Submitted: 29 June, 2022 Accepted: 02 August, 2022 Published: 15 December, 2022 DOI:10.22514/ejgo.2022.056

ORIGINAL RESEARCH

European Journal of Gynaecological Oncology

# Application and comparison of several machine learning methods in the prognosis of cervical cancer

Yawen Ling[1,†], Weiwei Zhang[2,†], Zhidong Li[1], Xiaorong Pu[1], Yazhou Ren[1,3,*]

[1]School of Computer Science and Engineering, University of Electronic Science and Technology of China, 611731 Chengdu, Sichuan, China
[2]Cancer prevention and treatment institute of Chengdu, Department of oncology, Chengdu Fifth People's Hospital/The Second Clinicical Medical College, Affiliated Fifth People's Hospital of Chengdu University of Traditional Chinese Medicine, 611137 Chengdu, Sichuan, China
[3]Institute of Electronic and Information Engineering of UESTC in Guangdong, 523808 Dongguan, Guangdong, China

***Correspondence**
yazhou.ren@uestc.edu.cn
(Yazhou Ren)

† These authors contributed equally.

## Abstract

Accurate prognosis of cervical cancer in the clinical setting is challenging because of the complexity of the causative factors. Considering the drawbacks of the widely used Cox proportional hazards model, such as the inability to fully use the information and the possible failure to achieve the best fit, several new attempts based on machine learning have been developed to find better prognostic prediction models. However, the application of these attempts is often limited, because they often rely on public databases. Therefore, for cervical cancer, there is a need to explore the value of machine learning in terms of its practical application in prognostic prediction. In this study, we introduced several machine learning methods including k-nearest neighbors (KNN), decision tree (DT), logistic regression (LR), support vector machine (SVM), random forest (RF), extreme gradient boosting (XGBoost) and light gradient boosting machine (LightGBM) to predict the survival of patients by using the real-world pathological data of 216 patients collected from the Fifth People's Hospital of Chengdu. The experimental results showed that these methods have a promising application value in the prediction of overall survival (OS) of patients with cervical cancer (KNN: F1-score = 0.95, Accuracy = 0.93, DT: F1-score = 0.94, Accuracy = 0.92, LR: F1-score = 0.92, Accuracy = 0.90, SVM: F1-score = 0.94, Accuracy = 0.92, RF: F1-score = 0.96, Accuracy = 0.95, XGBoost: F1-score = 0.96, Accuracy = 0.95, LightGBM: F1-score = 0.96, Accuracy = 0.95). Moreover, XGBoost and LightGBM gave the importance of the clinical indicators associated with cervical cancer, whose correlation with OS and progression-free survival (PFS) can be further obtained. Thus, the predictors of OS and PFS were successfully identified. Finally, the results were confirmed by the Cox proportional hazards model. These results indicated that machine learning methods can accurately predict the OS of patients with cervical cancer. Moreover, the methods can be used to analyze the correlation between clinical indicators and OS or PFS to help doctors make more accurate decisions in a clinical setting.

## Keywords

Cervical cancer; Machine learning; Prognosis; Overall survival; Progression-free survival

## 1. Introduction

Cervical cancer is one of the four most common gynecological cancers diagnosed globally and a significant threat to women's health [1]. Although cervical cancer can be effectively prevented by vaccination against human papillomavirus (HPV) [2], its incidence is still high in developing countries [3]. On the other hand, difficulties in organizing cervical cancer screening have led to low screening rates in developed regions [4]. The prognosis of cervical cancer is crucial for follow-up treatment. Currently, the clinical treatment for cervical cancer is mainly based on its stage and this approach has some limitations and is debatable. For example, whether patients in the IB1 stage need to continue treatment after surgery needs to be determined on the basis of postoperative high-risk factors. Although several indicators such as miR-216b level [5] and overexpression of microRNA-944 [6] have been reported to have a prognostic value for cervical cancer, detecting these indicators in body fluids is generally difficult and expensive. Therefore, we urgently need effective, cheap, and convenient-to-detect indicators for the accurate prognosis of patients with cervical cancer.

At present, the Cox proportional hazards model is widely used for predicting prognosis. The Cox model relies on the proportional risk assumption, but this assumption is usually violated in practice, leading to a possible underestimation or overestimation of the average relative risk [7]. In addition, because of its inherent linear assumption, it cannot reflect the nonlinear analysis of realistic clinical characteristics [8],

and thus, has limitations such as the inability to fully utilize the information and achieve the best fit. Therefore, finding better solutions for the accurate prognosis of cervical cancer is essential. Nowadays, modern medical and computer technologies are developing rapidly. Thanks to technological breakthroughs, machine learning is being widely used and has shown promise in dealing with various complex problems as well as demonstrates the ability that is close to or even surpasses human capabilities. Thus, there have been several attempts to use machine learning for medical requirements. For example, for a challenging problem such as whole slide image (WSI) classification for lung cancer, Wang *et al.* [9] used a patch-based fully convolutional network (FCN) to extract depth features and then used random forest (RF) for effective prediction. Liu *et al.* [10] proposed and validated a prognostic model for breast cancer based on extreme gradient boosting (XGBoost) and Cox proportional hazards model, which provides an essential means for clinical diagnosis and treatment for improving patient survival. To address the new problem of prostate cancer in China, Zhang *et al.* [11] built a decision tree (DT) based on prostate characteristics to help screen patients with prostate cancer. The addition of deep learning delivers impressive performance, Mansour designed a convolution neural network (CNN) based breast cancer detection model that outperforms other country-of-artwork techniques [12]. The development of ensemble deep-learning-enabled clinical decision support system for breast cancer diagnosis and classification (EDLCDS-BCDC) [13] and social engineering optimization with deep transfer learning-based breast cancer detection and classification (SEODTL-BDC) [14] techniques has also brought in significant help for biomedical image processing. However, deep neural networks often imply expensive computational overhead, and it is a black box for doctors whose interpretability remains a problem. On the other hand, some commonly used machine learning methods such as DT and support vector machine (SVM) have stronger interpretability and they are still promising for providing help with the decision-making process of doctors.

For cervical cancer prognosis prediction, the DT algorithm was applied on a public dataset of the University of California by Alam *et al.* [15]. Wu *et al.* [16] successfully used the SVM algorithm for the diagnosis and classification of malignant cervical cancer samples. Ijaz *et al.* [17] used RF as a classifier to classify potential patients and differentiate between the prognostic factors of cervical cancer after removing outliers from the data set and balancing the number of cases. Deng *et al.* [18] introduced three methods, SVM, XGBoost and RF, to diagnose the "Cervical Cancer Behavior Risk Data Set" from the "University of CaliforniaIrvine (UCI) Machine Learning Repository" and reported that XGBoost and RF have a better performance than SVM in their experiments. Moreover, Lu *et al.* [19] integrated multiple machine learning algorithms, including logistic regression (LR), DT, SVM, multilayer perceptron (MLP), and k-nearest neighbors (KNN), and used a voting strategy to predict the risk of cervical cancer.

