cannot achieve full recovery, however increasing number of advanced technologies are being applied in the medical field. The 5-year survival rates in late-stage cancer patients have risen from 2%~5% decades ago to 16%~23% today [9]. The accumulation of cancer patients' data, and artificial intelligence (AI) development and application can be advantageous over traditional medical models. The long-term AI utilization can drive the cancer rehabilitation diagnostics toward higher-end, personalized, precise, and intelligent approaches.

2.1 Machine learning applied to cancer

In assisted cancer diagnosis and treatment, the studies focus on data mining regarding the recurrence prediction after cancer surgery. It can mine the complex cancer risk factors to find the corresponding recovery patterns of cancer patients. It is useful for the rehabilitation treatment of patients if physicians can use data mining techniques to forecast the patient's condition. The big data techniques are employed for breast cancer prediction. Literature [10] considers two types of data, namely gene expression (GE) and DNA methylation (DM) to extend machine learning algorithms for the classification through separately and jointly applying each dataset.

In 2018, W. Yue *et al.* [11] overviewed the distinct machine learning algorithms from the Wisconsin Breast Cancer Database (WBCD), wherein the particle swarm algorithm and its improved machine learning algorithms pertaining to performance were examined. Kaur *et al.* [12] proposed a random forest technique based on multi-objective differential evolution to optimise the random forest parameters. Qi and Chen *et al.* [13] used a combination of random forests and Particle Swarm Optimization (PSO) algorithm, where PSO algorithm was employed for optimizing the parameters of random forests. Adnan and Islam *et al.* [14] utilised a genetic algorithm to find the optimal number of trees for random forest algorithm. S. Kabiraj and M. Raihan *et al.* [15] employed random forest and extreme gradient enhancement (XGBoost) to predict breast cancer via a breast cancer dataset. The random forest algorithm obtained 74.73% accuracy and the XGBoost achieved 73.63%. D. Yifan *et al.* [16] proposed a breast cancer classification prediction model to provide a benign or malignant diagnosis by fusing the random forest and AdaBoost algorithms. Test results shows that the prediction accuracy of integrated model was improved by an average of 4.3% and up to 9.8% over single algorithm model, which provided a new reference model for breast cancer prediction. M. Ranjan and A. Shukla *et al.* [17] investigated cancers in lung, liver, kidney, breast and brain. The proposed methods for classifying cancer malignancy were random forest classifier, convolutional neural network and ResNet50. The accuracies of 99%, 75.75%, 88.09%, 96% and 81% were attained for lung, liver, kidney, breast, and brain cancers, respectively. Z. Huang *et al.* [18] developed a hierarchical clustered random forest (HCRF) model. The decision trees were clustered using hierarchical clustering techniques by measuring the similarity between decision trees. The variable importance measure (VIM) method optimised the number of features selected for breast cancer prediction. The classification of HCRF-based algorithm using VIM as the feature selection method achieved the best accuracies of 97.05% and 97.76% compared to Adaboost and random forests decision trees. Ai-Min Yang *et al.* [19] proposed cuckoo algorithm to determine the optimal parameters of Deep Neural Networks (DNN) for optimal dependency region (DR) domain. They used improved TSVR algorithm to establish cancer recurrence prediction model which could predict various cancers with the accuracy of over 91%, *i.e.*, higher than that of Cable News Network (CNN) and e-TSVR models. The technology-enabled cancer healthcare industry could make accurate cancer diagnosis because of the multiplexed analysis and complementary

cancer big data combined with the experience of physicians in practical clinical applications. In 2022, Zezhong Ma and Meng Zhang *et al.* [9] combined the federated learning framework with convolutional neural network to build an assisted diagnostic model by taking physical examination indexes of patients as inputs and recurrence time and recurrence location as outputs. The joint prediction model based on CNN improvement achieved >90% accuracy for the data of five cancer types under joint modelling and simulation conditions.

The literature [20] proposed number of deep learning-based prediction models in stacked integration framework to improve the prediction of postoperative breast cancer from existing multimodal datasets with 90.2% accuracy. Successful breast cancer treatment increased the chances of survival in women. Machine learning could support the discovery of important patterns from medical data through cancer predictive models [21] and assisted in the conducive treatment of chest cancer after diagnosis. Results depicted that the Random Forest was a better model for "training data" and "test data" compared to J48 and Classification and Regression Tree (CART) models. Moreover, the Random Forest provided valuable information to identify the important patterns. The multi-objective optimisation solved more metric problems, and large number of nature-inspired metaheuristic algorithms were developed to solve the multi-objective problems [22]. A. Rajagopal *et al.* [23] introduced a new multi-objective particle swarm optimization model to evolve state-of-the-art deep CNNs for scene classification which generated non-dominated solutions in an automatic manner at Pareto front-end. Therefore, this manuscript employed a multi-objective particle swarm optimization random forest model for predicting the time to breast cancer recurrence.

2.2 Ancillary diagnostic models

The tumour treatments are increasing due the advancements in medical field, however, every method has some degree of damage to patient's body. The World Health Organisation has clarified that cancer is a lifestyle disease, and there are links between cancer pathogenesis and geographic impact, working environment, lifestyle habits and medication [24, 25]. It is therefore important to treat cancer alongside patient rehabilitation. Cancer rehabilitation is thus an emerging interdisciplinary area that combines rehabilitation medicine and oncology. In 1971, the oncology rehabilitation concept was first introduced in US National Cancer Program and was defined by Cromes [26] as helping the cancer patients to maximise their physical, social, psychological, and occupational capabilities pertaining to disease and its treatment. Cancer rehabilitation occurs throughout the process of diagnosis–treatment–post-treatment–end-stage–follow-up family support. Given the cancer gravity, the rehabilitation effects are complex regarding medical psychology, ethics, sociology, nutrition, and others. The mortality rate of cancer patients over the decades has increased and life expectancy has lengthened as a result. Oncology rehabilitation has become increasingly popular over the last 20 years. A model for predicting the time to recurrence in breast cancer patients has been developed, as shown in Fig. 1. This will ensure a smooth recovery process taking care of the patient's physical, mental, social, and work wellbeing.

3. MOPSO-RF model basis

3.1 Theoretical knowledge

A random forest (RF) regression model is a regressor consisting of multiple regression trees $\{h(X, \theta_k)\}$, where θ_k is an independently distributed random variable, and the final regression prediction of input vector X is determined by a vote of all the regression trees in forest (taking the average of predictions of all the regression trees). The RF regression model like the other regression models can explain the effects of independent variables on dependent variables. The decision tree algorithm is first understood for studying the

