ISSN 2959-6157

Research on the Application of Machine Learning in Cancer Prediction and Diagnosis

Chenyu Wang[1, *]

[1]Beijing Royal School, Beijing, China
*Corresponding author: wangchenyu@st.brs.edu.cn

Abstract:

Due to the continuous development and advancement of artificial intelligence technology, the application of machine learning in the medical field, particularly in cancer treatment and prediction, is becoming increasingly widespread and profound. By utilizing machine learning algorithms, healthcare professionals can analyze large volumes of medical data more accurately through specific models. This paper discusses the application of machine learning in cancer diagnosis and treatment, with special attention to prediction, diagnosis and treatment. By analyzing a large number of clinical and genomic data, machine learning, which provides support for early diagnosis, reveals the complex pattern of cancer development. Deep learning technology can help doctors diagnose and predict cancer more accurately in medical image analysis and genome data analysis. Ensemble learning methods such as the random forest model improve the accuracy of prediction. It also introduces semi supervised learning, which provides a new perspective for cancer prediction: enhancing model training by using unlabeled data. The paper also introduces the application of cox-net and survival support vector machine in survival analysis.

Keywords: deep learning; algorithm model; supervised learning; cancer prediction; survival analysis.

1. Introduction

Cancer is a disease characterized by abnormal cell growth and proliferation, which can form tumors in any part of the body. Lung cancer is one of the most common types, often associated with smoking and other factors. Early-stage lung cancer may not present obvious symptoms, but as the disease progresses, patients may experience symptoms such as coughing and difficulty breathing. Early detection and treatment of lung cancer are crucial for patient survival. Through machine learning, we can utilize extensive patient data, genetic data of individuals, and clinical observation data to predict cancer risks, diagnose tumor types, and forecast treatment responses and the likelihood of disease in suspicious patients. The advancement of machine learning brings new hope to cancer prediction, assisting doctors in accurately identifying patient risk factors, early tumor detection, and devising personalized treatment plans.

This paper explores the applications of machine learning and deep learning in cancer prediction and diagnosis. By analyzing vast clinical and genomic data, complex patterns of cancer development are unveiled, providing support for early diagnosis. The paper introduces the application of machine learning in cancer prediction, particularly how standardized patient data processing and analysis can forecast cancer risks. It also emphasizes the importance of data preprocessing, feature selection, and model training, as well as the necessity of balancing datasets to ensure model accuracy and reliability. Deep learning techniques exhibit significant potential in cancer research due to their ability to handle complex and high-dimensional data, aiding doctors in more precise cancer diagnosis and prognosis through genomic data analysis and medical image processing. This approach includes supervised learning and semi-supervised learning, where semi-supervised learning enhances model training by utilizing unlabeled data. In terms of models, the paper extensively discusses ensemble learning methods such as Random Forest and Extreme Random Trees, which combine multiple decision trees to enhance prediction accuracy. The paper also delves into the applications in survival analysis, such as COX-NNET and Survival Support Vector Machine, specialized models for handling survival time data.

2. Research Methodology

2.1 Data Preprocessing

The research methods for cancer prediction using machine learning typically follow a comprehensive process,

with many Python or Java projects available online for reference. The general process for collecting data used for prediction and model training usually involves gathering various data sources, such as clinical data from medical facilities or institutions where drugs or experiments are conducted, in a form that can be read by programs. This data is then standardized, cleaned, integrated, categorized, and subsequently processed for further programming.

GENDER	AGE	SMOKING	YELLOW_F	ANXIETY	PEER_PRES	CHRONIC I	FATIGUE	ALLERGY	WHEEZING	ALCOHOL	COUGHIN	SHORTNES	SWALLOW	CHEST PAI	LUNG_CAN
M	69	1	2	2	1	1	2	1	2	2	2	2	2	2	YES
M	74	2	1	1	1	2	2	2	1	1	1	2	2	2	YES
F	59	1	1	1	2	1	2	1	2	1	2	2	1	2	NO
M	63	2	2	2	1	1	1	1	1	2	1	1	2	2	NO
F	63	1	2	1	1	1	1	1	2	1	2	2	1	1	NO
F	75	1	2	1	1	2	2	2	2	1	2	2	1	1	YES
M	52	2	1	1	1	1	2	1	2	2	2	2	1	2	YES
F	51	2	2	2	2	1	2	2	1	1	1	2	2	1	YES
F	68	2	1	2	1	1	2	1	1	1	1	1	1	1	NO
M	53	2	2	2	2	2	1	2	1	2	1	1	2	2	YES
F	61	2	2	2	2	2	2	1	2	1	2	2	2	1	YES
M	72	1	1	1	1	2	2	2	2	2	2	2	1	2	YES
F	60	2	1	1	1	1	2	1	1	1	1	2	1	1	NO
M	58	2	1	1	1	1	2	2	2	2	2	2	1	2	YES
M	69	2	1	1	1	1	1	2	2	2	2	1	1	2	NO

Fig. 1 Standardized Classification Example for Predictive Data

The dataset in Figure 1 comprises standardized cancer patient data publicly available from American hospitals, covering factors that may influence cancer risk, such as levels of nutritional balance, duration of smoking, and the presence of respiratory abnormalities. These factors have undergone fuzzification processing, transforming them into data represented as 1 or 2 to enhance data readability, processing operability, and improve the accuracy of predictive results.

