Transactions on Computer Science and Intelligent Systems Research
ISSN: 2960-1800, eISSN: 2960-2238 | Volume 9
AIDML 2025

Application of Multiple Machine Learning Models based on Python in Predicting the Risk of Esophageal Cancer

Junhan Zhao

College of Basic Medical Sciences, Jilin University, Changchun, Jilin, China

Corresponding author: zhaoch9924@mails.jlu.edu.cn

Abstract. As health awareness rises and technology advances, machine learning has garnered significant attention in cancer prediction. This study focuses on esophageal cancer, a common and high-mortality digestive tract tumor, aiming to evaluate the predictive accuracy of different machine learning models, identify optimal models, and examine the role of hyperparameter tuning through random search in enhancing prediction accuracy for models with lower performance. The findings of the study indicate that Random Forest (RF), GradientBoosting(GB), and Extreme Gradient Boosting (XGBoost) perform best, with accuracy reach 1.00; the K-Nearest Neighbors (KNN) accuracy is 0.97; the Support Vector Classification (SVC) accuracy is about 0.58, and the SVC with random search for hyperparameter tuning reaches 0.97; the accuracy of logistic regression is 0.68, and after random search hyperparameter tuning, it can reach 0.83. This study provides meaningful insights into leveraging machine learning for cancer prediction, with the potential to enhance future diagnostic practices and therapeutic strategies.

Keywords: Machine learning; Python; esophagus cancer; prediction.

1. Introduction

Esophageal cancer (EC), which is the seventh most common cancer globally, presents a substantial challenge to public health worldwide [1]. Annually, over 600,000 individuals worldwide receive a diagnosis of esophageal cancer (EC) and regrettably, the five - year survival rate for those with EC remains below 20% [2]. About 75% of new cases and deaths from esophageal cancer worldwide occur in Asia, with most of them concentrated in China. In terms of gender differences, 71.5% of new cases of esophageal cancer worldwide are in men [3]. This gender imbalance is also reflected in the gender differences in the data set. Within Western countries, the principal risk factors contributing to esophageal cancer are tobacco use and alcohol consumption. Tobacco and alcohol interact synergistically, exerting a positive combined effect [4]. This interaction stimulates the process of carcinogenesis in normal tissues, ultimately leading to the development of esophageal cancer [2]. Tobacco use and alcohol intake are both considerations in the modeling process.

Over the past decade, remarkable progress has occurred in the domain of machine learning (ML), where the creation of highly complex algorithms and the refinement of data preprocessing methods have both witnessed substantial development [5]. Therefore, cancer prediction has undergone a profound transformation due to advancements in machine learning. Although there are many types of research on the prognosis, postoperative survival rate and metastasis rate of esophageal cancer by machine learning, there are few types of research on predicting whether to replace esophageal cancer, and most of them use a single model, while there are still few researches on using multiple models to predict and compare the best model. The capacity for systems to learn and enhance their performance is typically endowed by ML models through training, which eliminates the necessity for explicit programming [6]. As a result, researchers in other industries such as medicine can be possible to use machine learning to make predictions without needing the same programming skills as computer science professionals.

The principal aim of this study is to determine which machine learning (ML) algorithm exhibits the optimal performance when it comes to predicting esophageal cancer. The algorithms that were contrasted were logistic regression (LR), support vector classifier (SVC), Random Forest (RF),

gradient boosting (GB), Extreme Gradient Boosting (XGBoost), K-Nearest Neighbors (KNN), and Light Gradient Boosting Machine (LGBM). At the same time, for the model with low prediction accuracy, random search is used to perform hyperparameter tuning, and the two results are compared to observe the influence of hyperparameter tuning on the prediction rate before and after [7][8]. All machine learning algorithms are based on Python's sklearn library with the dataset from Kaggle.

Traditional machine learning algorithms show outstanding performance in specific tasks like medical diagnostics. In these fields, the amount of available data is restricted, and the process of feature engineering plays a vital and decisive role [9]. Therefore, this study focuses on the selection of data sets and selects public data sets that have been manually reviewed on Kaggle. In real - world data analysis, data imbalance and missing values are major challenges. Imbalanced datasets have majority and minority classes, which, along with missing data, affects standard classifiers, causing bias and inefficiency, for instance, the forecast results are more skewed towards the majority class [10].