Although some studies have attempted to use machine learning for cervical cancer prognosis, these applications have been mostly used on public databases. Satisfactory results on public databases often do not mean that the model is competent. In other words, effective methods are still a long way from their application. On the other hand, real-world data sets from a particular region are more complex and specific, and attempts for accurate prognoses on the basis of such data sets can better reflect the application potential of the method. To determine the accuracy of machine learning in cervical cancer prognosis in terms of real-world observations, this study collected the clinical data of patients with cervical cancer from the Fifth People's Hospital of Chengdu and analyzed it using some common machine learning methods including KNN, DT, LR, SVM, RF, XGBoost and LightGBM to identify the prognostic value of simple clinical indicators for this cancer. By analyzing the correlation of these clinical indicators with patient overall survival (OS) and progression-free survival (PFS), it can help doctors make more accurate decisions.

## 2. Materials and Methods

### 2.1 Study population

In this study, we selected the clinicopathological data of 216 patients who were diagnosed with cervical cancer by cervical biopsy or surgical pathology from 2005 to 2014 at the Fifth People's Hospital of Chengdu. All the patients with international federation of gynecology and obstetrics (FIGO) stage I-III cervical cancer were managed by radical hysterectomy and pelvic lymphadenectomy. Para-aortic lymph node dissection was performed in patients with suspicious para-aortic lymph node metastasis. Postoperative radiotherapy with or without concurrent platinum-based chemotherapy was treated depending on the postoperative pathology. The ethics committee of the Fifth People's Hospital of Chengdu approved the study protocol. Before the start of the study, all patients had signed an informed consent form. The features of the patients include age, stage, pathological type, vascular tumor thrombus, lymph node metastasis, interstitial invasion, tumor size, recurrence, recurrence time, death and survival time. The recurrence time, also called PFS, was calculated from the date of operation until the date of the first tumor recurrence. The OS was defined as the time from operation until death or the last follow-up.

### 2.2 Data preprocessing

In the original data set, there are a total of 216 patients' pathology data, including 85 records with missing data. In addition, the features of name, medical record number, and pathology number in the feature set could improve the purity of the training model, but the final trained model did not have a certain generalization ability and could not effectively predict when it faces new samples, and hence they were deleted. Particularly, it is easier to add and subtract discrete features than continuous features during algorithm training, which will speed up the convergence of the model while reducing the risk of model overfitting. Therefore, this study discretized some features based on other related works [20, 21].

### 2.3 Missing data processing

To allow each algorithm to compete fairly on the same data set, the missing data were classified into the following 3 categories:

**TABLE 1. The data set used for training and testing the models.**

| Feature name | Description | Value | Mean/Distribution |
|---|---|---|---|
| Age | Age at diagnosis | | Mean: 44.6 (29–66) |
| Stage | FIGO staging system | | |
| | | Stage I | 97 |
| | | Stage II | 68 |
| | | Stage III | 2 |
| Pathological type | Histological grading | | |
| | | High | 33 |
| | | Middle | 107 |
| | | Low | 20 |
| | | Other | 7 |
| Vascular tumor thrombus | Vascular tumor thrombus or not | | |
| | | Yes | 13 |
| | | No | 154 |
| Lymph node metastasis | Lymph node metastasis or not | | |
| | | Yes | 27 |
| | | No | 140 |
| Interstitial invasion | Degree of tumor invasion | | |
| | | $\leq$1/2 | 59 |
| | | $>$1/2 | 108 |
| Tumor size | Tumor diameter | | |
| | | $\leq$2 cm | 63 |
| | | (2, 4) cm | 58 |
| | | $>$4 cm | 46 |
| Recurrence | Recurrence or not | | |
| | | Yes | 60 |
| | | No | 107 |
| Recurrence time | Time of recurrence | | Mean: 46.1 (mon) |
| Survival time | Overall survival | | Mean: 53.1 (mon) |
| | | 5 yr or less | 109 |
| | | More than 5 yr | 58 |

*FIGO: international federation of gynecology and obstetrics.*

missing completely at random (MCAR), missing at random (MAR), and missing not at random (MNAR) [22, 23]. For continuous MCAR or MAR missing data, the mean was filled, and for discrete ones, the mode (the value which appears most in a set of values) was filled. MNAR data such as the tumor size were processed for deletion because it was uncertain whether they were missing or 0. In addition, the records with missing data in the target column were also eliminated. The preprocessed data set used for training and testing the models is presented in Table 1.

## 2.4 Machine learning methods and baselines

In this study, we adopted several common and widely applied machine learning methods, including KNN [24], DT [25, 26], LR [27], SVM [25, 28], RF [29, 30], XGBoost [31, 32] and LightGBM [33, 34], to predict the OS of patients, with the aim of exploring the value of machine learning methods in the prognosis of cervical cancer, so as to facilitate the diagnosis and improve the accuracy and efficiency of prediction.

Different machine learning models have different hyperparameters. In order to compare the generalizability of different models on the real cervical cancer data set, all hyperparameters were set using default parameters or based on experience, as deemed necessary, to ensure that they compete fairly on the data. In this study, the patient's continuous feature OS was first selected as the label to implement the regression prediction. Second, OS was discretized into two categories, *i.e.*, 5 years or less and more than 5 years, to perform the binary classification task regarding predicting patients' 5-year survival and evaluate the importance of features associated with OS. Finally, these

models were again trained to explore the significant predictors associated with the PFS. For the regression task, Lasso and Ridge regression were selected as the baseline tests.

All experiments were implemented through the Python programming language (Python Software Foundation, https://www.python.org/, version: 3.8.13), package LightGBM (https://lightgbm.readthedocs.io/, version: 3.3.2), package XGBoost (https://xgboost.readthedocs.io/, version: 1.6.1), and package scikit-learn (https://scikit-learn.org/stable/, version: 1.0.2).

## 2.5 Training methods

In order to mine as much useful information as possible from the data and simultaneously to better evaluate the generalization ability of the model, this study adopted the leave-one-out method. Indeed, the cost of the leave-one-out method is a greater computational burden. Although the computational burden may be heavy, the leave-one-out method is a reliable approach [35]. Considering the small sample size of the data set in this study, the increased computational burden can be tolerated; therefore, we utilized the leave-one-out method in this study.