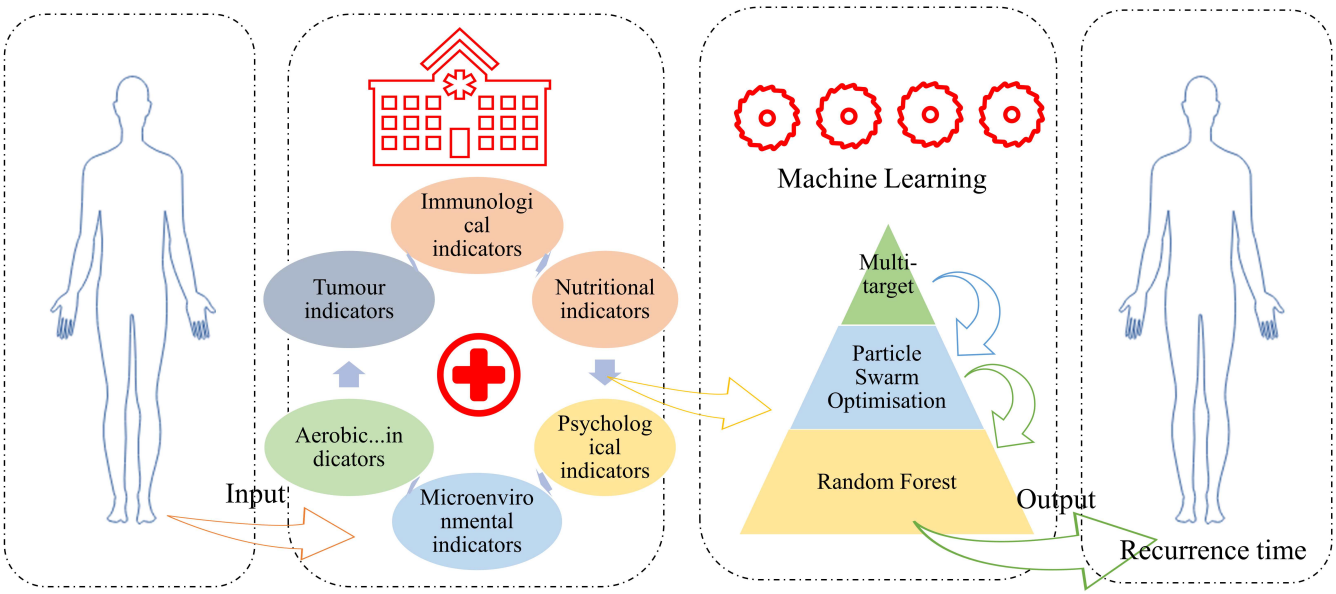

FIGURE 1. Prediction model of time to postoperative recurrence of breast cancer.

RF algorithm. A decision tree model is constructed based on pre-processed training sample data to use the decision tree algorithm for classification and regression. The constructed model is then employed to classify or regress the predicted sample data for prediction analysis. They are the powerful tools in data mining and machine learning to draw conclusions from the results of tree models. Decision trees are easy to understand as their results mimic the human brain, and it is thus easy to understand the rules derived from these results. These algorithms have an impact in the medical field and help doctors to make critical decisions on specific pathology reports because of their tree-based nature. Some popular decision tree models include ID3, C4.5, Chi-squared Automatic Interaction Detector (CHAID) and CART [27]. It is necessary in the decision tree construction to choose features for usage in the segmentation of training samples. The segmentation principle is that the "purity" of node after segmentation must be higher than before segmentation, which otherwise is a leaf node point without segmentation. The main indicators for evaluating nodes "purity" in the model include information gain, information gain rate and Gini coefficient.

The particle swarm optimization (PSO) algorithm is inspired by the foraging behaviour of bird flocks. Modelling is done in the way birds in a flock find food by sharing information among themselves and abstracting that way into an algorithm [28]. It is simple and easy to implement with few control parameters and excellent global search capabilities. It has become an evolutionary computer technology study. In the standard PSO algorithm [29], the swarm consists of N particles, each representing a potentially feasible solution in Dth dimensional search space. The ith particle is composed of position vector as shown in Eqn. 1 and velocity vector as shown in Eqn. 2

$$X_i(t) = \left(x_i^1(t), x_i^2(t), \cdots, x_i^D(t)\right) \tag{1}$$

$$V_i(t) = \left(v_i^1(t), v_i^2(t), \cdots, v_i^D(t)\right) \tag{2}$$

in Dth dimension. At $t+1$th generation, particle i updates the velocity equation at dth dimension as shown in Eqn. 3.

$$\begin{aligned} v_i^d(t+1) = w \cdot v_i^d(t) + c_1 \cdot r_1^d(t) \cdot \left(Pbest_i^d(t) - x_i^d(t)\right) \\ + c_2 \cdot r_2^d(t) \cdot \left(Gbest^d(t) - x_i^d(t)\right) \end{aligned} \tag{3}$$

Where $i = 1, 2, \cdots, N, d = 1, 2, \cdots, D$, N and D denote population size and search space dimension, respectively; acceleration factors c_1 and c_2 are the non-negative constants used to control the effect of self-learning and social learning terms in the search element; and r_1^d and r_2^d represent random numbers obeying a uniform distribution on the interval $[0,\ 1]$.

Multi-objective Optimization Problems (MOPs) have multiple conflicting objectives. Taking the minimisation objective as an example, MOPs can be described as in Eqn. 4.

$$\begin{aligned} \min F(x) = \left(f_1(x), f_2(x), \cdots, f_m(x)\right)^T \\ subject\ to\ x \in \Omega \end{aligned} \tag{4}$$

In equation (2), m is the number of objective functions; x is the decision variable; $f_i(x)$ is the ith objective function. If the decision variable x_1 completely dominates another decision vector x_2, it is denoted as $x_1 \prec x_2$. The conditions to be satisfied are shown in Eqn. 5.

$$\begin{aligned} \forall i \in (1, 2, ..., m) : f_i(x_1) \leq f_i(x_2) \quad and \\ \forall j \in (1, 2, ..., m) : f_j(x_1) \leq f_j(x_2) \end{aligned} \tag{5}$$

In MOPs, a solution is Pareto-optimal when it is not dominated by any other solution. The set of all Pareto-optimal solutions in the search space forms a trade-off surface called the Pareto front.

3.2 MOPSO-RF model

The parameters uncertainty caused by manual empirical settings of initial parameters in algorithm training model can be avoided by employing the initialised particle swarm algorithm to search for the optimal values of initial parameters in RF algorithm. The multi-objective particle swarm algorithm selects the appropriate number (n_estimators) and maximum depth (max_depth) of decision trees to ensure that the training model generalisation is enhanced, and the predictive capability of the model is improved. The n_estimators and max_depth of decision trees are chosen as the particle dimensions which reduce the running time of algorithm search and enhance search effect to a certain extent. The algorithm search area is constructed in two dimensions. The algorithm searches for parameter values to optimise the training model, while shortening the search run time, and thus ensuring the algorithm efficiency.

The multi-objective optimization goal is to obtain high-quality Pareto-optimal solution sets instead of just one optimal solution. When using MOPSO-RF algorithm to deal with multi-objective optimization problems, the relationship between six indicators of breast cancer patients and time to cancer recurrence is employed as the objective function. This finds the Pareto-optimal solution set for each objective function, *i.e.*, for each indicator to remove the influence of other indicators on the predicted time to breast cancer recurrence and construct a MOPSO-RF model.

The extension of MOPSO-RF algorithm like the evolutionary algorithm requires adjustments to the algorithm for handling the multi-objective optimization problems, which includes the following three points.

(a) How to save many Pareto optimal solutions searched by the algorithm from beginning of the run to the end;

(b) How to select the leading particles (individually optimal, globally optimal) to guide them towards the Pareto-optimal solution region and discover the solution that approximates the true Pareto frontier as closely as possible;

(c) How to balance the search and exploitation capabilities in the evolutionary process to maintain the population diversity

while avoiding its falling into local optimum.

As per the above three requirements, this manuscript proposes a MOPSO stochastic Senri algorithm. The pseudo-code of algorithm is described in Table 1.

TABLE 1. MOPSO-RF algorithm.

