Proceedings of the 2nd International Conference on Modern Medicine and Global Health
DOI: 10.54254/2753-8818/44/20240415

Diagnostic analysis of cancer based on machine learning

Xishi Wang

School of Big Data and Information Engineering, Guizhou University, Guizhou, 550025, China

2047665441@qq.com

Abstract. Cancer, as the second leading cause of death in the world, caused an estimated 9.6 million deaths in 2018, accounting for one-sixth of all deaths. Early detection and early treatment is the best solution for cancer, and now, through machine learning methods, we can achieve accurate judgment of cancer so as to realize precise treatment and reduce the mortality rate of cancer. During this discussion, we will focus on the applications of machine learning methods to diagnose breast cancer, prostate cancer, oral cancer, which use machine learning methods including neural convolutional networks, K-clustering, support vector machine (SVM), and so on. As of now, machine learning has achieved better results than other methods, but due to the importance and complexity of cancer diagnosis and the cost of human computational capacity for diagnosis, we still hope to find a more accurate and effective method to realize the accurate judgment of cancer, and then introduce it into real-life applications.

Keywords: machine learning, diagnosis analysis, cancer.

1. Introduction

Cancer is a leading cause of death worldwide, causing nearly 10 million (or nearly one in six) deaths in 2020 [1]. Early detection of cancer and timely and correct diagnosis and treatment can gain valuable treatment time for cancer patients, increase the cure rate of patients, and reduce the death rate of cancer.

Currently, diagnostic methods for cancer prediction include manual diagnosis, statistical diagnosis and machine learning diagnosis. Manual diagnosis has the disadvantages of low accuracy, long time, and more reliance on the experience of the diagnosing doctor, while statistical methods also have the problems of incomplete datasets and insufficient number of comparative databases. Machine learning maximizes diagnostic efficiency and accuracy compared to the first two methods. There are a lot of studies showed that machine learning methods can dramatically improve accuracy at about 15-25% when compared to traditional cancer diagnostic methods, and at a more fundamental level, machine learning can also help with our understanding of how cancer develop and progress [2]. More often than not, we see cancer diagnosis by combining statistical methods with machine learning methods, and in Xiangchun Xiong et al.'s report, data mining and statistical techniques were used to analyze breast cancer to achieve improved diagnostic accuracy, which can be achieved in a short period of time [3]. Of course, statistical methods are better at analyzing the overall situation, and James Ted McDonald et al. used the linked Census-Cancer Registry administrative database to statistically analyze cancer diagnoses to get the overall results confirming the existence of the healthy immigrant effect of cancer [4].

DOI: 10.54254/2753-8818/44/20240415

Mimic the human learning process through data- and algorithm-based judgments, machine learning is more superior than other methods, it can improve the cognitive ability of things through the corresponding methods, so that it can get more accurate results in a certain number of training, so that it can complete the diagnosis of cancerous parts of cancer patients in the shortest possible time, saving a lot of human and material resources. In the shortest time, it can provide doctors with the most accurate results, so that patients can get accurate treatment as soon as possible, and reduce the adverse consequences of cancer to cancer patients.

In this article, we focus on the contribution of machine learning in cancer diagnosis. We first introduced the basic definition of machine learning and the importance of machine learning in cancer diagnosis. Then, we specifically analyzed different cancers, mainly introducing the contribution of applying machine learning methods to cancer prediction, diagnosis and prognosis in breast cancer, prostate cancer and oral cancer.

2. Diagnostic Method

2.1. The basic definition of machine learning

Machine Learning is part of Artificial Intelligence and Computer Science which imitate the way our human learn by focusing on using data and algorithms, incrementally improving their own accuracy. Machine learning methods are generally categorized into three types, supervised learning, unsupervised learning and reinforcement learning, which are classified based on the expected outcome of the algorithm [5].

The history of machine learning dates back to 1959, Arthur Samuel, an employee created the word machine learning and become a leader in artificial intelligence and computer games [5]. The mainly cause of death worldwide is cancer [6]. Everyone is facing with the challenge of fighting against cancer [6]. According to the survey, there are about 96,480 people expected to die from skin cancer, for lung cancer are 142,670 people, for breast cancer are 42,260 people, for prostate cancer are 31,620 people, and for brain cancer are 17,760 people [6]. Every year, countless patients die of advanced cancer, and the earlier detection of cancer, the larger possibilities to save lives, but the detection of early cancer by human labor is time consuming and inaccurate results, therefore, we want to find a practical way to find diagnosis as well as evaluation of cancer, and machine learning methods have thus entered into the vision of various researchers. By using machine learning methods to diagnosis cancer for the first time was at the end of the 20th century, the same things happened to using the artificial neural network (ANN) methods and to using decision tree (DT) methods [7], since then, the application of machine learning is more extensive. Nowadays, through the application of machine learning methods, we can basically realize the diagnostic assessment of cancer for prediction and prognosis [7]. In this paper, we will introduce the prediction, diagnosis, evaluation and analysis of different types of cancer using different machine models and machine learning methods.