## 2.6 Statistical analysis

This study utilized the Cox proportional hazards model to conduct a univariate survival analysis of each clinical indicator that may affect the patient's OS. The Cox proportional hazards model is essentially a regression model that is commonly used for statistical analysis of medical research for investigating the association between the OS of patients and one or more predictor variables [36–38]. SPSS (Version 26.0, IBM Corp., Armonk, NY, USA) statistical software was applied for data analysis. All $p$ values were bilaterally distributed, and $p < 0.05$ was considered to indicate a statistically significant difference with a 95% confidence interval.

# 3. Results

## 3.1 Performance of different models

Based on the data set, the regression analysis was performed using RF, XGBoost and LightGBM, and Lasso and Ridge regression analyses were performed as baseline tests to predict the OS of patients, considering that the OS of patients is a continuous variable. The final results obtained are represented as root mean square error (RMSE) and are presented in Table 2.

**TABLE 2. Regression performance results of different models.**

| Methods | RMSE |
|---|---|
| Lasso | 0.76 |
| Ridge | 0.73 |
| RF | 0.90 |
| XGBoost | 1.02 |
| LightGBM | 0.78 |

*RF: random forest; XGBoost: extreme gradient boosting; LightGBM: light gradient boosting machine; RMSE: root mean square error.*

For the binary classification problem, patients were divided into two categories on the basis of their survival time (Table 1). Seven machine learning models, which are KNN, DT, LR, SVM, RF, XGBoost and LightGBM, were trained to diagnose the 5-year survival of these patients. The confusion matrix was used to calculate evaluation metrics such as precision, recall, specificity, F1-score and accuracy. Table 3 summarizes the performance evaluation results of the seven models. Each row of the table represents the performance of each model on different evaluation metrics. In addition, Table 4 gives the confusion matrix obtained for each model using which the receiver operating characteristic (ROC) curve of the models was determined (Fig. 1).

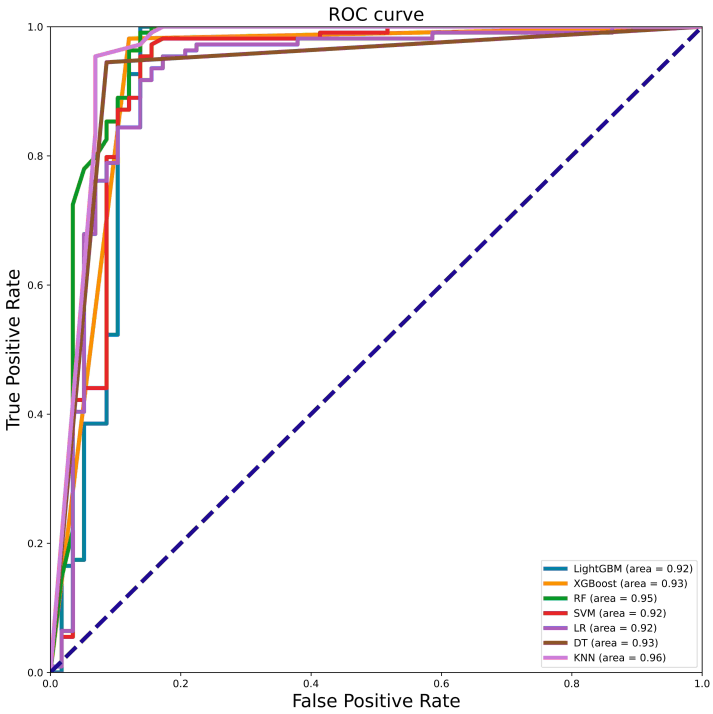


**FIGURE 1. The ROC curve of the models.** ROC: receiver operating characteristic; LightGBM: light gradient boosting machine; XGBoost: extreme gradient boosting; RF: random forest; SVM: support vector machine; LR: logistic regression; DT: decision tree; KNN: k-nearest neighbors.

Table 3 shows the classification performance of the seven models. In general, all the seven machine learning models showed good performance, with LightGBM scoring the highest (KNN: Accuracy = 0.93, DT: Accuracy = 0.92, LR: Accuracy = 0.90, SVM: Accuracy = 0.92, RF: Accuracy = 0.95, XGBoost: Accuracy = 0.95, LightGBM: Accuracy = 0.95). As summarized in Table 4, KNN, RF, XGBoost and LightGBM were highly accurate in predicting positive samples, whereas DT, LR, and SVM were relatively weak. On the other hand, when the samples were negative, all models had similar correct and error rates.

## 3.2 Feature importance

XGBoost and LightGBM were not only able to prognosticate the OS of patients with cervical cancer, but their power also

**TABLE 3. Performance evaluation results of different models.**

| Methods | Precision | Recall | Specificity | F1-score | Accuracy |
|---|---|---|---|---|---|
| KNN | 0.97 | 0.93 | 0.94 | 0.95 | 0.93 |
| DT | 0.93 | 0.95 | 0.87 | 0.94 | 0.92 |
| LR | 0.94 | 0.91 | 0.87 | 0.92 | 0.90 |
| SVM | 0.95 | 0.93 | 0.91 | 0.94 | 0.92 |
| RF | 1.00 | 0.92 | 1.00 | 0.96 | 0.95 |
| XGBoost | 0.98 | 0.94 | 0.96 | 0.96 | 0.95 |
| LightGBM | 1.00 | 0.94 | 1.00 | 0.96 | 0.95 |

*KNN: k-nearest neighbors; DT: decision tree; LR: logistic regression; SVM: support vector machine; RF: random forest; XGBoost: extreme gradient boosting; LightGBM: light gradient boosting machine.*