1	Initialize POP
2	Evaluate index
3	Initialize the Archive
4	For t = 1 to iter do
5	For i = 1 to nPop do
6	Select a for pop(i)
7	Update pop(i) using PSO
8	Multi-objective particle swarm optimization on pop(i)
9	MOPSO-RF update objective function
10	Handle constraints on pop(i)
11	Evaluate pop(i)
12	Update pbest(i)
13	End
14	Update the Archive
15	Recommended plan
16	End

PSO: particle swarm optimization; MOPSO-RF: multi-objective particle swarm optimised random forest.

3.3 Algorithm steps and flow

MOPSO-RF algorithm steps are:

(1) Initialization: initialization of the algorithm parameters and the velocity and position of the particles;

(2) Calculate the particle fitness value and save non-dominated solution; update the individual optimum (pbest);

(3) Update the velocity and position of particles;

(4) Calculate the objective function value of particle and update pbest by comparing to the pbest of previous iteration;

(5) Update the external archives;

(6) Multi-objective particle swarm optimization of random forest construction objective functions;

(7) Update the optimal solution for objective function;

(8) Help doctors find the right cure for breast cancer patients;

(9) Determine if termination is required. Output all the non-dominated solutions to the external file in that case, otherwise skip to step (3).

MOPSO-RF algorithm flow chart is given in Fig. 2, where N is both correct when relevant conditions are met to proceed, and Y stands for wrong when the relevant conditions are not met to proceed.

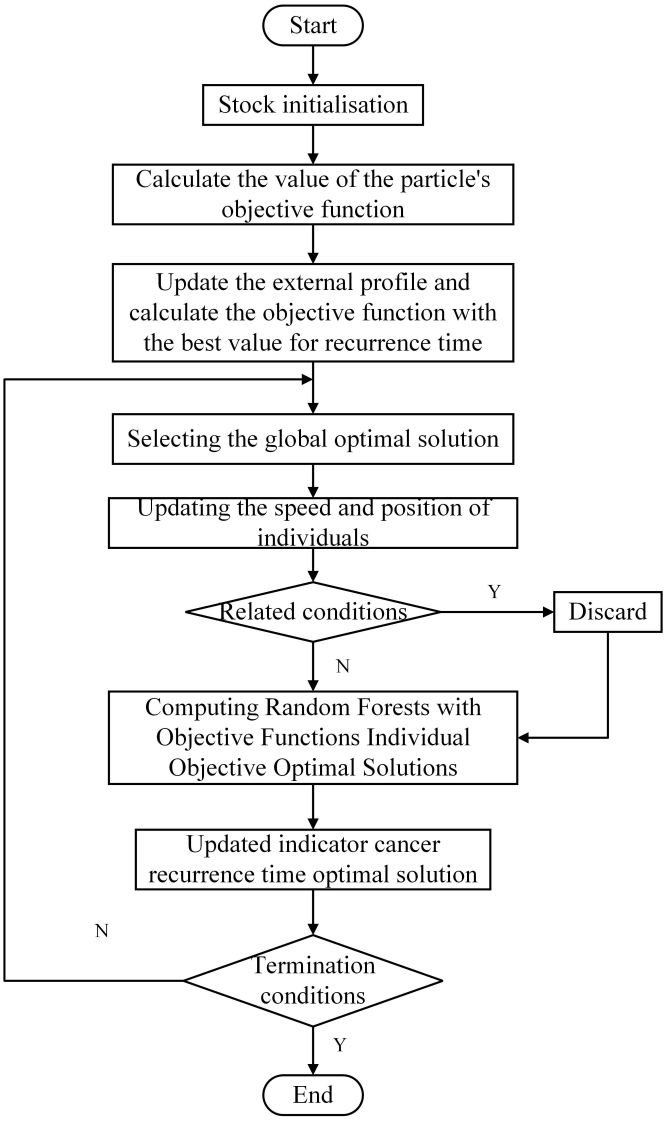

FIGURE 2. MOPSO-RF flow chart.

4. Experiments and results analysis

4.1 Data and pre-processing

Based on the long-term medical consensus, the 2018 ASC Carcinogenic Factors Study [30], and cancer assessment data from TIES.IO, there were multiple factors influencing the time to cancer recurrence. Twelve indicators were selected as a sample set to predict the time to recurrence via pre-processing of breast cancer data. The sample set had two categories, one related to self: gender, age, basal score, tumour score, immune score, and psychological score; and the other to external factors: basal nutrition score, nutrition comparison score, safe intake score, microenvironment score, total nutrition score and aerobic exercise score. The nutritional support data were pre-processed and time to relapse ranged from 6 to 60 months by considering that the time within 6 months might be in the treatment period. The cases were thus not considered as valid data where relapse was within 6 months. The age at cancer patients' surgery covered a wide range of ages. Cancer patients in 25 to 65 years of age were considered for the study as the model predictions were biased below 25 years age and other systemic diseases of above 65 years age could affect the patient data. Data from cancer patients were collected and scored. The correlation was studied using statistical correlation coefficients: Pearson correlation coefficient [31], and Spearman Rank correlation coefficient [32] for the sample input indicators.

Twelve input variables and recovery time were chosen to conduct the correlation analysis for dimensionality reduction, and to get the correlation between each variable and the recovery time. As in Table 2, the indexes were positively correlated with recurrence time for twelve input indexes, however, overall correlation factor was not remarkably high, wherein the age at surgery and gender had extremely low correlation coefficients with the recurrence time.

Spearman Rank correlation coefficient was not as stringent as the Pearson correlation coefficient regarding data conditions. Spearman Rank correlation coefficient could be employed regardless of the overall distribution pattern of two variables and the sample size if the observations of two variables were paired rank-rated information, or rank information obtained by transforming observations of continuous variables. Spearman's rank correlation coefficients and each of the sample input indicators are shown in Table 3.

TABLE 2. Pearson correlation coefficients between sample input indicator and time to recurrence.

Indicators	Coefficients
Tumour score	0.164
Base score	0.141
Total nutrition score	0.130
Immune score	0.117
Safe intake score	0.114
Nutritional comparison score	0.080
Psychological score	0.064
Aerobic exercise score	0.051
Basic nutrition score	0.041
Micro-environment score	0.040
Age of surgery	0.020
Gender	0.005

TABLE 3. Spearman correlation coefficients between sample input indicator and time to recurrence.

Indicators	Coefficients
Tumour score	0.175
Base score	0.143
Total nutrition score	0.130
Immune score	0.116
Safe intake score	0.113
Nutritional comparison score	0.084
Psychological score	0.071
Aerobic exercise score	0.058
Basic nutrition score	0.040
Micro-environment score	0.039
Age of surgery	0.011
Gender	0.006

The comparison of Pearson and Spearman in Tables 2 and 3 depicted that the final relevance coefficients were not identical, however they had the same overall features for twelve input indicators. It was seen that the ranked relationship between twelve input indicators and correlation was certain before the output, *i.e.*, the tumour score had the greatest impact on time to recurrence, while age at surgery and gender had small impact. The two items of age at surgery and gender could be disregarded when there were too many indicators during a follow-up study. It was found by combining Tables 2 and 3 that there were six indicators with the largest coefficients and the highest correlation with the time to breast cancer recurrence. These first-level six indicators were thus used in the manuscript. The remaining six indicators had relatively small correlation with the time to breast cancer recurrence and not calculated.