2.2 Variable Impact

Following data preprocessing, the subsequent analysis typically involves feature selection and extraction, as well as data processing using computational programs like n-spire. Different statistical methods are employed in various scenarios, such as t-tests for comparing differences in normally distributed data, Mann-Whitney U tests for comparing non-normally distributed data, and chi-square tests for comparing data sources from different disease conditions. By calculating p-values and comparing them with the significance level (α significant level), the correlation and impact between variables can be determined.

A correlation coefficient heatmap is a clear visualization method used to display the degree of impact of each variable on the summarized data.

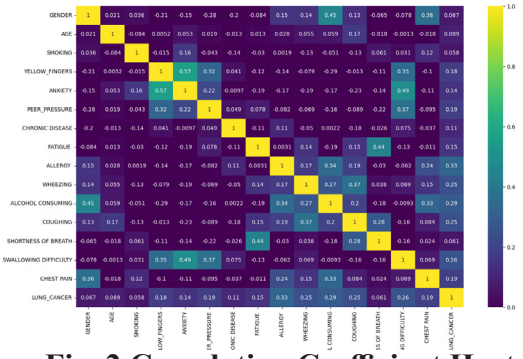

Fig. 2 Correlation Coefficient Heatmap

As shown in the example in Figure 2, only a few variables meet the criteria with p-values less than the given α level (in this case, the α level is 0.05), indicating that multiple variables have a minor impact on the overall outcome. In the subsequent model training process, variables with a significant impact on the outcome will be selected for prediction. Model selection relies on the ROC curve area under the curve (AUC) as a measure of model reliability. Key metrics considered include model recall rate, precision, F1 score, and accuracy[1].

3. Overview of Machine Learning

3.1 Deep Learning

Deep learning is a machine learning method that, through learning from large-scale data, can uncover patterns hidden within the data, providing more possibilities for early cancer diagnosis and treatment. It primarily utilizes deep neural networks and other deep structures to learn and understand data. Deep learning processes complex data through hierarchical abstract representations, such as medical imaging and genomic data, reducing the need for extensive feature engineering to some extent. The

application of deep learning in cancer prediction and treatment is extensive, where models trained using deep learning techniques like Random Forest and Block Forest provide more accurate depictions of cancer development trends. Genomic data analysis is one of the significant applications of deep learning in cancer prediction. Through deep learning techniques, doctors can analyze large-scale genomic data to discover gene variations and mutations related to cancer, thereby predicting patients' cancer risks. Personalized medicine will advance from this, enabling medical institutions to provide more precise treatment plans for patients. Targeted therapies based on specific cancer cell gene mutations are a leading trend in healthcare, and predictive models derived from extensive data using deep learning techniques are indispensable.

Furthermore, deep learning plays a crucial role in medical image analysis. By employing deep learning algorithms to analyze medical imaging data sets such as X-rays, MRIs, and CT scans, doctors can more accurately identify tumors, assess their malignancy, conduct early diagnosis, identify morphological features of tumor cells, evaluate tumor growth rates, and even predict the risk of tumor metastasis. This information is vital for formulating treatment plans and predicting patient survival rates.

In clinical data mining, deep learning helps doctors extract potential patterns and features related to cancer from vast clinical data. For instance, analyzing patients' clinical records and treatment response data can aid doctors in predicting the progression of patients' conditions and provide a basis for adjusting treatment plans. Deep learning can also be utilized for predicting tumor metastasis. By analyzing patients' clinical and imaging data, models can forecast the risk of tumor metastasis. For example, in lung cancer patients, the model can analyze tumor growth rates, morphological features, and imaging characteristics to predict the likelihood of tumor metastasis. An example is provided using a Random Forest model for learning. Random Forest is a widely used machine learning algorithm composed of multiple decision trees, each independently trained. Its main parameters include the number of trees, maximum tree depth, minimum samples per leaf, and minimum samples per split. These parameters sequentially decrease the complexity of the Random Forest model. Optimizing the model based on the area under the ROC curve (AUC), a search is conducted for the four main parameters, starting with the number of classifiers. During iteration, the maximum AUC is sought, and the optimal number of iterations is set. Grid search is performed for the maximum tree depth and minimum samples per split, and the best parameters are then applied to the subsequent model, as shown in Figure 3.