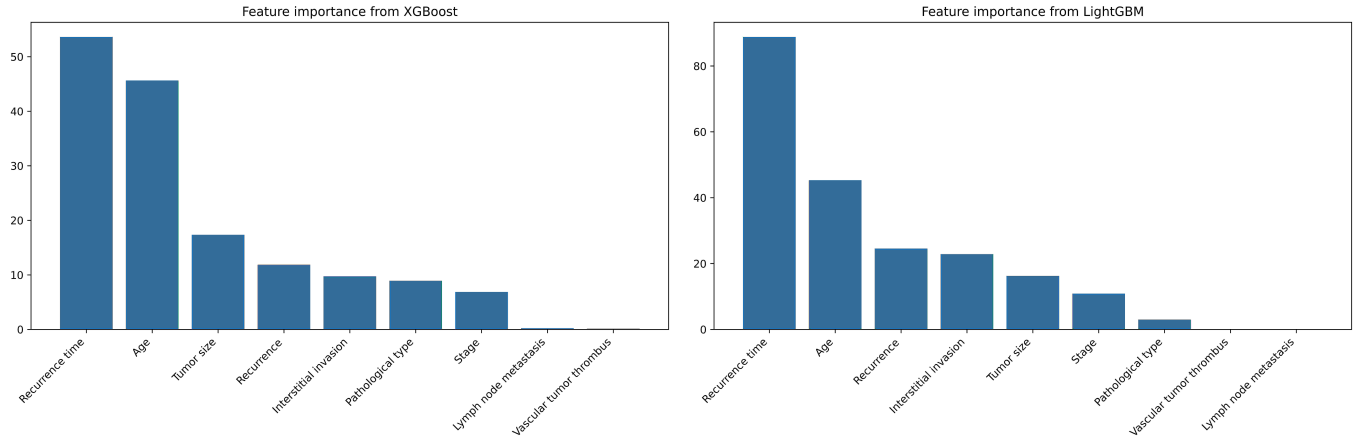


**FIGURE 2. Feature importance with OS.** Left: Feature importance from XGBoost. Right: Feature importance from LightGBM. XGBoost: extreme gradient boosting; LightGBM: light gradient boosting machine.

**TABLE 4. Confusion matrix of different models.**

| Methods | TP | FN | FP | TN |
|---|---|---|---|---|
| KNN | 106 | 8 | 3 | 50 |
| DT | 101 | 5 | 8 | 53 |
| LR | 102 | 10 | 7 | 48 |
| SVM | 104 | 8 | 5 | 50 |
| RF | 109 | 9 | 0 | 49 |
| XGBoost | 107 | 7 | 2 | 51 |
| LightGBM | 109 | 8 | 0 | 50 |

*KNN: k-nearest neighbors; DT: decision tree; LR: logistic regression; SVM: support vector machine; RF: random forest; XGBoost: extreme gradient boosting; LightGBM: light gradient boosting machine; TP: true positive; FN: false negative; FP: false positive; TN: true negative.*

lies in the mechanism by which they can count the correlation of each feature with the label. To determine the clinical indicators with a greater effect on the OS and PFS of these patients, feature importance was calculated using XGBoost and LightGBM, and the ranking of feature importance is given in Fig. 2 and Fig. 3.

## 3.3 Survival analysis

In order to verify the correctness of the machine learning outcomes, we conducted the Cox model of OS and PFS so as to validate the results from a statistical perspective. The survival curves are shown in Fig. 4 and Fig. 5, and the detailed results are listed in Table 5 and Table 6. In Fig. 4, the survival curve suggests that the OS of patients with recurrence (Fig. 4a), interstitial infiltration (Fig. 4b), lymph node metastasis (Fig. 4c), and age (>50, Fig. 4d) is significantly shorter than that of negative patients. With the later stage (Fig. 4e) and the larger the tumor size (Fig. 4f), the patient's OS can gradually decrease. Although vascular tumor thrombus (Fig. 4g) and pathological type (Fig. 4h) also affect the OS of patients, they have little effect and are not statistically significant ($p > 0.05$). Univariate survival analysis of OS implies that recurrence (Hazard ratios (HR) 276.429, 95% confidence interval (CI) 20.692–3692.888, $p < 0.0001$), interstitial invasion (HR 5.675, 95% CI 2.378–13.545, $p < 0.0001$), stage (HR 3.674, 95% CI 2.789–4.484, $p < 0.0001$), lymph node metastasis (HR 2.592, 95% CI 1.297–5.178, $p < 0.01$), tumor size (HR 3.245, 95% CI 1.595–4.315, $p < 0.0001$), and age (HR 2.623, 95% CI 1.595–4.315, $p < 0.001$) were significantly associated with OS. For PFS in Table 6, interstitial invasion (HR 4.742, 95% CI 2.394–9.394, $p < 0.0001$), stage (HR 2.874, 95% CI 2.172–3.803, $p < 0.0001$), tumor size (HR 2.343, 95% CI 1.599–3.433, $p < 0.0001$) and, age (HR 2.236, 95% CI 1.435–3.484, $p < 0.001$)

**TABLE 5. The results of Cox model of OS.**

| Feature name | b | SE | Wald $\chi$ | $p$ | HR | HR (95% CI) |
|---|---|---|---|---|---|---|
| Recurrence | 5.66 | 1.32 | 18.07 | *** | 276.43 | 20.69–3692.89 |
| Interstitial invasion | 1.74 | 0.44 | 15.30 | *** | 5.68 | 2.38–13.55 |
| Stage | 1.30 | 0.14 | 85.63 | *** | 3.67 | 2.79–4.48 |
| Lymph node metastasis | 0.95 | 0.35 | 7.28 | * | 2.59 | 1.30–5.18 |
| Tumor size | 1.18 | 0.25 | 22.15 | *** | 3.25 | 1.99–5.30 |
| Age | 0.96 | 0.25 | 14.43 | ** | 2.62 | 1.60–4.32 |
| Vascular tumor thrombus | −0.72 | 0.73 | 0.01 | 0.92 | 0.93 | 0.23–3.86 |
| Pathological type | 0.30 | 0.21 | 1.97 | 0.16 | 1.35 | 0.89–2.04 |

*b: Regression coefficient of the model; SE: Standard error; Wald $\chi$: Wald statistics; p: Level of significance: *p < 0.01. **p < 0.001. ***p < 0.0001. The value of p < 0.05 is considered statistically significant; HR: Hazard ratios; HR (95% CI): 95% confidence interval (CI) for HR.*

**TABLE 6. The results of Cox model of PFS.**

| Feature name | b | SE | Wald $\chi$ | $p$ | HR | HR (95% CI) |
|---|---|---|---|---|---|---|
| Interstitial invasion | 1.56 | 0.35 | 19.91 | *** | 4.74 | 2.39–9.39 |
| Stage | 1.06 | 0.14 | 54.53 | *** | 2.87 | 2.17–3.80 |
| Lymph node metastasis | 0.60 | 0.31 | 3.85 | 0.05 | 1.83 | 1.00–3.33 |
| Tumor size | 0.85 | 0.20 | 19.10 | *** | 2.34 | 1.60–3.43 |
| Age | 0.81 | 0.23 | 12.64 | ** | 2.24 | 1.44–3.48 |
| Vascular tumor thrombus | 0.05 | 0.52 | 0.01 | 0.92 | 1.05 | 0.38–2.91 |
| Pathological type | 0.29 | 0.19 | 2.22 | 0.14 | 1.33 | 0.91–1.95 |

*b: Regression coefficient of the model; SE: Standard error; Wald $\chi$: Wald statistics; p: Level of significance: *p < 0.01. **p < 0.001. ***p < 0.0001. The value of p < 0.05 is considered statistically significant; HR: Hazard ratios; HR (95% CI): 95% confidence interval (CI) for HR.*