The above analyses led to the identification of six level 1 indicators: tumour, immune, nutritional, psychological, microenvironmental, and aerobic exercise and advanced work; and 45 secondary indicators where each secondary indicator had maximum weight of 10 and minimum of 1. According to the secondary indicators and their weights, the primary indicators were calculated together with the calculation formula shown in Eqn. 6.

$$I = \sum_{i=1}^{n} x_i w_i / \sum_{i=1}^{n} w_i \tag{6}$$

Where x_i is the value of ith and w_i is the weight of ith second level indicator under the first level indicator. The correlations for each indicator score and their weights are provided in Table 4. The weights were given based on the experience of physician experts.

Cancer recurrence after the treatment affects long-term survival of cancer patients. Studying the factors affecting recurrent cancer and adopting clinical interventions can improve the survival rate of cancer patients. About 90% of cancer deaths are caused by primary tumour cells that have moved far from their original location. Metastasis is an inefficient process; however it causes most cancer-related deaths because millions of cells can leave the tumour each day. The growth of cancer cells at distant sites is likely to occur at some point even if only a small percentage of cells leaving the tumour survive to form a new tumour. The study of cancer recurrence time is thus important for clinical diagnosis and cancer treatment. In this manuscript, the historical case data of 1149 breast cancer patients were collected through the Cancer Data Analysis Laboratory with Stre in Beijing. The data were evaluated, quantified, and correlated with six influencing factors on time to recurrence, as shown in Fig. 3.

In Fig. 3, Imm: Immune indicators, Tum: Tumour indicators, Mic: Microenvironmental indicators, Psy: Psychological indicators, Nut: Nutritional indicators, Aer: Aerobic exercise and advanced work indicators, and Tim: Time to recurrence after breast cancer surgery. The relationship between indicators can be found by analysing the correlation between these indicators and the breast cancer, such as tumour indicator and other indicators: a positive correlation with the immune indicator as the tumour growth may inhibit the immune system function; a negative correlation with nutritional indicator as

TABLE 4. Indicators for cancer patients.

Indicators	Related terms and weights
Immune indicators	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).
Tumour indicators	Size (10); Placeholder (10); Violate the relationship (10); Angiogenesis (10); Pathological typing (3); CTC value (9); Differentiation (10); Mutation target (1).
Nutritional indicators	Total nutrition (6); Balanced nutrition (3); Nutrition safety assessment (5); Cancer cell proliferation (10); Immune cell proliferation (10); Angiogenesis (8); Amino acids evaluation (5); Proteomics evaluation (10).
Psychological indicators	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).
Microenvironmental indicators	O_2 (3); pH value (4); Interstitial pressure (2); Inflammatory response (7); Vascular permeability (6); CTC value (9); Proteomics analysis (8).
Aerobic exercise and advanced work indicators	Aerobic exercise (4); Advanced social work (3); Texas Social Behavior Questionnaire (3).

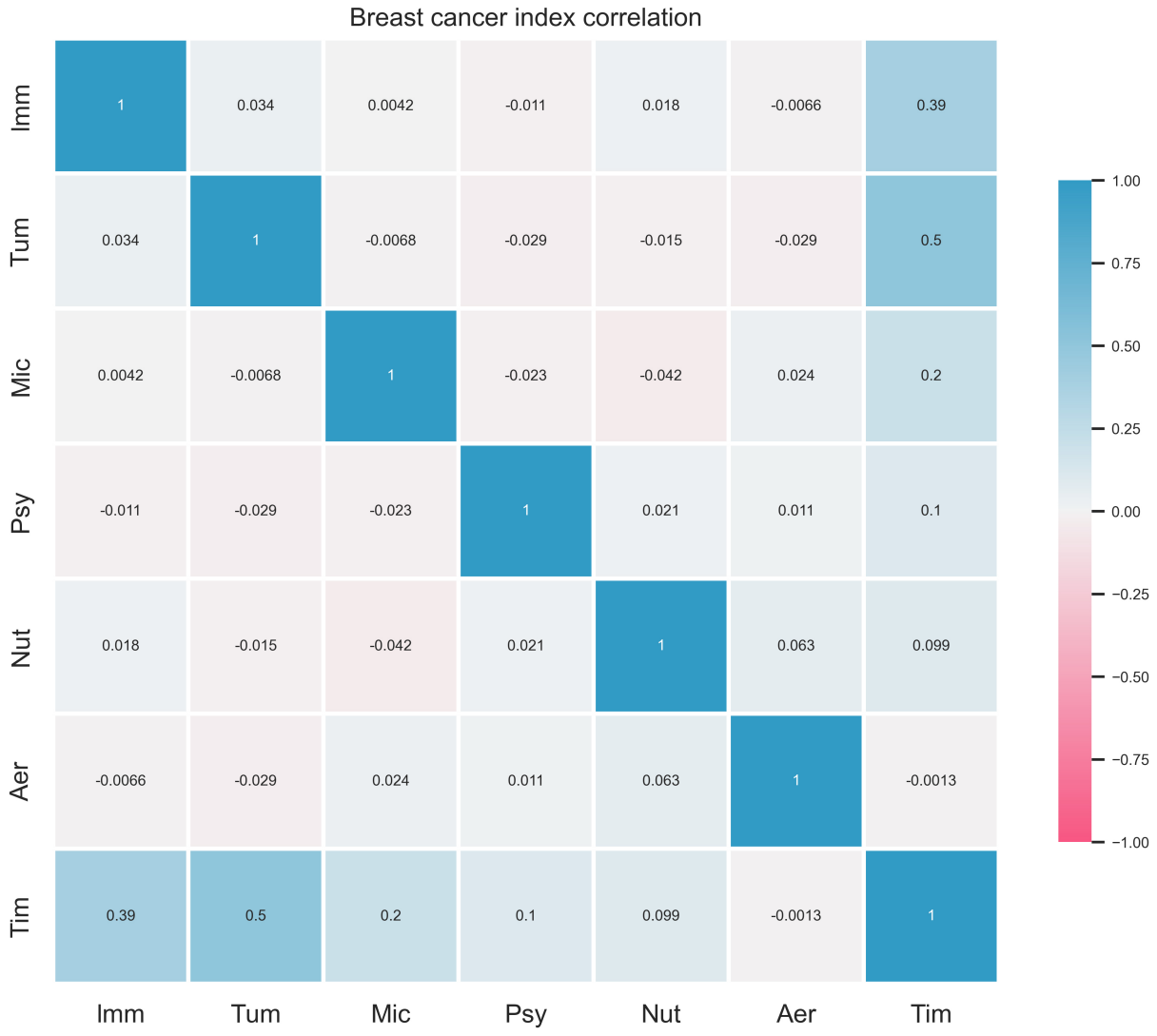

FIGURE 3. Breast cancer index correlation.

the tumour growth requires nutritional support; a negative correlation with psychological indicator as the psychological stress caused by tumour may affect psychological state; negative correlation with microenvironmental indicators as the tumours may affect surrounding microenvironment; and negative correlation with aerobic exercise and advanced work indicators because tumours may disrupt bodily functions and cause body damage. Therefore, the degree of correlation between its indicators and the degree of association with the time to breast cancer recurrence are studied herein.

4.2 Comparative experiments

In this manuscript, the accuracy was selected to evaluate the indicators for model performance. The confusion matrix was defined as shown in Table 5. The study employed six composite modular indicators affecting the patients recovering from breast cancer as input, and time to breast cancer recurrence as output to construct a model for predicting the time to breast cancer recurrence. The time to breast cancer recurrence was defined as 60 months, and the patients were designated as healthy or unhealthy around this boundary. The patients with recurrence time of over 60 months were recorded as healthy, and below 60 months as unhealthy. This study targeted the intervention for unhealthy breast cancer patients. MOPSO-RF model predicted the recurrence time of breast cancer. The predicted value within ± 6 months was regarded as the reasonable range. This was the permissible error range and recorded as Error Rate (ER).

