Transactions on Computer Science and Intelligent Systems Research
ISSN: 2960-1800, eISSN: 2960-2238 | Volume 5
AIDML 2024

A Research on Machine Learning for Predicting Survival Time in Cancer Patients

Binghua Xi

College of Information, Mechanical and Electrical Engineering, Shanghai Normal University, Shanghai, China

1000514516@smail.shnu.edu.cn

Abstract. Cancer remains one of the leading causes of mortality worldwide. Predicting the survival of cancer patients holds immense significance in guiding treatment decisions and enhancing patient quality of life. To provide a comprehensive understanding of the application of machine learning methods and to explore future research directions in survival prediction, this paper reviews the current landscape of machine learning algorithms in this domain. The paper discusses the significance and challenges associated with predicting the survival time of cancer patients. It begins with an introduction highlighting the importance of this prediction task in the medical domain. Traditional methods utilized for survival prediction are then explored, shedding light on their strengths and limitations. Following this, the paper delves into the application of traditional machine learning approaches for survival prediction. Subsequently, the focus shifts towards deep learning algorithms and their role in survival prediction. In conclusion, the paper identifies potential avenues for further research in this field, paving the way for continued advancements in cancer survival prediction.

Keywords: Survival prediction; machine learning; predictive Models.

1. Introduction

Medical technology is highly developed in today's society, cancer is still a cause of most death. According to the International Agency for Research on Cancer released by the International Agency for Research on Cancer on January 6, 2022, there were19.29 million new cancer cases worldwide in 2020 and 9.96 million cancer deaths worldwide in 2020.

Timely treatment and appropriate treatment strategies of cancer are very important to improve the survival rate of cancer patients. Although patients received prompt treatment, many of them will still die of a recurrence of cancer. According to Haojie's research, pancreatic cancer has an 80% recurrence rate [1]. In response to this issue, setting up a model to estimate the survival time of cancer patients can help medical workers to adjust their treatment strategies and medical resources allocation to satisfy patients' requirements best.

In the early 20th century, statistical modeling algorithms are developed to learn patterns in data and perform tasks such as classification and regression. With the increase of computing power, machine learning algorithms represented by decision tree are developed to train models to optimize predictive performance, which is focuses on inferring outputs by learning patterns from given input data and optimizing the model's generalization ability. In recent years, a breakthrough of deep learning algorithms has been made because of the appearance of large-scale data set and the application of GPU, which has made important contributions to the fields of natural language processing and image recognition.

In prior studies, predictive modeling algorithms are used widely for prediction of cancer patient's survival time and all of these algorithms have been proved that it has good application value in clinic. Karabacak et al developed machine learning algorithms based on decision trees and random forests to predict short-term postoperative outcomes following spinal tumor resections and indicates that machine learning algorithms can give fair predictions [2]. Haojie's research reveals deep learning algorithms and statistical models also have good accuracy and generalization [1].

This paper will review the current applications of predictive modeling algorithms on predicting survival time of cancer patients in China and foreign countries, analyze and compare the performance of each kind of algorithm in predicting survival time and give some suggestions to future researches. Based on the review, analysis and comparisons, readers can understand the core principle of relevant algorithms while learning the application of them in predicting the survival time of cancer patients.

2. Importance and Challenges of Predicting Cancer Patient Survival time

Survival predictions are a key factor for doctors when making treatment decisions for cancer patients. Doctors can rationally allocate medical resources under the guidance of survival prediction to avoid excessive waste of medical resources. For patients with short survival time, doctors can also take more care of patients' psychology health [3].

Until reliable predictive models emerged, predicting a patient's survival time depended on the judgment of doctors. But the subjective judgment can be influenced by the emotion or experience of the doctor, which can lead to errors.

Furthermore, predicting the survival time of cancer patients is a very comprehensive task. First, The survival time of cancer patients can be affected by many variables. Quannian's research used 43 types of patient's physical conditions as variables to fit the model for prediction [4]. Second, some patients may experience events that other patients haven't experienced during the research, which may make their survival time unknown [5].