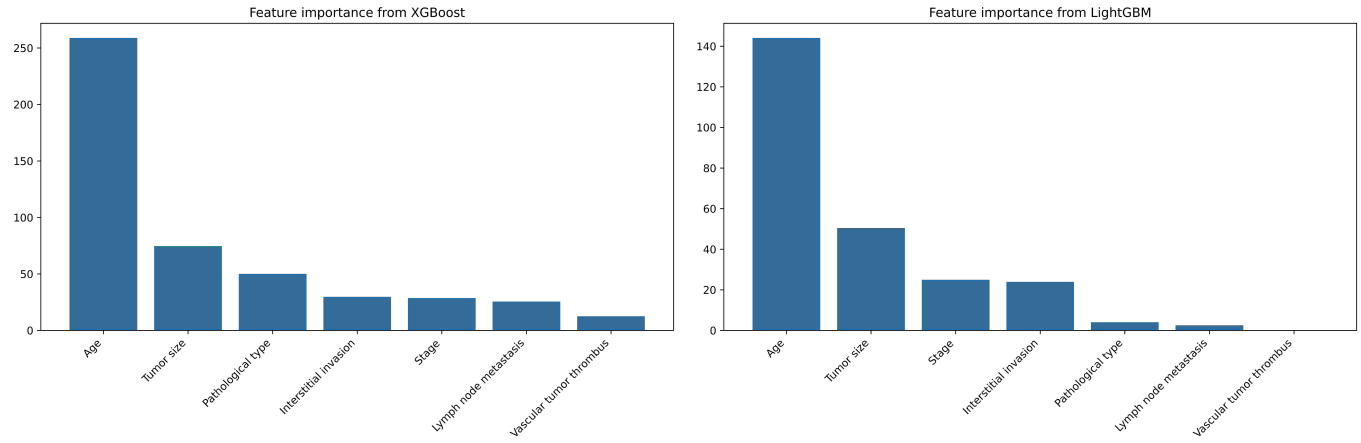


**FIGURE 3. Feature importance with PFS.** Left: Feature importance from XGBoost. Right: Feature importance from LightGBM. XGBoost: extreme gradient boosting; LightGBM: light gradient boosting machine.

were significantly associated with it.

## 4. Discussion

Cervical cancer is the fourth most common malignancy worldwide and the second leading cause of cancer death in women aged 20 to 39 years [1]. For detecting of cervical cancer lesions, screening strategies consist of cytological methods, colposcopy, and many other methods. However, most of these screening methods are highly dependent on the physician's expertise and subjective experience is involved in the decision-making process. In developing countries, healthcare resources are extremely limited and patients often have difficulty adhering to routine screening because of the low level of awareness of the problem. Therefore, predicting the risk of patients with cervical cancer is important. Compared with clinical methods, machine learning methods can classify these patients more easily and accurately. When a set of risk factors for cervical cancer is known, such as interstitial invasion and recurrence time, machine learning methods can train classifiers on the basis of these risk factors and predict outcomes when new cases with these risk factors are identified.

In this study, based on the real-world clinical data of patients with cervical cancer, we applied several machine learning methods to predict the OS of these patients. First, we performed a regression task by using RF, XGBoost and LightGBM to determined the OS and compared it that determined using classical statistical methods. As summarized in Table 2, three machine learning models performed close to Lasso and Ridge regression, whose average RMSE is about 0.9 (months).

However, in reality, patients and doctors are often not sensitive to survival times that are precise to months or days, regardless of whether the RMSE is 0.7 or 0.9 months. In addition, various underlying factors affect patients, and the unavoidable errors in model predictions may not be what patients and doctors need. Therefore, we discretized the OS of patients and transformed it into a binary classification problem to obtain results that are more acceptable to patients and physicians with high accuracy.

Fig. 1 shows the ROC curve indicating the satisfying performance of the machine learning methods for predicting the OS in patients with cervical cancer. Thereafter, using XGBoost and LightGBM, the correlation between each clinical indicator and the survival of these patients was calculated. As shown in Fig. 2, both models concluded that recurrence time, age, interstitial invasion and tumor size significantly affect the OS of patients with cervical cancer, which is consistent with the clinical experience of the doctors.

The statistical analysis results using the Cox model successfully confirmed the effectiveness and correctness of the machine learning methods. Both XGBoost and LightGBM identified that recurrence, age, interstitial invasion and tumor size were critical for the OS of patients and that these clinical indicators were statistically significant. Similarly, recurrence, age, interstitial invasion and tumor size were also statistically significant in the Cox model, which also reflected a strong correlation with the machine learning models. On the other hand, together with the results obtained from the machine learning, the univariate survival analysis suggested that clinical indicators of the patients including recurrence, age, interstitial invasion, and tumor size are important prognostic factors for determing OS.

Among these important clinical OS prognostic indicators, recurrence had a direct relationship with OS. However, in the clinical setting, this clinical indicator is often difficult to measure and predict. If the predictors of recurrence can be found, it will help clinicians to make accurate decisions and then administer targeted treatment. As shown in Fig. 5 and summarized in Table 6, the univariate survival analysis of PFS suggested that, interstitial invasion, stage, tumor size, and age were significantly associated with patient recurrence. Moreover, XGBoost and LightGBM highlighted the importance and association of these clinical indicators with recurrence. As shown in Fig. 3, patient age and tumor size were important factors affecting recurrence, whereas other clinical indicators showed some degree of correlation. This observation is consistent with the results obtained using the Cox model, implying that age and tumor size can be considered a reliable basis for the decision-making process by clinicians.

In summary, the obtained results and validation indicate the value of machine learning methods in the prognosis of cervical cancer. On one hand, the accuracy of machine learning methods is close to that of classical statistical methods. On the other hand, they are highly accurate and can identify prognostic factors significantly associated with patient OS and PFS, which can assist doctors in making accurate decisions. Moreover, compared to other related studies based on public datasets [15–18], in the present study, we attempted to apply machine learning methods to the real-world clinical data of patients with cervical cancer. We believe that machine learning methods could be more widely and effectively applied to the prognosis of cervical cancer in the future.

Because this was a retrospective study, some limitations and shortcomings are inevitable. The size of the data set and the number of missing values often affect the performance of the machine learning models. A larger training set with fewer missing values can help in developing more accurate and generalized models. Hence, including patient clinical data from different hospitals and regions in the subsequent study is essential. Importantly, machine learning models are somewhat overwhelmed when faced with new extreme cases and may give wrong predictions. Hence, regardless of a model's accuracy, its results cannot be a decisive factor in a clinician's decision. In addition, although there have been initial attempts of applying machine learning in medicine, including but not limited to medical decision making, medical imaging, and medical information, the extension of the results derived from this study to other practical clinical applications is possible but still requires further rigorous exploration and validation.