TABLE 5. Confusion matrix.

	TRUE	Predictions
Time	Real Time (RT)	Predicted Time (FT)
Number	Real Number (RN)	Predicted Number (PN)

Accuracy is one of the measures for good model results and the model accuracy for this experiment is calculated as in Eqn. 7.

$$Accuracy = \begin{cases} 1 & , \quad |FT - RT| \leq ER \\ \frac{RN-PN}{RN} 100\% & , \quad |FT - RT| > ER \end{cases} \tag{7}$$

The model accuracy is calculated in a way that no calculation is done for those in the error range. Only the error rate outside error range is calculated and accuracy is derived. The model is compared to attain at the best model for optimisation.

GBDT, LR (Linear Regression), LR (Logistic Regression), SVM (Support Vector Machine), XGBoost and RF algorithms were selected for the comparison experiments. In the dataset, 80% of samples were used as the training set and 20% as the test set with the target output type of the model being recurrence time. The equipment model used for the experiments was Intel i7 processor (3.40 GHz), 16 GB RAM, GPU 3060, 12 Gigabytes, Window 10, and Python 3.8. Results of the breast cancer prediction models for six classes of algorithms are shown in Fig. 4A–F.

The experimental results are shown in Table 6 as derived from the comparison of six algorithm types in plots (A) to (F) of Fig. 4. The results further strengthen the choice of RF model for breast cancer.

The prediction of breast cancer recurrence time was made by employing six machine learning algorithms and their accuracies. It was found that the conventional and machine learning algorithms for predicting breast cancer recurrence time had good accuracy, however not up to the mark as desired. The integrated learning algorithm Random Forest had the highest accuracy compared to others. Its accuracy was related to the number (n_estimators) and maximum depth (max_depth) of the decision trees. So, the Random Forest (RF) was optimized.

4.3 Experimental results and analysis

MOPSO was constructed after comparing the machine learning for predicting recurrence time after breast cancer surgery. After initialising the population, the random forest with particle swarms was optimised by setting c_1 to 2.4, c_2 to 2.2 and w to 0.90. The fitness function was defined with the individual extremes being the optimal solutions found for each particle, and then finding a global value from these optimal solutions called the current global optimum. This was compared with the historical global optimum and updated, and the results shown in Fig. 5.

The multi-objective optimisation particle swarms were used after finding the parameters of particle swarm optimisation for random forests. MOPSO algorithm was employed to search for the relatively better parameters of RF model under test dataset to improve the predictive ability of RF regression tree model. It was found in the process of analysis that the number of particles in depth of the decision tree was always converged to more than 8. For more training of this result, the total number of multi-objective particles was set to 200, and the number of iterations to 500. After the experiments, the MOPSO algorithm search process of initial state and the end of iterative state, as shown in Fig. 6, resulted in the optimal Random Forest Decision Tree, the number of decision trees, and the decision tree depth.

As shown in Fig. 6, MOPSO generated number of random particles in the initial state where each represented a set of decision trees (n_estimators) and maximum depth (max_depth). They were updated iteratively to determine the relatively better parameters. Fifty trials were performed under the set conditions and average was taken as the final result. The manuscript after compiling the results of many experiments exhibited the best value wherein the number of decision trees (n_estimators) was 412 and the maximum depth of decision tree (max_depth) was 9. The random forest model performance under these parameters was relatively better.

MOPSO-RF experiment was started according to the parameter settings in Fig. 6 to verify whether the parameters of breast cancer recurrence time prediction were reasonable. It was found that RF accuracy after MOPSO was improved for the prediction of time to breast cancer recurrence, as shown in Table 7.

After the experiments, MOPSO-RF was used to predict the recurrence time after breast cancer surgery, however the degree of correlation between each index and recurrence time was different. The multi-objective particle swarm algorithm was used to optimize the variables in RF model for predicting

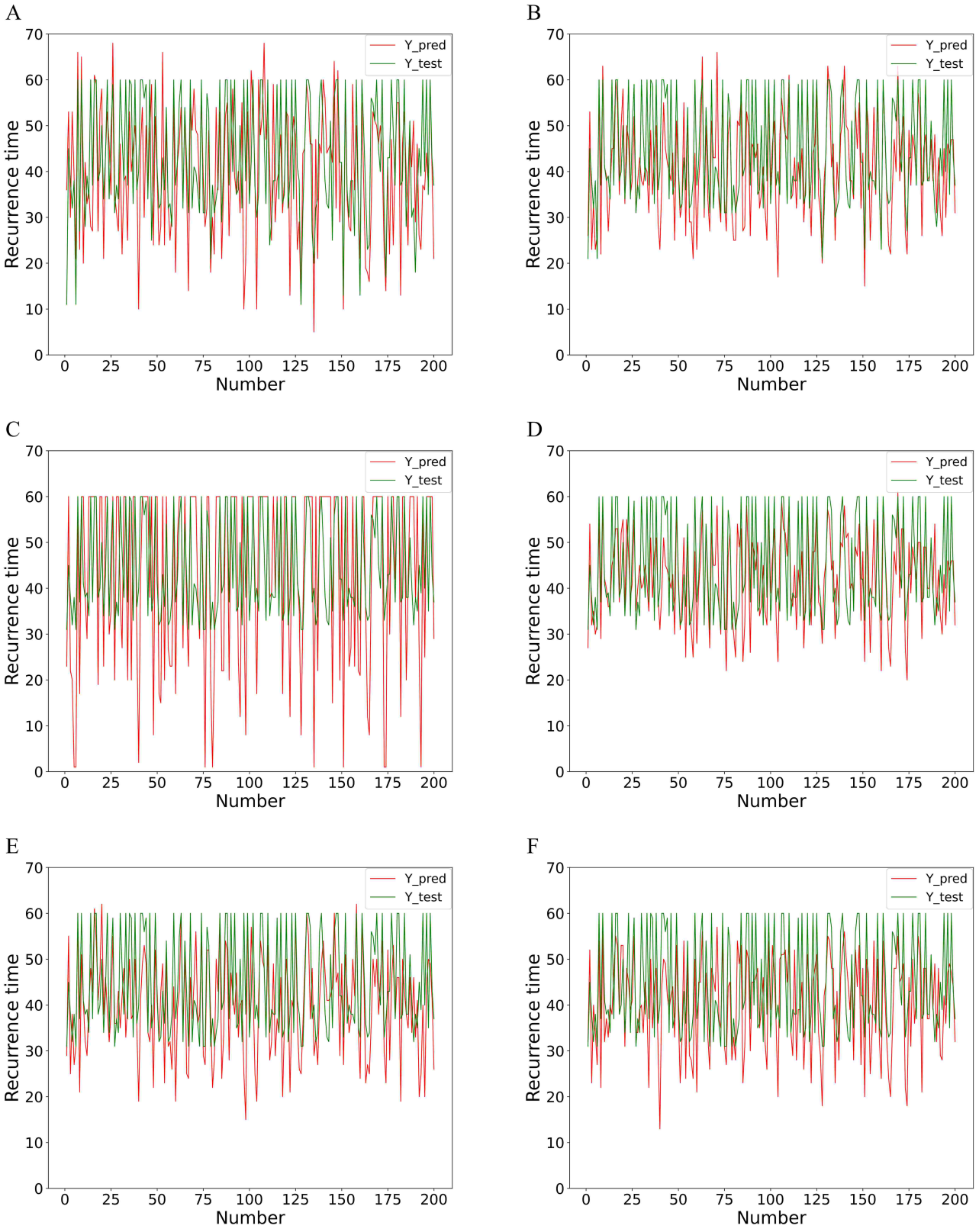

FIGURE 4. Experimental diagrams of model comparison. (A) GBDT; (B) LinearRegression; (C) LogisticRegression; (D) SVM; (E) XGBoost; (F) Random Forest.