2. Datasets and Models

2.1. Datasets

This study was based on a free public dataset containing comprehensive clinical data on esophageal cancer uploaded on Kaggle's official website, including a total of 3985 anonymous patient profiles, including patient demographics, clinical data, and cancer-specific attributes, which were one-dimensional data. Between 2012 and 2015, data were collected from patients aged 27-90 years from 10 countries by completing forms.

2.2. Methods

Whether to have cancer can be regarded as a binary classification problem in machine learning, which can be based on Patient demographics (e.g., age, gender), Tumor histology and staging information, Treatment history to predict Lymph node examination results. Therefore, the following models can be used, and by comparing the accuracy of different models, the model that is more suitable for predicting esophageal cancer can be selected. This section provides a brief introduction to the model.

Logistic Regression (LR) is widely used in medicine because of its strong interpretability and its applicability to multivariate modeling.

Support Vector Machine (SVM) a well-suited model to find specific features that influence diagnosis results within the massive volumes of data, can be used for classification or regression.

Random Forest (RF) can reduce multi-source and multi-dimensional data, which is mainly used for classification and regression.

Gradient Boosting (GB) mainly focuses on data processing accuracy and speed, and is one of the machine learning algorithms proven to be efficient in various research fields.

Extreme Gradient Boosting (XGBoost), is featured in high efficiency and robustness, with high accuracy in multiple fields. Overfitting can be mitigated by adding regularization terms to the model.

K-Nearest Neighbours(KNN) is widely used in text classification, image processing and other fields and is an algorithm for data mining. It can handle multiple types of data (such as numerical, nominal, ordinal, etc.), and has a certain robustness to missing values and outliers.

Light Gradient Boosting Machine (LGBM) has the characteristics of high efficiency and speed in processing large-scale and high-dimensional data.

3. Dataset Description and Pre-processing

A total of 3985 original data were labeled and converted into integers; original numerical data were still used for continuous variables; and categorical variables (such as gender, target category, etc.)

were sequentially encoded to ensure that they were properly formatted in the modeling process. Columns with more than 10% missing values are deleted and the remaining missing values are filled with the mean. Since the index in the column of 'person neoplasm cancer status' in the dataset is "WITH TUMOR" (0) with a total of 2235 cases, "TUMOR FREE "(1) with 1415 cases in total, not balanced. In order to make the subsequent model converge, positive and negative samples of the dataset should be balanced, that is, the number of patients with the disease and the data without the disease should be roughly consistent. Therefore, after processing the missing data, 1400 cases were randomly selected in the two categories through a random library, so that the ratio of cancer patients to cancer patients was 1:1, which was taken as the data included in the study, and the random number seeds were fixed to ensure sustainability.

After processing the missing data, the dataset included 2,800 entries representing individuals with and without esophageal cancer. It includes 27 attributes, of which 26 are predictors for identifying potential risk factors. The numerical characteristics and proportion of each attribute are shown in Table 1.

Table 1. The partial characteristics of predictors

Feature	Summary Statistics
weight	Mean: 74.8, SD: 17.0, Range: 41.0-138.0
stage_event_tnm_categories	Mean: 12.5, SD: 13.0, Range: 0-48
primary_pathology_age_at_initial_pathologic_diagnosis	Mean: 62.8, SD: 12.1, Range: 27-90
gender	0: 2409 (86%), 1: 391 (14%)
person_neoplasm_cancer_status	0: 1400 (50%), 1: 1400 (50%)
vital_status	1: 1732 (62%), 0: 1068 (38%)
tobacco_smoking_history	1.0: 801 (29%), 4.0: 554 (20%), 2.0: 553 (20%), 3.0: 549 (20%), 2.3261398176291794: 343 (12%)
alcohol_history_documented	1: 1929 (69%), 0: 831 (30%), 2: 40 (1%)
has_new_tumor_events_information	1: 1543 (55%), 0: 1257 (45%)
has_drugs_information	0: 2219 (79%), 1: 581 (21%)
has_radiations_information	0: 2130 (76%), 1: 670 (24%)
primary_pathology_esophageal_tumor_cental_location	0: 1981 (71%), 1: 736 (26%), 2: 70 (2%), 3: 13 (0%)
primary_pathology_initial_pathologic_diagnosis_method	1: 1662 (59%), 0: 613 (22%), 2: 445 (16%), 3: 80 (3%)
primary_pathology_primary_lymph_node_presentation_assessment	0: 2080 (74%), 2: 436 (16%), 1: 284 (10%)

The dataset variables are shown in Table 2., which provides a brief overview of each attribute included in the dataset, providing an initial understanding of factors potentially associated with esophageal cancer risk.