To summarize, we used several machine learning methods to identify the prognostic factors associated with the OS and PFS of patients with cervical cancer and verified the conclusions using the Cox statistical analysis model. We found that recurrence time, age, interstitial invasion, and tumor size are crucial prognostic factors for cervical cancer, whereas age and tumor size could be predictors of recurrence.

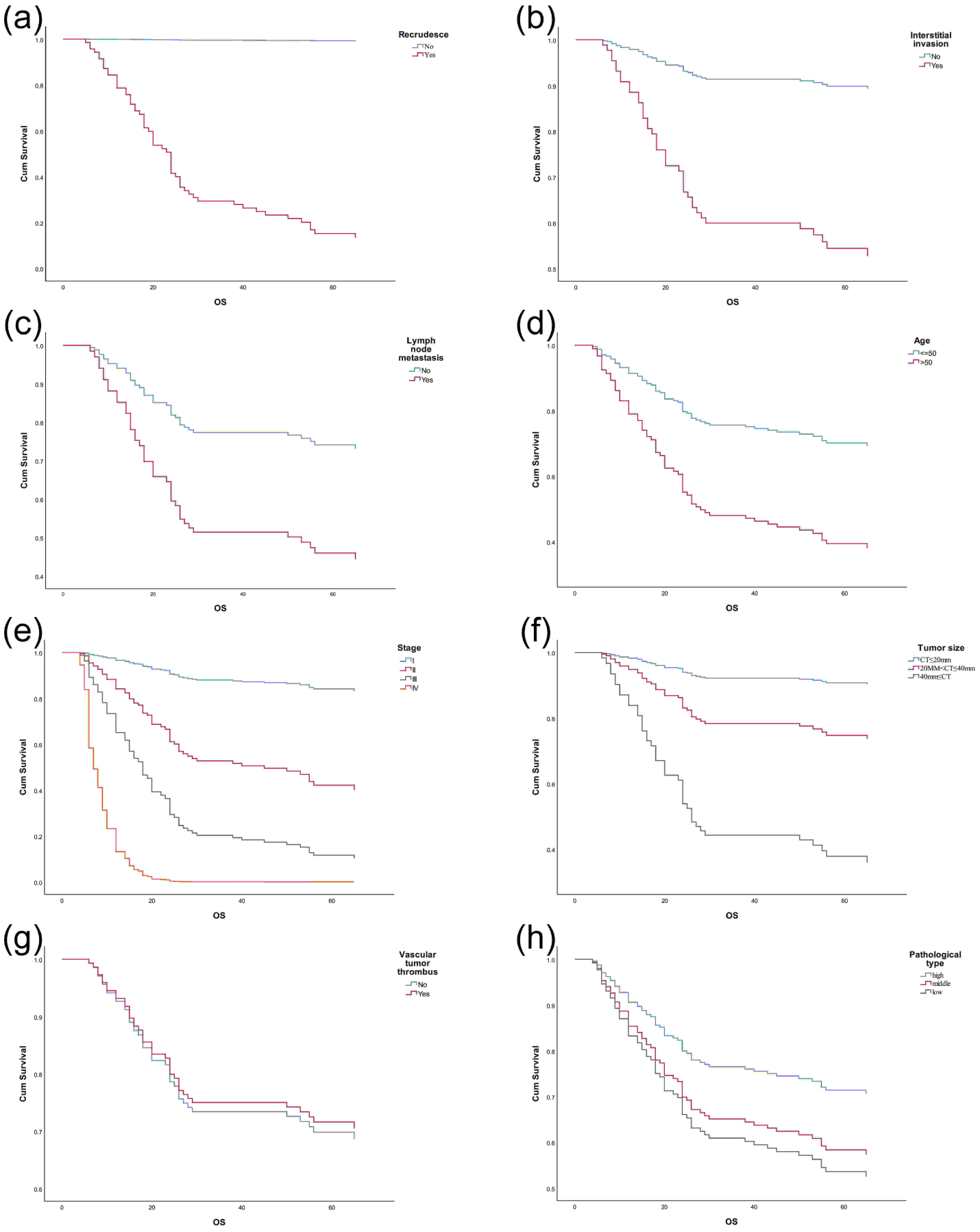


**FIGURE 4. Cox model of OS.** (a) Recurrence. (b) Interstitial invasion. (c) Lymph node metastasis. (d) Age. (e) Stage. (f) Tumor size. (g) Vascular tumor thrombus. (h) Pathological type. The prognostic value of the features is visually examined through the model. OS: overall survival; CT: computed tomography.