TABLE 6. Six types of algorithmic model accuracy.

	GBDT	Linear Regression	Logistic Regression	SVM	XGBoost	RF
Breast cancer	71.74%	79.83%	53.48%	83.70%	85.65%	88.70%

GBDT: Gradient Boosted Tree; SVM: Support Vector Machine; RF: random forest.

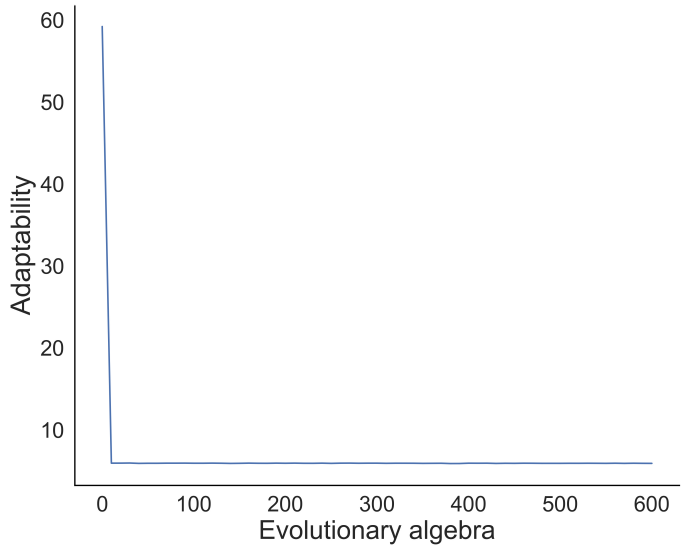

FIGURE 5. Particle swarm optimization of random forests.

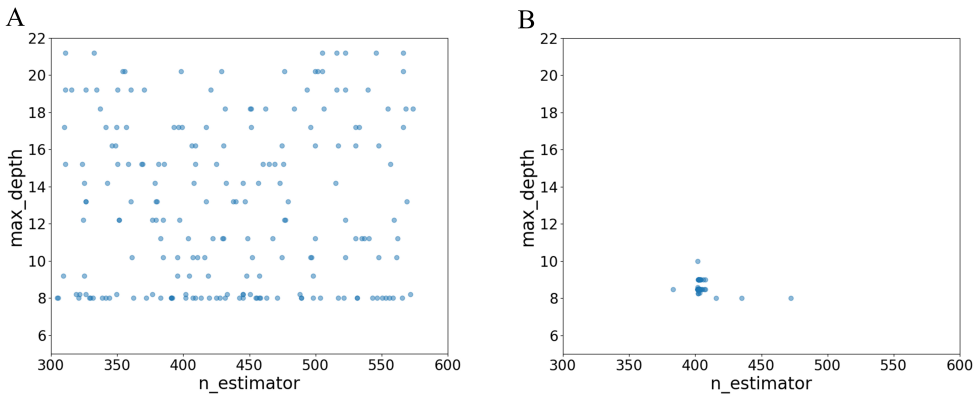

FIGURE 6. Multi-objective particle swarm optimization for random forests. (A) Iterating over the initial state; (B) End-of-iteration state.

TABLE 7. Algorithm comparison.

	RF	MOPSO-RF
Breast cancer	88.70%	92.17%

RF: random forest; MOPSO-RF: multi-objective particle swarm optimised random forest.

the correlation between each index and recurrence time. The schematic diagrams were made for the optimizations of each objective function composed of six types of indicators of cancer patients and the recurrence time of cancer. The results of algorithm search as shown in Fig. 7 included the effect of each indicator on the recurrence time as analysed according to the objective functions (A) to (F). The size of each point was the model accuracy in predicting the time with smaller points representing higher accuracy and red points representing errors.

As shown in Fig. 7, the algorithm had a strong search capability. The indicators represented between each objective function and the cancer recurrence time were not directly related. There was a conflicting relationship where one objective function of cancer indicator became better at the expense of another which turned worse. The objective function values of all cancer indicators could not be made optimal at the same time. The six level 1 indicators were the biggest factors affecting the recurrence time after breast cancer surgery within the error running range. They could provide suggestions to the doctors for recurrence time of breast cancer after surgery. Moreover, they could assist in suggesting nutritional support to cancer patients after surgery for slowing down the time to cancer recurrence.

It was found through experiments that each index had different optimal value for predicting the recurrence time after breast cancer surgery. There were many Pareto optimal solutions in the objective function represented by each index. This manuscript screened the Pareto optimal solution scheme to achieve concise and clear optimization results. The screening condition was imposed in a way that the objective function values of influences of six indicators on recurrence time were better than the objective function values of initial scheme. The typical screening scheme was selected from all the Pareto optimal solutions that met the screening conditions, as shown in Table 8.