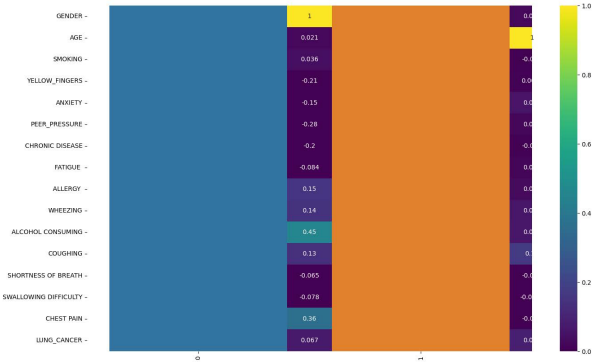

Fig. 3 Heatmap Showing the Degree of Influence of Different Variables on the Outcome.

3.2 Semi-Supervised Learning

Semi-supervised learning algorithms are a machine learning technique that utilizes training data containing both labeled and a significant amount of unlabeled data. Typically, supervised learning only utilizes labeled data for model training, but in semi-supervised learning, unlabeled data can also be leveraged, especially in cases where labeled data is limited. Unlabeled data plays a crucial role in boosting model performance and generalization in supervised learning, while in unsupervised learning, it helps the model discover intrinsic structures or patterns from unlabeled data. These unlabeled data, such as medical imaging or genomic data, do not directly lead to specific outcomes but can aid in identifying potential patterns or features, thereby enhancing predictive model performance. Labeled data is essential for training supervised learning models as the model learns the relationship between input data and corresponding labels. This enables the model to predict unseen data and make reasonable predictions for new input samples. Although many semi-supervised learning algorithms include a large amount of unlabeled data in the training data, there are some strict requirements for the training data in many semi-supervised learning algorithms. For instance, unlabeled data typically comes from labeled data of known categories rather than belonging to other categories or containing multiple categories; the labels of labeled data should be accurate; unlabeled data should maintain class balance, with a similar number of samples for each category; and the distribution of unlabeled data should be similar to labeled data, among others. In general, semi-supervised learning algorithms can be categorized into several types: self-training algorithms, graph-based semi-supervised algorithms, and Semi-Supervised Support Vector Machine (S3VM). Below is a brief introduction to two of these algorithms:

Self-training: Initially, a classifier is trained using labeled

data, and then the classifier is applied to classify unlabeled data, generating pseudo-labels or soft labels. Subsequently, unlabeled samples deemed correctly classified are selected and used to retrain the classifier.

Co-training: This method is essentially a form of self-training but with a more sophisticated approach. It assumes that each data point can be classified from different perspectives, allowing the training of different classifiers. These classifiers are then used to classify unlabeled samples from different perspectives, selecting high-confidence unlabeled samples to add to the training set. As these classifiers are trained from different perspectives, they complement each other, thereby improving classification accuracy, much like understanding things from different angles.

3.3 Models

3.1.1 Random Forest Model

The Random Forest model is an ensemble learning method consisting of multiple decision trees, each independently trained. The core idea of Random Forest is to combine the predictions of multiple decision trees to obtain more accurate and robust overall predictions. For classification problems, the majority voting method is typically used, where the class with the most votes from each tree's predictions is considered the final prediction. For regression problems, the average method is commonly employed, where the predictions of each tree are averaged to obtain the final prediction[2]. Extreme Random Trees (ET) is another ensemble learning technique that aggregates results from multiple decor-related decision trees in the forest to output classification results. Each tree in the Extreme Random Trees model is built from the original training samples, with a random sample at each test node containing k features. Each decision tree must select the best feature from this feature set, then split the data based on certain mathematical metrics (usually the Gini index)[3]. In scikit-learn, the classifier for Extreme Random Trees is ExtraTreesClassifier.

3.1.2 Cox-NNET Model

In 1995, Faraggi and Simon proposed a predictive model called COX-NNET. This model uses neural networks to predict the survival rate of prostate cancer using four clinical pieces of information[4]. However, the accuracy of this model is similar to traditional methods. Subsequently, other researchers proposed different modeling methods, some of which simplify survival prediction into a binary classification problem or discretize survival time into different time intervals. However, due to information loss, these methods may lead to reduced prediction accuracy. Travers Ching et al. introduced an artificial neural network model called COX-NNET for predicting survival rates from high-dimensional genomic data. COX-NNET is an extension of the COX proportional hazards regression model, capable of handling complex nonlinear relationships in the data. Its model structure includes an input layer, hidden layers, and a Cox regression layer[5].