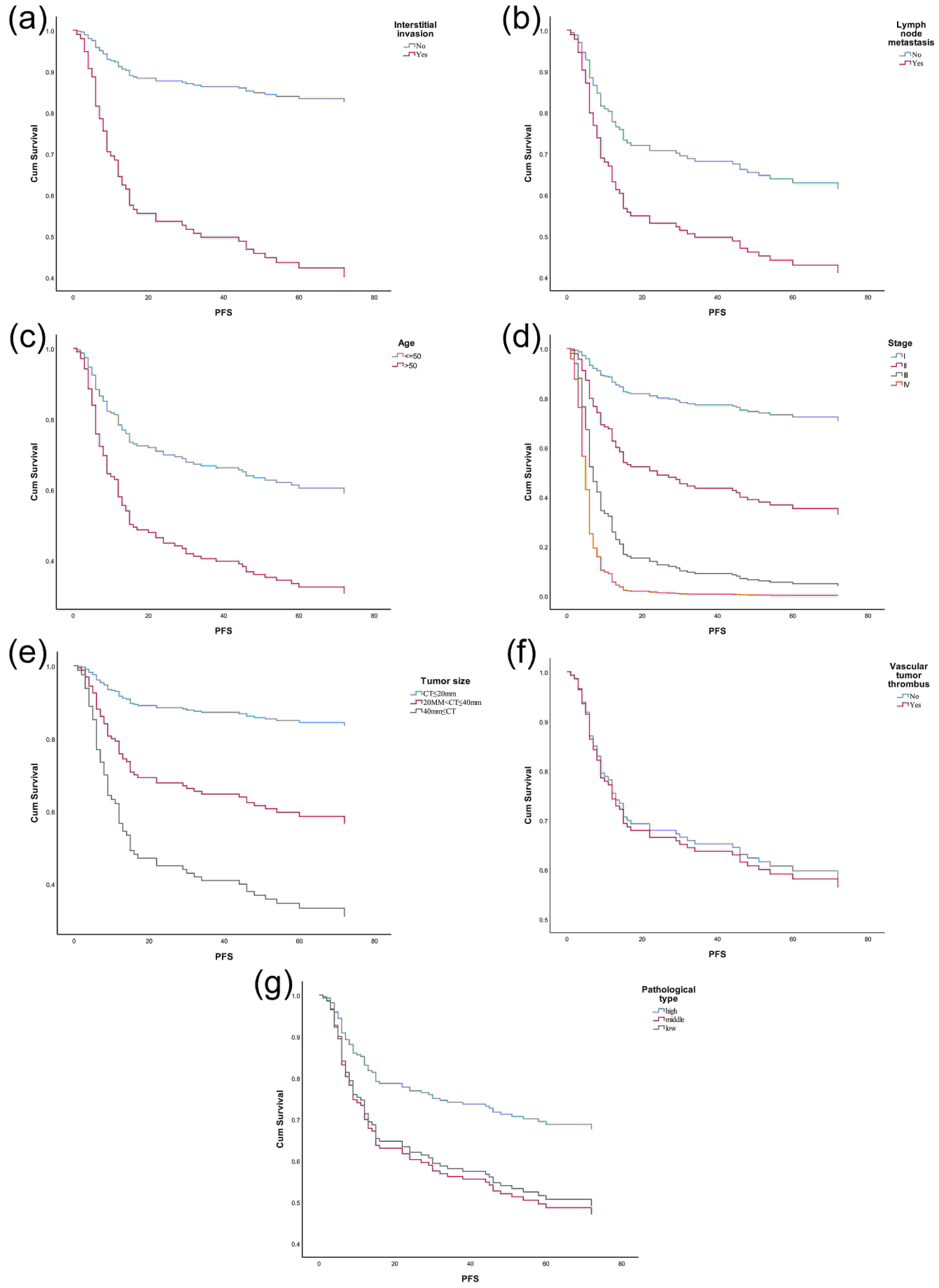


**FIGURE 5. Cox model of PFS.** (a) Interstitial invasion. (b) Lymph node metastasis. (c) Age. (d) Stage. (e) Tumor size. (f) Vascular tumor thrombus. (g) Pathological type. PFS: progression-free survival; CT: computed tomography.

## 5. Conclusion

In this study, several machine learning methods were successfully applied to identify the prognostic factors associated with the OS and PFS of patients with cervical cancer. Compared to other related studies that used public databases, this study collected more complex and real-world pathological data from the Fifth People's Hospital in Chengdu. Experiments using real-world pathological data demonstrated the immense potential of machine learning methods for cervical cancer prognosis (KNN: Accuracy = 0.93, DT: Accuracy = 0.92, LR: Accuracy = 0.90, SVM: Accuracy = 0.92, RF: Accuracy = 0.95, XGBoost: Accuracy = 0.95, LightGBM: Accuracy = 0.95), indicating that these methods have an immense application value in this field and could help doctors make more accurate and faster decisions. In the future, performance optimization of machine learning prognostic models and their generalizability are of interest to our research.

### AUTHOR CONTRIBUTIONS

YZR, WWZ and XRP—designed the study; WWZ—collected the data; YWL and ZDL—conducted the experiments; All authors analysed the results and reviewed the manuscript.

### ETHICS APPROVAL AND CONSENT TO PARTICIPATE

The present study was approved by the Ethics Committee of the Fifth People's Hospital (Chengdu, China, approval number: 2016-002-01) and was conducted according to the Declaration of Helsinki. Patients provided informed written consent at the time of data collection.

### ACKNOWLEDGMENT

We acknowledge all the clinicians, nurses, lab technicians, and interviewees who agreed to participate and give their opinions in this study.

### FUNDING

This work was supported in part by National Natural Science Foundation of China (No. 61806043), Sichuan Science and Technology Program (Nos. 2021YFS0172, 2022YFS0047, and 2022YFS0055), Guangdong Basic and Applied Basic Research Foundation (No. 2020A1515011002), Guangzhou Science and Technology Program (No. 202002030266), Medico-Engineering Cooperation Funds from University of Electronic Science and Technology of China (No. ZYGX2021YGLH022), Opening Funds from Radiation Oncology Key Laboratory of Sichuan Province (No. 2021ROKF02), and Chengdu Medical Research Project (Project No. 2020035). The funders had no role in design, data collection, analysis or publication related decisions of the study.

### CONFLICT OF INTEREST

The authors declare no conflict of interest.

### REFERENCES

[1] Siegel RL, Miller KD, Jemal A. Cancer statistics, 2020. CA: A Cancer Journal for Clinicians. 2020; 70: 7–30.

[2] Prandi GA, Cocchio S, Fonzo M, Furlan P, Nicoletti M, Baldo V. Towards the elimination of cervical cancer: HPV epidemiology, real-world experiences and the potential impact of the 9-valent HPV vaccine. European Journal of Gynaecological Oncology. 2021; 42: 1068–1078.

[3] Passos CM, Sales JB, Maia EG, Caldeira TCM, Rodrigues RD, Figueiredo N, *et al.* Trends in access to female cancer screening in Brazil, 2007–16. Journal of Public Health. 2021; 43: 632–638.

[4] Sahasrabuddhe VV, Parham GP, Mwanahamuntu MH, Vermund SH. Cervical cancer prevention in low- and middle-income countries: feasible, affordable, essential. Cancer Prevention Research. 2012; 5: 11–17.

[5] He S, Liao B, Deng Y, Su C, Tuo J, Liu J, *et al.* MiR-216b inhibits cell proliferation by targeting FOXM1 in cervical cancer cells and is associated with better prognosis. BMC Cancer. 2017; 17: 673.

[6] Park S, Kim J, Eom K, Oh S, Kim S, Kim G, *et al*. MicroRNA-944 overexpression is a biomarker for poor prognosis of advanced cervical cancer. BMC Cancer. 2019; 19: 419.

[7] Dunkler D, Ploner M, Schemper M, Heinze G. Weighted cox regression using the R package coxphw. Journal of Statistical Software. 2018; 84: 1–26.

[8] She Y, Jin Z, Wu J, Deng J, Zhang L, Su H, *et al*. Development and validation of a deep learning model for non-small cell lung cancer survival. JAMA Network Open. 2020; 3: e205842.

[9] Wang X, Chen H, Gan C, Lin H, Dou Q, Tsougenis E, *et al*. Weakly supervised deep learning for whole slide lung cancer image analysis. IEEE Transactions on Cybernetics. 2020; 50: 3950–3962.

[10] Liu P, Fu B, Yang SX, Deng L, Zhong X, Zheng H. Optimizing survival analysis of xgboost for ties to predict disease progression of breast cancer. IEEE Transactions on Biomedical Engineering. 2021; 68: 148–160.