Compared to the traditional PSO algorithms, this manuscript

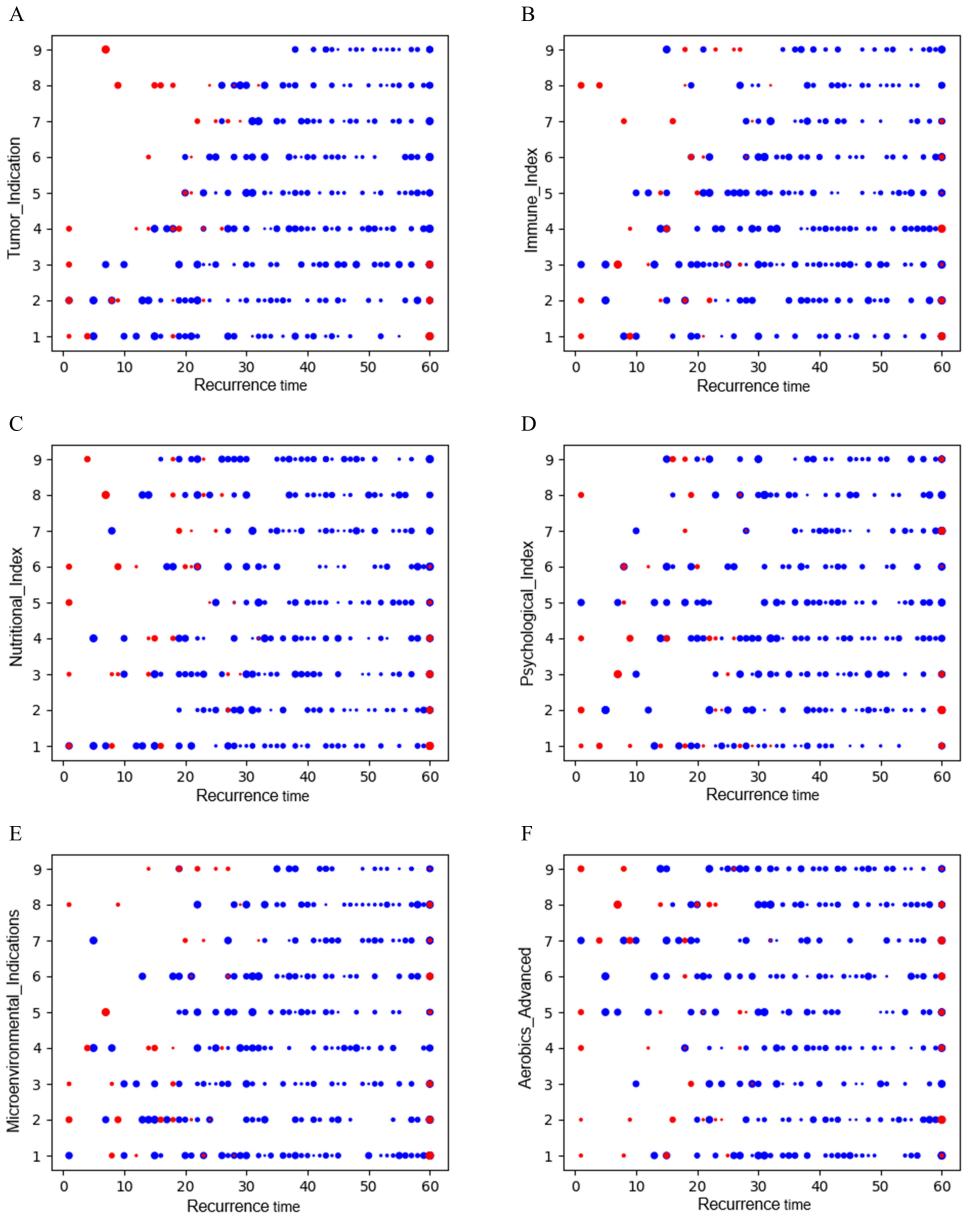

FIGURE 7. Optimisation results of each indicator function. (A) Tumour indicators; (B) Immune indicators; (C) Nutritional indicators; (D) Psychological indicators; (E) Microenvironmental indicators; (F) Aerobic exercise and advanced work targets.

TABLE 8. Screening program.

	Tumour indicators	Immune indicators	Nutritional indicators	Psychological indicators	Microenvironmental indicators	Aerobic exercise and advanced work indicators
Range of values	40–60	40–60	40–60	40–60	35–55	35–55
Screening program	50	45	56	55	49	47

used MOPSO algorithm to optimise the random forests. Multiple optimisation variables could be optimised for multiple objective functions. They used an external archive for storing updated non-inferior solution sets and used a shrinkage factor to improve the search capability and convergence speed of the algorithm. The optimisation results were significant compared to the initial solution. The multi-objective optimisation approach helped doctors to provide reasonable treatments for breast cancer patients in post-operative rehabilitation diagnosis. In combination with Pareto solution set for each indicator, multi-objective optimisation could be achieved for breast cancer patients to predict the time to cancer recurrence. In the study of breast cancer recurrence time with MOPSO-RF model, the maximum value of tumor indicators was optimized the most followed by the immune indicators and nutritional indicators, and the least was for microenvironmental indicators, indicating that the tumor indicators had the greatest influence on recurrence time of breast cancer. It was thus necessary

to focus on the influence of tumor and immune indicators on the optimization target of breast cancer recurrence time in the subsequent optimization analysis. A reasonable optimisation was required based on the symptoms of indicators for the post-operative recovery of cancer patients. Different cancers possessed different indicators for focusing and thus improvements were crucial in all such areas.

5. Conclusions

Cancer is a malignant disease that is difficult to completely cure. Cancer management is limited in improving the life quality and prolonging the life span of patients. The emergence of data mining can lead to the development of cancer diagnosis and treatment. This manuscript is aimed at the investigation of machine learning algorithmic models in data mining to achieve accurate prediction of cancer recurrence time. Firstly, the basic content of cancer research and related studies is introduced, followed by constructing a set of indicators affecting the cancer recurrence through physical health assessment of cancer patients by expert doctors. The gradient boosting tree (GBDT), support vector machine (SVM), linear/logistic regression (LR), XGBoost and random forest (RF) algorithm models are compared. It is concluded that the prediction accuracy of RF algorithm is higher than the other models. The particle swarm optimization random forest model is employed to find the optimal number of decision trees and decision tree depth. Multi-objective optimisation optimises six types of cancer indicators and recurrence time and finds the effect of each indicator on the cancer recurrence time. Optimising one indicator alone without considering the others can result in one indicator receiving the effect of others. The best result of multi-objective optimisation is to eliminate the effect of other indicators, and to find the optimal Pareto solution of each objective function which can predict the time to cancer recurrence for each indicator, and help doctors to provide accurate treatment plans for breast cancer patients after surgery. The time to cancer recurrence for breast cancer patients can be indexed by each indicator with precision. The results show that MOPSO-RF model based on this manuscript is suitable for predicting the time to recurrence of breast cancer and improving the survival rate of patients through appropriate and timely clinical interventions.

Manuscript experiments are purely applied to the prediction of time to recurrence after breast cancer surgery. There will be some limitations for other cancers, and the correlation of each index is different for different cancers. The influencing factors are also different. In future work, the dataset will continue to be optimised and will subsequently improve the algorithm to enhance the model accuracy. Moreover, various metrics for each cancer will be optimised which can guide the clinicians in choosing the right treatment plan for cancer patients. This will further improve their survival time and life quality.

AVAILABILITY OF DATA AND MATERIALS

The data that supports the findings of this study are available from Beijing Sitai Rui Health Technology Co., Ltd but restrictions apply to the availability of these data, which were used under license for the current study, and so are not publicly available. Data are however available from the authors upon reasonable request and with permission of Beijing Sitai Rui Health Technology Co., Ltd.

AUTHOR CONTRIBUTIONS

JW—writing-first draft, algorithm analysis, experimental code; HL—writing-review and editing, data editing; SJY—formal analysis, software; FCL—data collation, verification; AMY—supervision, project management, funding acquisition; DBH—resources, data centers.

ETHICS APPROVAL AND CONSENT TO PARTICIPATE

The study was approved by the Shanxi Provincial People's Hospital (12140000405704516B), Taiyuan, China. All patients provided written informed consent.

The study was approved by The Fourth Affiliated Hospital of Hebei Medical University (121300004017003821), Shijiazhuang, China. All patients provided written informed consent.

Participants:

1. The clinical data of this study were collected at the two hospitals mentioned in the ethics statement.
2. The data pre-processing process of this research was carried out in Beijing Stairui Health Technology Version 07 June 2023 submitted to Journal Not Specified 17 of 18 Co., Ltd. (TIES.IO).
3. The construction of the research model and the realization of the algorithm were completed by the Tangshan Key Laboratory of Engineering Computing (North China University of Science and Technology.

ACKNOWLEDGMENT

Here, I would like to express my gratitude to Aimin Yang for providing the experimental platform and providing guidance throughout the entire experimental process. I would also like to express my gratitude to Shujuan Yuan for helping me complete my manuscript writing. I would also like to express my gratitude to Hao Li for helping me complete the experiment and organize the data together. I would also like to express my gratitude to Fengchun Liu. Thank you very much for the experimental data provided by Dianbo Hua.

FUNDING

1. This research was funded by Key Science and Technology Project of Hebei Provincial Department of Education (North China University of Science and Technology, Project Number: JYG2020001);
2. This research was funded by Natural Science Foundation of Hebei Province (North China University of Science and Technology, Project Number: E2021209024).

CONFLICT OF INTEREST

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

REFERENCES

[1] Siegel RL, Miller KD, Fuchs HE, Jemal A. Cancer statistics, 2022. CA: A Cancer Journal for Clinicians. 2022; 72: 7–33.

