ISSN 2959-6157

Exploring the Application of Machine Learning to Cancer Prediction

Xianwen Jiang

Department of Animation, Tokyo Polytechnic University,2-9-5 Honcho, Nakano, Tokyo,Japan
*Corresponding author: A2225045@st.t-kougei.ac.jp

Abstract:

This paper explores the wide range of applications of machine learning techniques in the field of cancer, with a particular focus on their specific use in the diagnosis and classification of important cancer types such as lung, oral and breast cancer. The paper concludes that machine learning algorithms can assist physicians in detecting cancerous lesions earlier and improve the accuracy of diagnosis. In addition, the paper explores the importance of machine learning in the early detection and treatment of cancer and its potential for collaboration with clinicians. In the future, collaborations across datasets and across healthcare institutions will drive further development of machine learning algorithms, providing more possibilities for personalized medical diagnosis and treatment plans to maximize patient survival and quality of life. The research in this paper can give relevant readers with insight into the potential and application of machine learning in the field of cancer, as well as its important role in improving the efficiency and quality of healthcare services.

Keywords: Machine learning,Prediction,Cancer.

1. Introduction

A wide range of illnesses that can affect almost any organ or tissue in the body are included in the term “cancer.” It develops when aberrant cells multiply uncontrollably, beyond normal boundaries, and then invade neighboring body regions or spread to distant organs. This process is called metastasis. The main factor contributing to cancer-related death is this metastatic spread[1].

Cancer stands as a significant public health challenge globally, contributing substantially to the burden of disease. According to WHO’s 2020 estimates, cancer ranks among the primary or secondary causes of death before the age of 70 in most nations. Since 2010, cancer has held the leading position in mortality rates in China, witnessing an escalation in morbidity, mortality, and overall burden. GLOBOCAN’s 2020 database reveals that nearly 10 million deaths worldwide are attributed to cancer annually, with China accounting for approximately 30.15% of these fatalities. With the aging demographic, cancer-related deaths are anticipated to continue rising both globally and within China, further exacerbating the public health crisis. While early detection and effective treatment can render many cancers curable, the imperative lies in the timely and precise prediction of cancer occurrence[2].

Major public health issues like cancer can be resolved by integrating machine learning techniques into clinical practice, which also increases the effectiveness and caliber of healthcare provided. These days, the phrase “big data” is frequently used to refer to data that is produced quickly and in large quantities, making it impossible to process and analyze manually. In clinical practice, big data can be gathered from many sources, including diagnostic imaging and electronic health records. Furthermore, a lot of wearable technology (such as sensors, mobile apps, and medical equipment) continuously gather data and upload it to pertinent databases[3].Machine learning techniques utilize these data to discover relevant information and assist in clinical decision-making.

This study will explore and summarize the application of machine learning techniques in cancer. This paper will introduce different techniques and differences in machine learning. And introduce the way machine students are used in cancer prediction.

2. machine learning definition

Within the fields of computer science and artificial intelligence (AI), machine learning (ML) focuses on using data and algorithms to help AI systems mimic human learning processes, gradually improving their accuracy [4].

Machine learning is classified into four types based on the nature of the problem and the available data: supervised learning, semi-supervised learning, unsupervised learning, and reinforcement learning. Among these, supervised learning and unsupervised learning are the most prevalent. In the realm of machine learning, supervised learning holds a dominant position, accounting for approximately 70% of applications, particularly in tasks such as image classification. Conversely, unsupervised learning comprises around 10.00-20.00% of machine learning applications[5].

Supervised learning involves training algorithms using labeled datasets to accurately classify data or predict outcomes. As input data is provided to the model, it adjusts its weights until achieving proper fitting, a crucial aspect of the cross-validation process aimed at preventing overfitting or underfitting. This methodology enables organizations to address various real-world challenges at scale, such as sorting spam into separate folders from the inbox.

On the other hand, unsupervised learning makes use of machine learning algorithms to examine and group datasets that are label-free. Without human assistance, these algorithms work autonomously to uncover hidden patterns or groups within the data. Because of this feature, unsupervised learning is very suitable for tasks like pattern and image recognition, customer segmentation, and exploratory data analysis. Moreover, dimensionality reduction, or reducing the amount of features in a model, is made easier by unsupervised learning.