In addition to that, clinical data is often missing, heterogeneous, and truncated, which can not be treated effectively by the traditional statistical methods[6]. Therefore, setting up a reliable model for cancer patients' survival analysis is significantly important but challenging.

3. Traditional Methods for Survival Prediction

Traditionally, the Cox Proportional Hazards model (CPH) is used most widely in predicting the survival time of cancer patients[1,4,5]. It is a semi-parametric regression model proposed by British statistician D.R. OX in 1972, which can simultaneously study the relationship between multiple risk factors and the occurrence and occurrence time of event outcomes, thus overcoming the deficiency of single factor limitation in simple survival analysis.

However, the CPH model has some limitations. Firstly, the CPH model assumes that the effect of covariates on risk remains is constant throughout the follow-up period, but this hypothesis is not always valid in actual applications, which would cause mistakes.

Next, the CPH model is not suitable for high-dimensional datasets. When the number of covariates greatly exceeds the sample size, the estimation results of the CPH model may become unstable and prone to overfitting [5].

As a traditional method, the CPH model is usually the first choice in survival time prediction. In papers that use deep learning and machine learning to make predictions, the CPH model is often used as references. Haojie and Quannian's research used different machine learning models and deep learning models to predict patient's survival after pancreatic ductal adenocarcinoma, and compared these results with that given by the CPH model. In these cases, the CPH model provides reference value for evaluating the achievements of machine learning and deep learning [1,4].

Today the CPH model is gradually replaced by machine learning algorithms and deep learning algorithms in survival prediction, because machine learning algorithms and deep learning algorithms can perform better in their respective areas of expertise. But the CPH model is still an important method for survival prediction [1,7].

However, studies about the CPH model are still going on. The CPH model improved with different methods or combined with other models can perform better than the original CPH model and it is widely used in survival prediction. Zhou et al set up a prognostic nomogram based on a CPH model applied with LASSO method to achieve variable selection and dimensionality reduction, which can

improve the model's performance [8]. Xiaoyu's research proposed a new survival analysis algorithm called XGBENC, which combined the CPH model with machine learning methods. This model showed better accuracy and robustness in survival prediction [6]. Wangwang's research combined the CPH model with deep learning algorithms and also has a good performance [9].

4. Traditional Machine Learning Methods for Survival Prediction

Machine learning algorithms are aim to train models with real world data so that it can extract features and patterns from it and solve classification and regression problems. Compared to the CPH model, machine learning models can perform better in processing data with non-linear relationships, which can make the survival prediction more accurate.

Traditional machine learning methods such as logistic regression, support vector machine, decision tree, gradient boosting tree and random forests are simpler in model structures, so that they take less computational resource and they are advantageous in interpretability and iteration capability. Ensemble learning algorithms, with random forest as a representative, demonstrate particular advantages in handling high-dimensional data [4].

Traditional Machine learning algorithms also have broad applications in survival prediction. The application of the mentioned major machine learning algorithms in survival prediction is quite common and models have been adapted and modified in practical applications to suit survival prediction needs. Among machine learning algorithms, the random forest is the most popular algorithm [1,2,3,4,6,10].

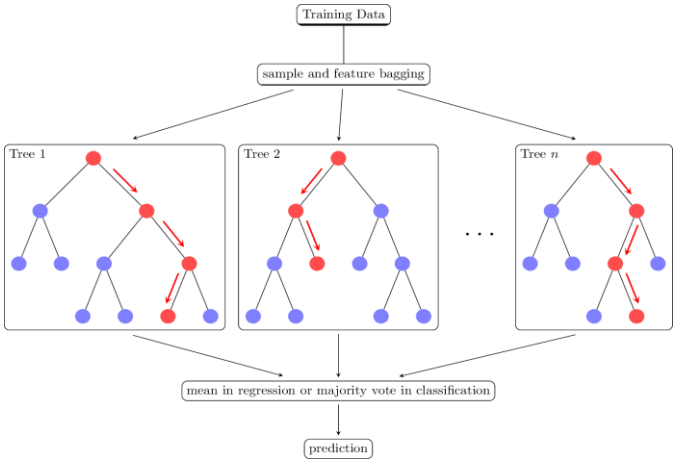

Figure 1. The Architecture of Random Forest (Photo/Picture credit :Original)