Table 2. Some key numeric classification labels represent the meaning

Feature	Description
gender	0 for MALE,1 for FEMALE
person_neoplasm_cancer_status	0 for WITH TUMOR,1 for TUMOR FREE
vital_status	0 for Dead,1 for Alive
tobacco_smoking_history	0 to 4 for each individual
alcohol_history_documented	0 for No,1 for Yes
has_new_tumor_events_information	0 for Yes,1 for No
has_drugs_information	0 for No,1 for Yes
has_radiations_information	0 for No,1 for Yes
primary_pathology_esophageal_tumor_cental_location	0 for Distal, 1 for Mid, 2 for Proximal,3 for nan
primary_pathology_initial_pathologic_diagnosis_method	0 for Other method, specify, 1 for Endoscopic Biopsy,2 for Surgical Resection, 3 for nan
primary_pathology_primary_lymph_node_presentation_assessment	0 for Yes,1 for nan,2 for No

Fig. 1 shows the distribution of samples divided by age and cancer risk. It shows that people aged 50-60 or 70-80 years have a higher risk of developing esophageal cancer. In addition, the samples with cancer have a larger outlier in terms of age than those without esophageal cancer.

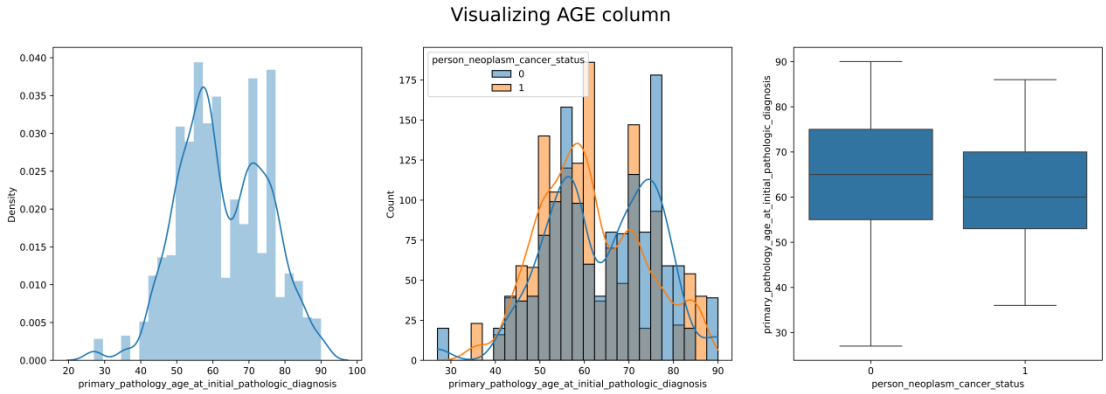

Fig. 1 Age distribution of patients with and without esophageal cancer (Picture credit: Original)

The above data were divided through random library into training sets 1:1120 (80%) and 0:1120 (80%) along with test sets 1:280 (20%), 0:280 (20%). Random seeds are fixed to ensure the repeatability of the experiment.

4. Experiment

4.1. Experiment Configuration

The models used are introduced in 2.2, all of which are called through Python's sklearn. SVC and logistic regression are used to construct models without using hyperparameter tuning with random search for hyperparameter tuning to compare the impact of hyperparameter tuning on prediction accuracy. Regularization terms are added to the XGBoost model to mitigate overfitting. Precision, recall, and f1-score were selected as evaluation indicators, and the results were visualized using the confusion matrix.

4.2. Experimental Results and Analysis

In this study, seven algorithms were used for training to find a more suitable model for esophageal cancer prediction. The data set provided by Kaggle is cleaned and processed for algorithm training. The training results are presented in Table 3.

Table 3. Evaluation indexes and accuracy of the model

Model		precision	recall	f1-score	accuracy
Logistic Regression	0	0.70	0.63	0.66	0.68
	1	0.66	0.73	0.70	
SVC	0	0.60	0.48	0.53	0.58
	1	0.57	0.68	0.62	
Random Forest	0	1.00	1.00	1.00	1.00
	1	1.00	1.00	1.00	
Gradient Boosting	0	1.00	1.00	1.00	1.00
	1	1.00	1.00	1.00	
XGBoost	0	1.00	1.00	1.00	1.00
	1	1.00	1.00	1.00	
KNN	0	0.96	0.98	0.97	0.97
	1	0.98	0.96	0.97	
LGBM	0	1.00	1.00	1.00	1.00
	1	1.00	1.00	1.00	
SVC (hyperparameter tuning)	0	0.96	0.97	0.97	0.97
	1	0.97	0.96	0.97	
logistic regression (hyperparameter tuning)	0	0.83	0.83	0.83	0.83
	1	0.83	0.83	0.83	