Semi-supervised learning bridges the gap between supervised and unsupervised approaches by utilizing smaller labeled datasets to guide classification and feature extraction from larger unlabeled datasets during training. This methodology addresses the issue of insufficient labeled data often encountered in supervised learning.

Reinforcement machine learning shares similarities with supervised learning, but instead of being trained on example data, the algorithm learns through iterative trials. By iteratively reinforcing successful results, the model evolves to develop optimal recommendations or policies for given problems.[4].

Machine learning technologies serve various functions, including aiding in diagnosis generation and treatment selection, predicting risks and stratifying diseases, minimizing medical errors, and enhancing productivity through classification or screening tools. For instance, these technologies can analyze radiology images to prioritize interpretation by human radiologists based on disease probability. Similarly, they can examine retinal images to identify patients with vision-threatening diseases, facilitating referrals to ophthalmologists [6] .

A broad spectrum of data is accessible for machine learning, encompassing clinicopathologic data, sociodemographic information, lifestyle factors, histological details, and imaging parameters[7].

3. Specific Applications of Machine Learning in Cancer

3.1 Application of Machine Learning in Lung Cancer Diagnosis

The following three case studies are all based on machine learning approaches to classify lung cancer. Although all three studies perform lung cancer classification, different datasets are used. Seenivasagam et al. used sputum color images for classification.Tabitha Peter et al. and Riel et al. used ct tumor images for classification.

Seenivasagam et al. introduced a machine learning strategy to classify lung cancer stages utilizing big data. This approach merges machine learning algorithms with Apache Spark architecture for active categorization, facilitating the effective management of sizable datasets. Apache Spark, a rapid cluster computing framework, evolves from Hadoop MapReduce and broadens its functionalities to encompass diverse computational tasks, spanning Interactive Queries and Stream Processing. The dataset comprises sputum color images acquired from a microscopy laboratory, initially processed via the MapReduce framework with image dimensions of 330 × 330. These images encompass a spectrum of cell types, yet the study concentrates on identifying lung cancer-relevant cells such as eosinophils, bronchial mucus, and squamous carcinoma. To achieve precise lung cancer stage classification, the proposed approach harnesses T-BMSVM [8].

A machine learning algorithm for disease classification based on lung cancer screening image data was proposed by Tabitha Peter et al. The classification of binary outcomes, distinguishing between malignant and benign states, is achieved through an ensemble of predictive classification models, including both linear and nonlinear approaches, integrated with various feature selection techniques. Elastic networks and support vector machines were integrated with feature selection methods based on linear combination or correlation. The dataset utilized comprised 200 CT scans from patients' lungs obtained at the University of Iowa Hospitals. Eligible patients were identified through a review of pathology and radiology reports, meeting specific criteria: a) presence of isolated lung nodules ranging from 5-30 mm, and b) confirmation of histopathologically malignant or benign nodules, or nodules exhibiting size stability for a minimum of 24 months. The investigation utilized 416 radiomics features to quantify nodal characteristics observed in CT images collected from diverse scanner protocols at the University of Iowa Hospitals. Radiomics, a technique for extracting numerous features from radiological images using data characterization algorithms, was employed. The integration of radiomic biomarkers with machine learning methods presents a promising diagnostic approach for tumor classification. [9].

Riel et al. used a linear regression model trained on a sample of 300 CT tumor images with model performance AUC (0.706-0.932). Feature inputs were clinical factors + nodal features on CT images. The final out conclusion linear regression model can perform malignant and benign classification of tumors as well as human observers [10].

The three case studies above demonstrate three things: Machine learning algorithm combined with Apache Spark design can classify the staging of lung cancer with high accuracy.Patients with histopathologically confirmed malignant nodules or histopathologically confirmed benign nodules or size stability for at least 24 months after disease classification of lung cancer screening image data using machine learning methods.Linear regression models can perform malignant and benign classification of tumors as well as human observers. Both the high-precision classification of lung cancer stages and the malignant and benign classification of nodules are crucial in cancer. The classification of lung cancer staging determines the cure rate of the patient, and the earlier the lung cancer is staged the higher the likelihood of cure rate. The earlier the lung cancer is staged, the higher the chance of cure. Highly accurate cancer staging gives patients an earlier and higher chance of cure.