Quannian's research applied decision tree, support vector machine, random forest (Figure 1) and logistic regression in a same dataset and demonstrates that the random forest and the logistic regression exhibited well performance [4]. Among them, random forest showed significantly better performance. However, Quannian's research about random forest may exist overfitting, the AUROC of random forest is very close to 1, which is too high. Karabacak et al developed an online survival prediction tool with XGBoost, LightBGM, CatBoost and random forest. These are frameworks for solving supervised learning problems based on gradient boosting tree and decision tree. According to Karabacak's research, the mean AUC of XGBoost, LightBGM, CatBoost and random forest are 0.704, 0.729, 0.726 and 0.743 respectively, which showed the random forest performs better than boosted tree methods. These frameworks are applied in the same dataset, which is come from SEER databas [2], so that the comparison is fair and has value of reference.

5. Deep Learning Algorithms for Survival Prediction

Deep learning algorithms are special kinds of machine learning algorithms. Deep learning aims to build multi-layer neural network models that achieve effective representation and pattern recognition of input data through successive layers of feature extraction and transformation. The complexity of the

model can be extended by increasing the number of layers and the number of neurons to accommodate larger data and more complex tasks.

The deep learning algorithms also have some disadvantages. First, the models given by deep learning are considered as black box models, which are difficult to interpret. Second, deep learning networks usually needs a large amount of annotated data for training to achieve better performance, and that will cost great manpower resources, money and computing resources.

Due to deep learning's excellent feature extraction capabilities, deep learning algorithms have become one of the most popular methods in survival prediction [2,3,11].

The most common use of deep learning for survival prediction is prediction through image recognition. The Convolutional Neural Network (CNN) is mainly applied in image recognition because of its obvious advantages in unstructured data.