[11] Zhang Y, Li Q, Xin Y, Lv W. Differentiating prostate cancer from benign prostatic hyperplasia using PSAD based on machine learning: single-center retrospective study in China. IEEE/ACM Transactions on Computational Biology and Bioinformatics. 2019; 16: 936–941.

[12] Mansour RF. A robust deep neural network based breast cancer detection and classification. International Journal of Computational Intelligence and Applications. 2020; 19: 2050007.

[13] Ragab M, Albukhari A, Alyami J, Mansour RF. Ensemble deep-learning-enabled clinical decision support system for breast cancer diagnosis and classification on ultrasound images. Biology. 2022; 11: 439.

[14] Althobaiti MM, Ashour AA, Alhindi NA, Althobaiti A, Mansour RF, Gupta D, *et al*. Deep transfer learning-based breast cancer detection and classification model using photoacoustic multimodal images. BioMed Research International. 2022; 2022: 1–13.

[15] Alam TM, Khan MMA, Iqbal MA, Abdul W, Mushtaq M. Cervical cancer prediction through different screening methods using data mining. International Journal of Advanced Computer Science and Applications. 2019; 10: 388–396.

[16] Wu W, Zhou H. Data-driven diagnosis of cervical cancer with support vector machine-based approaches. IEEE Access. 2017; 5: 25189–25195.

[17] Ijaz MF, Attique M, Son Y. Data-driven cervical cancer prediction model with outlier detection and over-sampling methods. Sensors. 2020; 20: 2809.

[18] Deng X, Luo Y, Wang C. 'Analysis of risk factors for cervical cancer based on machine learning methods'. 2018 5th IEEE International Conference on Cloud Computing and Intelligence Systems (CCIS). 23–25 November 2018. IEEE: Nanjing, China. 2018.

[19] Lu J, Song E, Ghoneim A, Alrashoud M. Machine learning for assisting cervical cancer diagnosis: an ensemble approach. Future Generation Computer Systems. 2020; 106: 199–205.

[20] Turan T, Yildirim BA, Tulunay G, Boran N, Kose MF. Prognostic effect of different cut-off values (20 mm, 30 mm and 40 mm) for clinical tumor

size in FIGO stage IB cervical cancer. Surgical Oncology. 2010; 19: 106–113.

[21] Kato T, Takashima A, Kasamatsu T, Nakamura K, Mizusawa J, Nakanishi T, *et al*. Clinical tumor diameter and prognosis of patients with FIGO stage IB1 cervical cancer (JCOG0806-A). Gynecologic Oncology. 2015; 137: 34–39.

[22] Graham JW. Missing data analysis: making it work in the real world. Annual Review of Psychology. 2009; 60: 549–576.

[23] Karadaghy OA, Shew M, New J, Bur AM. Development and assessment of a machine learning model to help predict survival among patients with oral squamous cell carcinoma. JAMA Otolaryngology-Head & Neck Surgery. 2019; 145: 1115–1120.

[24] Song C, Li X. Cost-sensitive KNN algorithm for cancer prediction based on entropy analysis. Entropy. 2022; 24: 253.

[25] Afolayan JO, Adebiyi MO, Arowolo MO, Chakraborty C, Adebiyi AA. Breast cancer detection using particle swarm optimization and decision tree machine learning technique. Intelligent Healthcare. 2022; 286: 61–83.

[26] Idris NF, Ismail MA. Breast cancer disease classification using fuzzy-ID3 algorithm with FUZZYDBD method: automatic fuzzy database definition. PeerJ Computer Science. 2021; 7: e427.

[27] Thabtah F, Abdelhamid N, Peebles D. A machine learning autism classification based on logistic regression analysis. Health Information Science and Systems. 2019; 7: 12.

[28] Lilhore UK, Simaiya S, Pandey H, Gautam V, Garg A, Ghosh P. Breast cancer detection in the iot cloud-based healthcare environment using fuzzy cluster segmentation and svm classifier. Ambient Communications and Computer Systems. 2022; 356: 165–179.

[29] Wang X, Zhai M, Ren Z, Ren H, Li M, Quan D, *et al*. Exploratory study on classification of diabetes mellitus through a combined random forest classifier. BMC Medical Informatics and Decision Making. 2021; 21: 105.

[30] Liu YH, Jin J, Liu YJ. Machine learning-based random forest for predicting decreased quality of life in thyroid cancer patients after thyroidectomy. Supportive Care in Cancer. 2022; 30: 2507–2513.

[31] Chen T, Guestrin C. XGBoost: a scalable tree boosting system. Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. 13–17 August 2016. Association for Computing Machinery: San Francisco, CA. 2016.

[32] Li Y, Zou Z, Gao Z, Wang Y, Xiao M, Xu C, *et al.* Prediction of lung cancer risk in Chinese population with genetic-environment factor using extreme gradient boosting. Cancer Medicine. 2022; 00: 1–10.

[33] Ke G, Meng Q, Finley T, Wang T, Chen W, Ma W, *et al.* Lightgbm: a highly efficient gradient boosting decision tree. Advances in Neural Information Processing Systems. 2017; 30: 3146–3154.

[34] Huang K, Zhang J, Yu Y, Lin Y, Song C. The impact of chemotherapy and survival prediction by machine learning in early elderly triple negative breast cancer (eTNBC): a population based study from the SEER database. BMC Geriatrics. 2022; 22: 268.

[35] Yang C, Zhu X, Ahmad Z, Wang L, Qiao J. Design of incremental echo state network using leave-one-out cross-validation. IEEE Access. 2018; 6: 74874–74884.

[36] Mittal S, Madigan D, Burd RS, Suchard MA. High-dimensional, massive sample-size cox proportional hazards regression for survival analysis. Biostatistics. 2014; 15: 207–221.

[37] Halle MK, Sødal M, Forsse D, Engerud H, Woie K, Lura NG, *et al*. A 10-gene prognostic signature points to LIMCH1 and HLA-DQB1 as important players in aggressive cervical cancer disease. British Journal of Cancer. 2021; 124: 1690–1698.

[38] Sheng W, Bai WP. Identification of hypoxia-related prognostic signature for ovarian cancer based on cox regression model. European Journal of Gynaecological Oncology. 2022; 43: 247–256.

**How to cite this article:** Yawen Ling, Weiwei Zhang, Zhidong Li, Xiaorong Pu, Yazhou Ren. Application and comparison of several machine learning methods in the prognosis of cervical cancer. European Journal of Gynaecological Oncology. 2022; 43(6): 34-44. doi: 10.22514/ejgo.2022.056.