3.2 Machine Learning in Oral Cancer

Alhazmi created an artificial neural network (ANN) model in 2021 with the goal of forecasting an individual's probability of developing oral cancer through the use of data pertaining to risk factors, systemic medical problems, and clinic-pathological aspects. Seventy-three cases with pathology reports and confirmed diagnoses were included in the study. Utilizing the ANN, a popular data mining algorithm, an artificial intelligence (AI) based prediction model was created. 29 patient-related variables were included in the model's construction. A training set of 54 cases (75% of the dataset) and a testing set of 19 cases (25% of the dataset) were randomly selected from the dataset. Based on the available datasets, the results highlight the promise of machine learning techniques in oral cancer diagnosis and screening. The results demonstrate how well the ANN can predict malignancy probability and improve positive predictive value, which makes it easier to forecast an individual's risk of oral cancer based on relevant risk factors, systemic medical problems, and clinic-pathological data[11].

3.3 Machine Learning in Breast Cancer

Using a network pretrained on around 1 million non-medical images from ImageNet, Huynh et al. pioneered the use of transfer learning in breast cancer imaging.This strategy has become more well-liked because of its established benefits. Likewise, given the significant overlaps between digital mammography (DM), digital breast tomosynthesis (DBT), and digitized screen-film mammograms (SFM), deep convolutional neural networks (DCNNs) can be trained on a blend of these modalities, leading to larger datasets and more effective computer-aided diagnoses (CADs). For example, studies have demonstrated that employing transfer learning between digitized SFM and DM can significantly improve the area under the curve (AUC) of DM classification. In another investigation, training on both SFM and DBT images was observed to markedly enhance the AUC from 0.81 to 0.90 ($P < 0.05$) compared to training solely on DBT images.

Using both DM and DBT scans, researchers investigated how well an artificial intelligence (AI) system performed on its own for breast cancer detection. According to their findings, the AI system can lower the recall rate while achieving sensitivity comparable to radiologists in a DM screening situation. In addition, it was shown that the AI system performed on DBT scans just as well as radiologists, but with a higher recall rate[12].

Nowadays, due to the multifactorial nature of cancer, it has become difficult to rely on the judgment of a single healthcare professional for cancer classification. Therefore, there is an increasing demand for intelligent algorithms to aid in cancer diagnosis.Haitham Elwahsh et al. proposed a deep neural learning based cancer prediction model (DNLC).DNLC consists of several phases:

In the initial phase, a deep network (DN) is used to select the most relevant features from the dataset. Subsequently, the deep neural network is trained using genomic or clinical data samples. Finally, the efficacy of the DNLC approach in early cancer detection is evaluated.

Five cancer datasets—covering colon, lung adenocarcinoma, squamous cell carcinoma, breast cancer, and leukemia tumors—are used by DNLC for categorization. The average accuracy of the DNLC approach is 93%, which is superior to alternative methods in all circumstances, according to the results[13].

3.4 Recommendations

The great potential of machine learning algorithms in the medical field lies in their ability to utilize complex medical images and clinical data to help doctors detect cancer lesions earlier and improve the accuracy of cancer diagnosis. By analyzing large-scale datasets, machine learning algorithms are able to identify potential cancer features and assist doctors in making more accurate diagnostic and treatment decisions.

In the future, the development of machine learning will require collaboration across different datasets and healthcare organizations. Creating larger and more diverse datasets by sharing data will allow machine learning algorithms to better understand different types of cancers and improve the accuracy and robustness of the algorithms. In addition, close collaboration with clinicians is critical. Machine learning algorithms can provide clinicians with more comprehensive patient information and personalized medical advice, leading to improved patient outcomes and survival rates.

Overall, the application of machine learning in healthcare will contribute to revolutionizing cancer diagnosis and treatment, providing patients with more timely and effective healthcare services, and improving their survival rates and quality of life.

4. Conclusion

In this paper, machine learning algorithms have been enumerated in the range of cancer applications, including cavity cancer screening and diagnosis, early cancer detection, and CT tumor image classification. Machine learning some of these methods are not inferior to specialized doctors, for example, the DNLC method in the paper has an average accuracy of 93% in its ability to detect cancer at an early stage, and the DNLC method outperforms other methods in all cases. But machine learning is ultimately a machine subject to error and there is no way to achieve 100% accuracy, but as a medical classification tool it can be very effective in predictive classification and assessing risk. For example, using machine learning for ct image classification, to learn the degree of risk of their own cancer, only medium and high risk of manual confirmation by the doctor, and low and medium risk for conservative treatment and observation. This can greatly save the labor cost of social medical resources and the time people spend in the hospital. In today’s society, going to the doctor is a very time-consuming affair. Machine learning cannot replace traditional healthcare. However it can be used as a medical tool to improve the efficiency and accuracy of medical treatment.