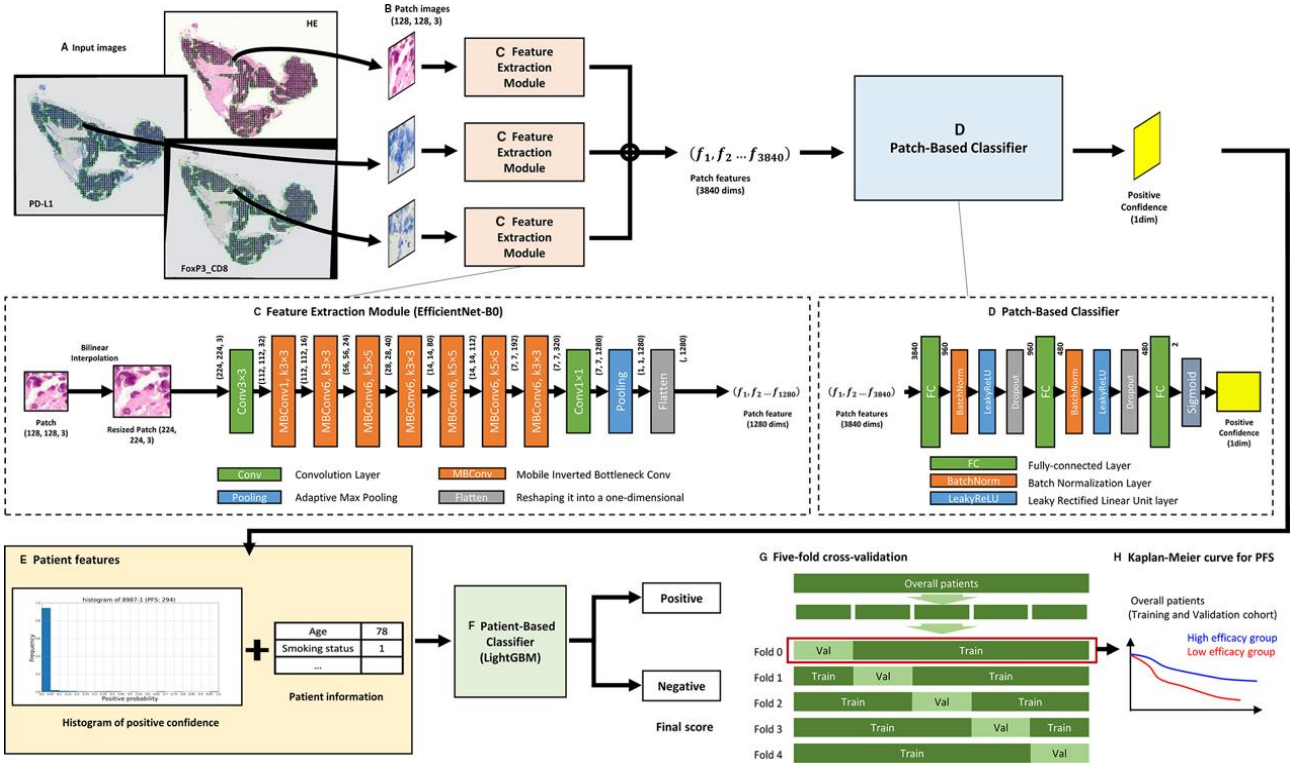

Figure 2. Overview of Shibaki's model [12]

The research of Shibaki et al built an survival prediction based on image recognition using CNN (Figure 2) [12]. The model used the EfficientNet-B0 model to process patients' image to obtain confidence scores for 1-year PFS for each patch. This is an application of large pre-trained CNN models, which is widely used in CNN-based methods [11]. Then the LightGBM model is used to perform the final classification prediction. This is a representative example of using CNN for image recognition to perform survival prediction. The mean AUC of this model is 0.782, which illustrated the well performance of the model. However, the training dataset was relatively small, and the model lacked validation in real world data, which may lead to overfitting.

Except for the two-dimensional pictures, CNN can extract features from high dimensional data such as CT scans. Mahootiha's research used the 3D CNN architecture trained with discrete LogisticHazard-based loss to perform the survival prediction and gets a mean AUC of 0.8 [13]. Furthermore, Mahootiha et al used the spearman score and random forest to select clinical variables. The combination of deep learning network and traditional machine learning method gave explainable importance of clinical variables, which made up for the low interpretability of deep learning algorithms. Shibaki and Mahootiha's research showed great ability of CNN in medical image analysis and survival prediction.

Deep learning algorithms represented by Artificial Neural Network (ANN) and Recurrent Neural Network (RNN) can also applied in survival prediction based on patient data. Especially, the RNN can

memorize short-term memory, so that it can be applied to methods which have requirements of dealing with sequential data. In the survival prediction of bladder cancer patients, the ANN is the most commonly used machine learning algorithm because of its ability to detect complex nonlinear relationships [11]. However, the performance of deep learning methods may be worse than other machine learning algorithms. In Quannian's research, the ANN algorithm performs worse than the random forest [4].

In general, different kinds of deep learning algorithms have applications in survival prediction, and different architectures of neural network algorithms have their own advantages in different areas. Researchers can choose deep learning algorithms carefully according to their actual requirements.

6. Comparisons between Survival predicting methods

6.1. Common public datasets

Table I presents the basic information of different datasets.

Table 1. The description of common public datasets.