Logistic regression models have a higher recall rate for class 1 (0.73) but a lower accuracy (0.66), indicating that the model tends to predict more samples to class 1 and there may be false positives. The overall performance is mediocre and suitable as a baseline model. After hyperparameter tuning, the performance is significantly improved, but there is still a gap compared with other models, which may be related to the limitations of data distribution or linear assumptions of the model itself.

The performance of untuned SVC is poor, especially since the recall rate of class 0 is low, which may be due to an unreasonable classification boundary caused by kernel function or parameter optimization. After hyperparameter tuning, SVC performance is greatly improved and close to perfect classification, indicating that parameter optimization (such as kernel function, regularization term) is very important for the SVC effect.

Tree models including Random Forest, gradient lift, XGBoost and LGBM have recall of 1.00, F1 scores of 1.00 and accuracy of 1.00 in all categories. The perfect metric may suggest overfitting or data leakage, but the training set does not overlap with the test set, and feature importance indicates that the maximum value is 0.1187, and that there is no value close to 1. KNN is excellent and balanced, indicating that the local structure of the data is clear and suitable for distance-based classification.

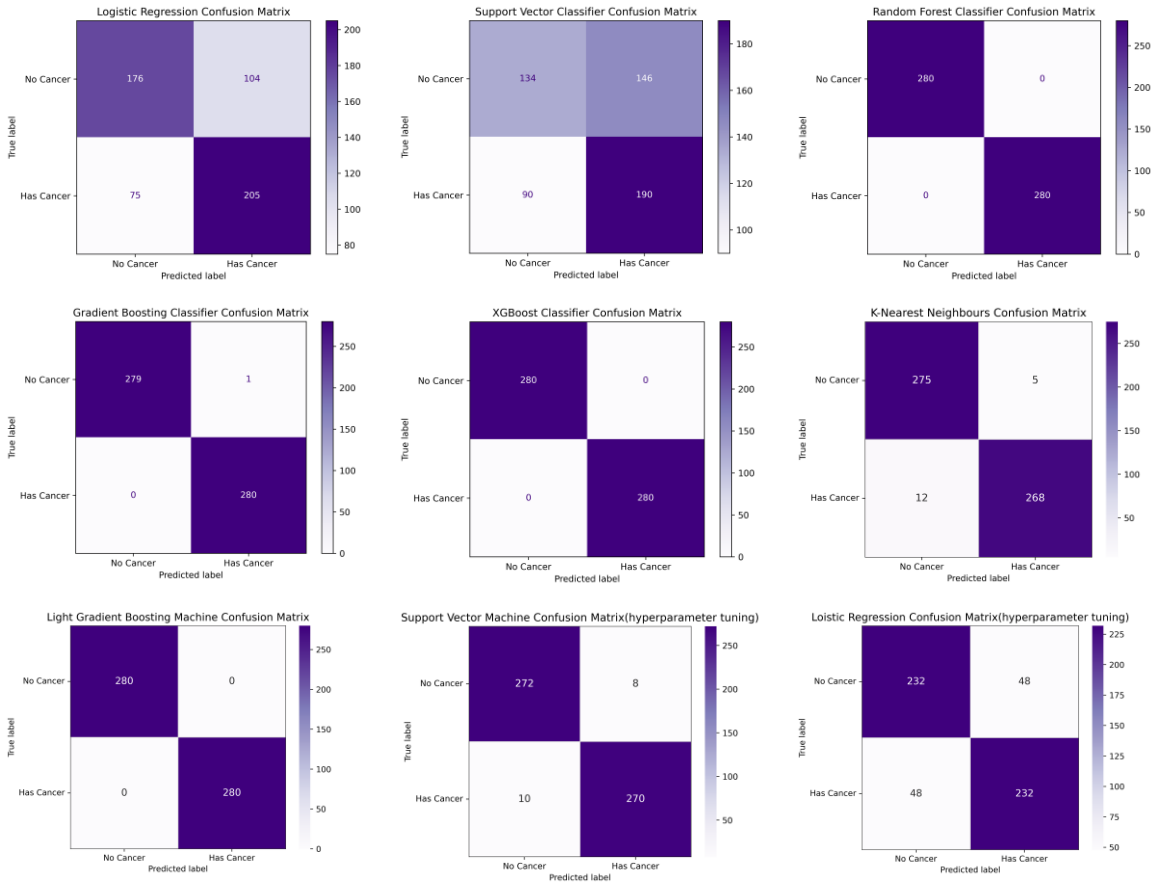

Fig. 2 Confusion matrix of model prediction results (Picture credit: Original)