A new angle on cancer research is provided by machine learning, which creates opportunities for the creation of decision support systems that will enhance precision oncology. Early detection and monitoring have become essential elements in the management of cancer due to the fundamental relevance of early treatment. Present-day prediction models provide highly valuable screening and examination suggestions to doctors and high-risk patients, enabling them to make informed decisions. These models maximize the distribution of resources for public health while simultaneously improving the quality of life for patients. Through the integration of several parameters, including genetics, biomarkers, and demographics, these models provide doctors with more accurate guidelines for monitoring high-risk groups. Targeted screening and monitoring provide more effective treatment routes and enable the prompt identification of possible disorders.

More advances in cancer diagnosis and treatment can be expected with the continuous development and application of machine learning technology in the medical field. In the future, machine learning algorithms will be more intelligent and personalized, able to provide customized diagnosis and treatment plans based on a patient’s specific situation. In addition, as medical data accumulates and is shared, collaboration across datasets and organizations will become more common, thus improving the accuracy and reliability of machine learning algorithms. At the same time, machine learning technology may also be combined with other advanced technologies such as gene editing and immunotherapy to bring more treatment options and hope to cancer patients.

5. Reference

[1]WorldHealthOrganization.health-topics/cancer.[DB/OL].(2024/4/21).https://www.who.int/zh/health-topics/cancer#tab=tab_1

[2]WorldHealthOrganization.news-room/fact-sheets/detail/cancer.[EB/OL].(2024/4/21).https://www.who.int/zh/news-room/fact-sheets/detail/cancer

[3]Nwanosike, E.M., Conway, B.R., Merchant, H.A., & Hasan, S.S. (2021). Potential applications and performance of machine learning techniques and algorithms in clinical practice: A systematic review. International journal of medical informatics, 159, 104679 .

[4]IBM.Meachine-learning[DB/OL](2024/4/21),https://www.ibm.com/topics/machine-learning

[5]Wen, X., Guo, X., Wang, S., Lu, Z., & Zhang, Y. (2024). Breast cancer diagnosis: A systematic review. Biocybernetics and Biomedical Engineering.

[6]He, J.Y., Baxter, S.L., Xu, J., Xu, J., Zhou, X., & Zhang, K. (2019). The practical implementation of artificial intelligence technologies in medicine. Nature Medicine, 25, 30 - 36.

[7]Lopez-Perez, L., Georga, E., Conti, C., Vicente, V., García, R., Pecchia, L., Fotiadis, D., Licitra, L., Cabrera, M.F., Arredondo, M.T., & Fico, G. (2024). Statistical and machine learning

methods for cancer research and clinical practice: A systematic review. Biomedical Signal Processing and Control.

[8]Supriya, M., & Deepa, A.J. (2020). Machine learning approach on healthcare big data: a review. Big Data and Information Analytics.

[9]Delzell, D. A. P., Magnuson, S., Peter, T., Smith, M., & Smith, B. J. (2019). Machine Learning and Feature Selection Methods for Disease Classification With Application to Lung Cancer Screening Image Data. Frontiers in oncology, 9, 1393.

[10]Li, Y., Wu, X., Yang, P., Jiang, G., & Luo, Y. (2022). Machine Learning for Lung Cancer Diagnosis, Treatment, and Prognosis. Genomics, Proteomics & Bioinformatics, 20, 850 - 866.

[11]Hegde, S., Ajila, V., Zhu, W., & Zeng, C. (2022). Artificial intelligence in early diagnosis and prevention of oral cancer. Asia-Pacific journal of oncology nursing, 9(12), 100133.

[12]Balkenende, L., Teuwen, J., & Mann, R. M. (2022). Application of Deep Learning in Breast Cancer Imaging. Seminars in nuclear medicine, 52(5), 584–596.

[13]Haitham Elwahsh, Medhat A. Tawfeek, A.A. Abd El-Aziz, Mahmood A. Mahmood, Maazen Alsabaan, Engy El-shafeiy,A new approach for cancer prediction based on deep neural learning,Journal of King Saud University - Computer and Information Sciences,Volume 35, Issue 6,2023,101565.