Dataset	Description
Surveillance, Epidemiology, and End Results (SEER)	Data on cancer incidence, survival, and treatment across the United States to support cancer research and epidemiological investigations.
The Cancer Genome Atlas (TCGA)	patient samples from multiple cancer types, providing clinical characteristics, gene expression data, mutation information, and more.
Gene Expression Omnibus (GEO)	clinical information and gene expression data from cancer patients

6.2. A comprehensive performance metric

A predictive model needs metrics to estimate their performance in predicting real world data. Currently, the most commonly used method to estimate the performance of a model is to utilize the AUROC (Area Under the Receiver Operating Characteristic curve).

AUROC is derived from plotting the Receiver Operating Characteristic (ROC) curve, which utilizes the true positive rate (also known as sensitivity or recall) as the y-axis and the false positive rate as the x-axis. The AUROC value ranges between 0 and 1, where 0.5 represents a model that performs no better than random guessing, and 1 indicates a perfect classifier. A higher AUROC value implies better model performance, indicating a greater ability to accurately distinguish between positive and negative instances.

An AUROC value of 1.0 signifies a flawless discriminator, while values ranging from 0.90 to 0.99 are considered excellent, 0.80 to 0.89 as good, 0.70 to 0.79 as fair, and 0.51 to 0.69 as poor[2]. The AUROC can estimate model's performance comprehensively, so that this paper will use the AUROC as the standard for model performance comparison.

6.3. Analysis and comparison of experimental results

According to current research about machine learning algorithms, the table II shows the mean AUROC of each common algorithm. The data comes from paper [3].

Table 2. The mean AUROC of each common algorithm.

ML Category	ML Algorithm	Mean AUROC
Traditional ML Algorithm	Random Forest	0.8084
	Boosted Tree	0.7876
	SVM	0.7633
	Regularization Methods	0.7164
	Naïve Bayes, KNN or MLP	0.7899
Deep Learning Algorithm	ANN	0.7999
Traditional Method	CPH Model	0.77

The mean AUROC of Naïve Bayes, KNN or MLP in Table I represents the mean AUROC of these three algoritms. Table I illustrates that both the random forest and ANN outperform other algorithms, with the random forest exhibiting superior performance compared to ANN. Following these, Boosted Tree methods, SVM, Naïve Bayes, KNN, or MLP, along with the traditional Cox proportional hazards (CPH) model, demonstrate similar performance, occupying the second position in terms of AUROC. Notably, regularization methods exhibit inferior performance among the algorithms.

Studies [2] and [5] corroborate that the random forest outperforms Boosted Tree methods and other algorithms. Random Forest and Boosted Tree methods share a similar underlying principle as both are ensemble learning methods based on decision trees. The flexibility of decision trees in adapting to nonlinear relationships between features enables better capture and modeling of such relationships. Conversely, SVM and regularization methods excel in handling linear data. The superior performance of the random forest over Boosted Trees may stem from their distinct integration approaches. Each decision tree in Boosted Tree methods is generated based on the prior tree, while in random forest, each decision tree is generated independently. The iterative adjustment of sample weights in Boosted Tree methods during training may lead to certain features receiving excessive attention across multiple decision trees, resulting in the underutilization of relevant features in the model. In contrast, the random forest randomly selects feature subsets, ensuring relatively independent feature selection for each decision tree. This property enables better management of feature correlations in high-dimensional data and mitigating the risk of overfitting, particularly relevant in survival prediction tasks characterized by high-dimensional data.

Deep learning methods offer advantages in effectively representing and recognizing patterns in input data. However, training on large-scale datasets and complex network structures demands extensive time and computational resources. In contrast, the training process of the Random Forest is relatively straightforward, and computational efficiency can be enhanced through parallel processing. Therefore, in scenarios with large data volumes or limited computing resources, Random Forest may offer a more advantageous solution.

While the CPH model demonstrates performance comparable to traditional machine learning methods, it lags behind many machine learning algorithms such as Boosted Tree methods, random forest, and deep learning methods. This suggests that the advantages of machine learning methods are not universally evident [14]. However, the CPH model's ability to model covariate effects and provide risk ratio estimates holds significance for understanding and interpreting model results.