According to Fig. 2, results can be intuitively acquired that Random Forest, GradientBoosting, and XGBoost effect best, with accuracy reaching 1. The KNN accuracy is 0.97. The SVC accuracy is about 0.58, and the SVC with random search for hyperparameter tuning reaches 0.97. The accuracy of logistic regression is 0.68, and after random search hyperparameter tuning, it can reach 0.83. Through comparison, it is found that a random search for hyperparameter tuning is significantly helpful in improving the accuracy of esophageal cancer prediction.

5. Conclusion

In conclusion, this essay aims to explore and investigate the accuracy of various machine - learning models in the prediction possibility of cancer which is crucial for improving public health. It aims to identify more appropriate models and also to deliberate on how hyperparameter tuning via random search can enhance the prediction accuracy of models with a relatively low initial prediction rate. To recapitulate, the significance of this research lies in its ability to provide multiple model predictions based on the same data set to facilitate the search for a more suitable model and. This research contributes to the field by using machine learning to predict cancer and finds suitable models

including Random Forest, GradientBoosting, XGBoost and KNN. The implications of this study are manifold, particularly in medical artificial intelligence. It is important to acknowledge the limitations of this study, which include the lack of a recent dataset, and the generalizability of its findings. Further investigation is warranted to use more models to make predictions across a wider set of data and improve the accuracy of esophageal cancer prediction. Ultimately, the findings underscore the importance of promoting the development of medical artificial intelligence to improve people's health. The results of this research prompt us to think about what else can machine learning do in interdisciplinary practice, and to push people to research more artificial intelligence to help in medicine.

References

[1] N. Deboever, C. M. Jones, K. Yamashita, J. A. Ajani, and W. L. Hofstetter, Advances in diagnosis and management of cancer of the esophagus, BMJ, vol. 385, 2024.

[2] J. Li, J. Xu, Y. Zheng, Y. Gao, S. He, H. Li, et al., Esophageal cancer: Epidemiology, risk factors and screening, Chinese Journal of Cancer Research, vol. 33, no. 5, pp. 535, 2021.

[3] Y. J. Zheng, Y. Teng, S. Y. He, M. D. Cao, Q. R. Li, N. P. Tan, et al., Epidemiological characteristics of esophageal cancer worldwide and in China, 2022, China Cancer, vol. 34, no. 3, pp. 165-170, 2025.

[4] X. Yang, Z. Tang, J. Li, and J. Jiang, Esophagus cancer and essential trace elements, Frontiers in Public Health, vol. 10, p. 1038153, 2022.

[5] C. Janiesch, P. Zschech, and K. Heinrich, Machine learning and deep learning, Electronic Markets, vol. 31, no. 3, pp. 685-695, 2021.

[6] I. H. Sarker, M. M. Hoque, M. K. Uddin, and T. Alsanoosy, Mobile data science and intelligent apps: Concepts, AI-based modeling and research directions, Mobile Networks and Applications, vol. 26, no. 1, pp. 285-303, 2021.

[7] R. El Shawi, M. Bahman, and S. Sakr, To tune or not to tune? An approach for recommending important hyperparameters for classification and clustering algorithms, Future Generation Computer Systems, vol. 163, p. 107524, 2025.

[8] J. Kossen, N. Band, C. Lyle, A. N. Gomez, T. Rainforth, and Y. Gal, Self-attention between datapoints: Going beyond individual input-output pairs in deep learning, Advances in Neural Information Processing Systems, vol. 34, pp. 28742-28756, 2021.

[9] D. Bhattacharyya, B. Dinesh Reddy, N. M. J. Kumari, and N. T. Rao, Comprehensive analysis on comparison of machine learning and deep learning applications on cardiac arrest, J Med Pharm Allied Sci, vol. 10, no. 4, pp. 3125-3131, 2021.

[10] H. Cheng, KNN-SVM classifiers in complex diagnosis, Journal of Physics: Conference Series, vol. 2694, no. 1, p. 012081, 2024.