3.1.3 Survival SVM Model

Survival SVM Model: Support Vector Machines (SVM) proposed by Corinna and Vapmik in 1995 have been widely used in the classification and regression fields due to their solid theoretical foundation and numerous good properties. The main challenge in applying Support Vector Machine models in survival analysis is the presence of censored samples. Ignoring censored samples can lead to underestimation of failure time while treating censored samples as events not occurring can result in biased models. To introduce Support Vector Machines into survival analysis, Evers et al. proposed a Support Vector Machine method for survival analysis that includes ranking constraints, transforming the prediction problem into a ranking problem for solving. Van Belle et al., after comparing different methods, proposed a Survival Support Vector Machine (Survival SVM) that includes both ranking and regression constraints, which has good predictive performance while having a solid theoretical foundation.

4. Cancer Identification

Identifying biologically significant genes using the COX-ResNet model can reveal crucial processes underlying disease mechanisms. Here is an example: Based on the analysis of the Cox-ResNet model for the bladder cancer (BLCA) dataset, a significant gene analysis heatmap yielded the following conclusions: by dividing patient samples into high-risk and low-risk groups based on prognostic index scores, comparing the Kaplan-Meier survival curves and Log-Rank test p-values of the two risk groups, it was found that the survival curves of the two risk groups had a statistically significant difference (p-value ≤ 0.01). Using the Guided Grad-CAM method to generate heatmaps for high-risk group patients highlighted important areas in the prediction process, ultimately determining prognostic genes through heatmap analysis. In the heatmap of the high-risk population, a trend in gene significance scores was identified, and through slope analysis, the top 500 genes were determined to be effective prognostic genes, consistent with a small fraction of biomarkers among the total genes[6].

Another example involves researchers using a mixed machine learning method of Random Forest and Decision Trees for predicting the classification of primary tumors. Through 10-fold cross-validation and gene selection,

they found that when Random Forest used 300 genes, the mixed classifier provided optimal performance, achieving an accuracy rate of 88.5%. In terms of decision tree construction, they employed the CART algorithm and conducted post-pruning to prevent overfitting. On the training set, they achieved a classification prediction accuracy of 88.9%. Additionally, the results of classifying 20 known primary metastatic cancers showed an accuracy rate of 85%. Independent validation results indicated an overall accuracy rate of 42% on a dataset of 48 GEO samples[7].

5. Conclusion

Although machine learning shows great potential in cancer prediction, this study also identified some drawbacks. The performance of machine learning models highly depends on the quality and quantity of data. In cancer prediction, obtaining high-quality, accurately labeled medical data is often challenging, limiting the training and generalization capabilities of models and forcing researchers to use a large amount of unlabeled data. Additionally, some extensively trained models in machine learning are prone to overfitting on training data (overfitting occurs when a model is too tailored to the given data, making the trained model less generalizable), which may lead to performance degradation in practical applications. Overfitting issues are particularly severe when data is scarce or lacks diversity. Cancer development is influenced by various factors, and these factors may change over time. Machine learning models need regular updates to adapt to these changes, or methods such as using control groups when collecting data to balance variations, otherwise, the accuracy of predictions may decrease, leading to serious consequences.

Despite its drawbacks, machine learning and deep learning provide new tools and methods for early cancer diagnosis and treatment. These technologies not only improve diagnostic accuracy but also offer more personalized treatment plans for patients, ultimately enhancing patient survival rates and quality of life. Future research should continue to explore new applications of these technologies and address challenges encountered in practical applications: data accessibility, model interpretability, and effective integration of different types of data.

References

[1] Yuan Ke. Research on Liver Cancer Prediction Model Based on Interpretable Machine Learning. Jiangxi University of Finance and Economics, 2023.

[2] Dou Yuyang. Prediction and Analysis of Synthetic Lethal Gene Combinations in Cancer Based on Machine Learning and Statistical Inference. Nanjing University of Posts and Telecommunications, 2023.

[3] Shi Lihua. Application of Enriched Signaling Pathway p53-MDM2 Mediated by miRNA-29a in the Dynamics and Machine Learning in Breast Cancer Treatment Research. Yunnan Normal University, 2023.

[4] Mo Minghui. Construction of an Early Postoperative Recurrence and Metastasis Prediction Model for Colorectal Cancer Based on Machine Learning. Qingdao University, 2023.

[5] Lan Ning. Application of Block Forest in Integrating Clinical and Omics Data of Cancer Patients to Build Prognostic Prediction Model. Shanxi Medical University, 2023.

[6] Chen Wangwang. Research on Cancer Diagnosis and Prognosis Based on Deep Learning. Xi‘an University of Architecture and Technology, 2023.

[7] Liang Xin. Efficient Tumor Tracing Prediction Research Based on Hybrid Machine Learning Methods. Hainan Normal University, 2021.