7. Conclusion

This paper provides a comprehensive overview of the current applications of machine learning algorithms in predicting the survival of cancer patients, offering comparisons among various predictive methods. It is observed that traditional COX regression, conventional machine learning algorithms, and deep learning techniques are extensively utilized in this domain, yielding commendable performance. Notably, the random forest algorithm and ANNs are prominently featured and

demonstrate superior accuracy. Additionally, COX regression, particularly when enhanced through diverse methodologies or integrated with machine learning frameworks, remains prevalent in clinical studies and exhibits satisfactory performance.

The investigation of predictive models serves to inform clinicians in treatment strategies and resource allocation effectively. However, there remain ample avenues for further exploration in future research endeavors. Firstly, researchers are encouraged to validate their models using clinical data to identify potential discrepancies. Collaboration across multiple centers facilitates the expansion of data scale and sample diversity, thereby enhancing the stability and reliability of the models. Secondly, there is a pertinent need for the integration of predictive models with clinical practice. The amalgamation of theoretical models with clinical insights can bolster medical decision-making processes.

Reference

[1] H. Zhang. Construction of prognosis survival model of pancreatic cancer after radical surgery based on machine learning. Southern Medical University, 2024.

[2] M. Karabacak, K. Margetis A Machine Learning-Based Online Prediction Tool for Predicting Short-Term Postoperative Outcomes Following Spinal Tumor Resections. Cancers (Basel). 2023 Jan 28;15(3):812.

[3] PN. Butow, JM. Clayton, R. Epstein. Prognostic Awareness in Adult Oncology and Palliative Care. J Clin Oncol. 2020 Mar 20;38(9):877-884.

[4] Q. Shao. Clinical study on the application value of LMR and AFR in predicting long-term survival after pancreatic ductal adenocarcinoma . Lanzhou University, 2023.

[5] Y. Huang et al. Application of machine learning in predicting survival outcomes involving real-world data: a scoping review. BMC Med Res Methodol. 2023 Nov 13;23(1):268.

[6] X. Hou. Study on Survival Analysis Model based on XGBoost. Daliann Maritime University,2023.

[7] L. Xu ,L Cai ,Z Zhu , G.Chen . Comparison of the cox regression to machine learning in predicting the survival of anaplastic thyroid carcinoma. BMC Endocr Disord. 2023 Jun 5;23(1):129.

[8] D Zhou et al. A prognostic nomogram based on LASSO Cox regression in patients with alpha-fetoprotein-negative hepatocellular carcinoma following non-surgical therapy. BMC Cancer. 2021 Mar 8;21(1):246.

[9] W. Chen. Cancer diagnosis and prognosis based on Deep learning. Xi'an University of Architecture and Technology,2023.

[10] E. Badisy I et al. Risk factors affecting patients survival with colorectal cancer in Morocco: survival analysis using an interpretable machine learning approach. Sci Rep. 2024 Feb 12;14(1):3556.

[11] Y. Liu et al. Use of machine learning to predict bladder cancer survival outcomes: a systematic literature review. Expert Rev Pharmacoecon Outcomes Res. 2023 Jul-Dec;23(7):761-771.

[12] R. Shibaki et al. Machine learning analysis of pathological images to predict 1-year progression-free survival of immunotherapy in patients with small-cell lung cancer. J Immunother Cancer. 2024 Feb 15;12(2):e007987.

[13] M. Mahootiha et al. Multimodal deep learning for personalized renal cell carcinoma prognosis: Integrating CT imaging and clinical data, Computer Methods and Programs in Biomedicine, Volume 244,2024,107978.

[14] Y. Ma et al. Comparison of differentiation performance between survival data machine Learning and COX model based on benchmark experiment. Modern preventive medicine,2023,50(13):2344-2348+2368.