[2] Xia C, Dong X, Li H, Cao M, Sun D, He S, *et al.* Cancer statistics in China and United States, 2022: profiles, trends, and determinants. Chinese Medical Journal. 2022; 135: 584–590.

[3] Mohri M, Rostamizadeh A, Talwalkar A. Foundations of machine learning. MIT press: Cambridge. 2018.

[4] Ilyas QM, Ahmad M. An enhanced ensemble diagnosis of cervical cancer: a pursuit of machine intelligence towards sustainable health. IEEE Access. 2021; 9: 12374–12388.

[5] Lohani BP, Thirunavukkarasan M. A review: application of machine learning algorithm in medical diagnosis. 2021 International Conference on Technological Advancements and Innovations (ICTAI). Tashkent, 10–12 November 2021. IEEE: Tashkent, Uzbekistan. 2021.

[6] Waks AG, Winer EP. Breast cancer treatment. JAMA. 2019; 321: 288.

[7] Mohsin MY, Ali MR, Yousif M, Chaudhary ST, Tahir W, Wattoo WA. Accuracy improvement for the diagnosis of breast cancer using different techniques of machine learning. 2022 International Conference on Emerging Trends in Smart Technologies (ICETST). Karachi, 23–24 September 2022. IEEE: Karachi, Pakistan. 2022.

[8] Cheville A, Smith S, Barksdale T, Asher A. Cancer rehabilitation. In David X. Cifu (ed.) Braddom's physical medicine and rehabilitation (pp. 568–593). 6th edn. Elsevier: Richmond. 2021.

[9] Ma Z, Zhang M, Liu J, Yang A, Li H, Wang J, *et al*. An assisted diagnosis model for cancer patients based on federated learning. Frontiers in Oncology. 2022; 12: 860532.

[10] Alghunaim S, Al-Baity HH. On the scalability of machine-learning algorithms for breast cancer prediction in big data context. IEEE Access. 2019; 7: 91535–91546.

[11] Yue W, Wang Z, Chen H, Payne A, Liu X. Machine learning with applications in breast cancer diagnosis and prognosis. Designs. 2018; 2: 13.

[12] Kaur M, Gianey HK, Singh D, Sabharwal M. Multi-objective differential evolution based random forest for e-health applications. Modern Physics Letters B. 2019; 33: 1950022.

[13] Qi C, Chen Q. Evolutionary random forest algorithms for predicting the maximum failure depth of open stope hangingwalls. IEEE Access. 2018; 6: 72808–72813.

[14] Adnan MN, Islam MZ. Optimizing the number of trees in a decision forest to discover a subforest with high ensemble accuracy using a genetic algorithm. Knowledge-Based Systems. 2016; 110: 86–97.

[15] Kabiraj S, Raihan M, Alvi N, Afrin M, Akter L, Sohagi SA, *et al*. Breast cancer risk prediction using XGBoost and random forest algorithm. 2020 11th international conference on computing, communication and networking technologies (ICCCNT). Kharagpur, 01–03 July 2020. IEEE: Kharagpur, India. 2020.

[16] Yifan D, Jialin L, Boxi F. Forecast model of breast cancer diagnosis based on RF-AdaBoost. 2021 International Conference on Communications, Information System and Computer Engineering (CISCE). Beijing, 14–16 May 2021. IEEE: Beijing, China. 2021.

[17] Ranjan M, Shukla A, Soni K, Varma S, Kuliha M, Singh U. Cancer prediction using random forest and deep learning techniques. 2022 IEEE 11th International Conference on Communication Systems and Network Technologies (CSNT). Indore, 23–24 April 2022. IEEE: Indore, India. 2022.

[18] Huang Z, Chen D. A breast cancer diagnosis method based on VIM feature selection and hierarchical clustering random forest algorithm. IEEE Access. 2021; 10: 3284–3293.

[19] Yang AM, Han Y, Liu CS, Wu JH, Hua DB. D-TSVR recurrence prediction driven by medical big data in cancer. IEEE Transactions on Industrial Informatics. 2020; 17: 3508–3517.

[20] Arya N, Saha S. Multi-modal classification for human breast cancer prognosis prediction: proposal of deep-learning based stacked ensemble model. IEEE/ACM Transactions on Computational Biology and Bioinformatics. 2022; 19: 1032–1041.

[21] Kutrani H, Eltalhi S. Decision tree algorithms for predictive modeling in breast cancer treatment. 2022 IEEE 2nd International Maghreb Meeting of the Conference on Sciences and Techniques of Automatic Control and Computer Engineering (MI-STA). Sabratha, 23–25 May 2022. IEEE: Sabratha, Libya. 2022.

[22] Alkebsi K, Du W. A fast multi-objective particle swarm optimization algorithm based on a new archive updating mechanism. IEEE Access. 2020; 8: 124734–124754.

[23] Rajagopal A, Joshi GP, Ramachandran A, Subhalakshmi RT, Khari M, Jha S, *et al*. A deep learning model based on multi-objective particle swarm optimization for scene classification in unmanned aerial vehicles. IEEE Access. 2020; 8: 135383–135393.

[24] Juweid ME, Cheson BD. Positron-emission tomography and assessment of cancer therapy. New England Journal of Medicine. 2006; 354: 496–507.

[25] Knobf MT, Ferrucci LM, Cartmel B, Jones BA, Stevens D, Smith M, *et al*. Needs assessment of cancer survivors in Connecticut. Journal of Cancer Survivorship. 2012; 6: 1–10.

[26] Cromes Jr GF. Implementation of interdisciplinary cancer rehabilitation. Rehabilitation Counseling Bulletin. 1987; 21: 230–237.

[27] Pathak S, Mishra I, Swetapadma A. An assessment of decision tree based classification and regression algorithms. 2018 3rd International Conference on Inventive Computation Technologies (ICICT). Coimbatore, 15–16 November 2018. IEEE: Coimbatore, India. 2018.

[28] Kennedy J, Eberhart R. Particle swarm optimization. Proceedings of ICNN'95-international conference on neural networks. Perth, 27 November 1995–01 December 1995. IEEE: Perth, Australia. 1995.

[29] Eberhart RC, Shi Y. Tracking and optimizing dynamic systems with particle swarms. Proceedings of the 2001 congress on evolutionary computation (IEEE Cat. No. 01TH8546). Seoul, 27–30 May 2001. IEEE: Seoul, Korea (South). 2001.

[30] Bray F, Ferlay H, Soerjomataram I, Siegel RL, Torre LA, Jemal A. Global cancer statistics 2018: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. CA: A Cancer Journal for Clinicians. 2018; 68: 394–424.

[31] Mujahid AK, Thirumalai C. Pearson correlation coefficient analysis (PCCA) on adenoma carcinoma cancer. 2017 International Conference on Trends in Electronics and Informatics (ICEI). Tirunelveli, 11–12 May 2017. IEEE: Tirunelveli, India. 2017.

[32] Scherer P, Trębacz M, Simidjievski N, Viñas R, Shams Z, Terre HA, *et al*. Unsupervised construction of 558 computational graphs for gene expression data with explicit structural inductive biases. Bioinformatics. 2022; 38: 1320–1327.

How to cite this article: Jian Wang, Hao Li, Shujuan Yuan, Fengchun Liu, Aimin Yang, Dianbo Hua. Breast cancer recurrence time prediction based on the MOPSO-RF model. European Journal of Gynaecological Oncology. 2025; 46(5): 79-91. doi: 10.22514/ejgo.2025.068.