2.2. Diagnosing specific cancers through machine learning methods

2.2.1. Breast cancer. Breast cancer is becoming the biggest culprit in female mortality, which is caused by a variety of clinical, lifestyle, social, and economic variables [7], and early detection of breast cancer and its intervention can reduce mortality, and through machine learning tools, we can predict and diagnose breast cancer more accurately. Breast cancer detection uses mammograms, ultrasound, biopsy, and machine learning methods tests to find out if the disease is present [8]. Existing ways to diagnose the presence of breast cancer using machine learning are mitosis-focused detection, transcriptomics analysis and mammogram image analysis for different detection methods will be used for different machine learning methods for their diagnosis.

In the case of detection of breast cancer by dominant divisions, we have selected the following studies. Albayrak et al designed a model to extract features from convolutional neural networks (CNNs) which are used for training support vector machine and to detect mitotic divisions in the breast[8].

DOI: 10.54254/2753-8818/44/20240415

AlexNet was used to construct a CNN to classify benign mitotic divisions from malignant mitotic divisions using histopathological images, Hao Chen et al. designed a model to diagnose breast cancer by determining mitotic progression by training three networks which have different settings but are totally connected [9]. The results are in the form of scores, and by averaging those scores we can get a final output. The transcriptomics area aims to find out various aspects of RNA (ribonucleic acid) biology, which includes what are RNA sequence and structure and how RNA transcript, translate and their cellular function [9]. Transcriptomics analysis is used to classify cancers into certain molecular subtype with Diagnosis of clinical significance, therapeutic choice relevance or else [10]. DeepCC, a neural network model trained on the data of breast cancer, was tested on microarray data which can express gene independently and it turns out to be more efficiently and with higher accuracy compared to traditional machine learning methods [11]. Mammography image analysis is one of the more commonly used methods to detect breast cancer. In machine learning, there have existed many published studies that have shown that can be used to achieve the investigation and prediction of early breast cancer by SVM (support vector machine), decision trees, artificial neural networks, deep learning methods and so on. In convolutional neural networks, two effective training methods exist, one using pre-trained weights and the other using a stochastic procedure evaluated on two independent datasets [7]. The findings show that the pre-trained network accomplishes its task more efficiently.

2.2.2. Prostate cancer. Prostate cancer is the third leading cause of death in men, with highly likelihood of diagnosis in men [6]. Prostate cancer can be detected by several methods, such as rectal fingerprinting, PSA testing, and TRUS-guided prostate biopsy, but all of these methods have the disadvantage of being somewhat invasive or having low accuracy [12]. In machine learning, we usually use models constructed by multi parametric magnetic resonance imaging (mp-MRI) combined with multiple machine learning methods to diagnose and assess the characteristics and how serious is the prostate cancer.

In the case of detecting prostate cancer using linear/logistic regression methods, we find the following research results. There is a model of logistic regression designed by Iyanna and her team which can tell the difference between the transition zone of cancer and benign prostatic hyperplasia (BPH) by using mp-MRI [12], and also computed the area under the curve (AUC) of the logistic regression model by calculating the AUC with logistic regression methods, which was then used in a study by Kan et al. who reported that the linear/logistic regression method was shown to have diagnostic performance [12]. As a part of supervised learning, support vector machines have contributed equally to the diagnosis of prostate cancer. The biggest advantage of support vector machines lies in the classification problem, which can be achieved by maximizing the classification boundary and thus the hyperplane. Tang's team [12] had reviewed 64 cases that being examined by mp-MRI before having radical prostatectomy, and found that the detection rate of SVM-mpMRI was 74.9%, with an AUC of 0.82, which is the easiest way to demonstrate that the support vector machine method's superiority over manual detection. As one of the most popular deep neural networks Convolutional Neural Network (CNN), it has excellent performance in dealing with problems with image data. Convolutional neural networks have many layers which include convolutional, nonlinear, pooling and fully connected layers. In the full convolutional neural network developed by Wang et al [12] using mp-MRI, by using the images of volumetric prostate MR which obtained from seventy-nine patients for the development of a dataset for tumor detection, the final result 0.85 in generated accuracy in CNN methods was obtained as a promising result. In addition to the above applications of machine learning methods, we found that k-Nearest Neighbors, Decision Tree/Random Forest, and Naive Bayes are also able to diagnose prostate cancers. Anderson et al. [12], by designing a model combining logistic regression method and the nearest neighbor classification method, finally achieved 79 % predictive accuracy, identifying the most dangerous cancers at about 82% accuracy. AUC values for distinguishing cancerous tissues from non-cancerous tissues and for distinguishing cancerous tissues from suspicious tissues, respectively.

DOI: 10.54254/2753-8818/44/20240415

2.2.3. Oral cancer. Oral cancer, a kind of cancer about head and neck. It will do harm to our throat, mouth, lipsand tongue. As one of the most common cancer claimed by World Health Organization, oral cancer has been diagnosed 300,000 new cases every year worldwide [13]. In this section, we will focus on the K-means Clustering method and deep generative models in deep learning to demonstrate the detection and classification judgment for oral cancer.

The basic idea of K-means Clustering is to find a partitioning scheme of K clusters in an iterative manner, such that the clustering result corresponds to the minimization of the cost function, and the k-means clustering algorithm is relatively scalable and efficient for large datasets. The use of K-means Clustering algorithm is mentioned in Harnale et al. where each point in the data is compared with each central of it and the results are grouped using k-means [13]. Deep learning as a kind of machine learning method contains many branches, including Deep Generative Models. Generative models can randomly generate samples by learning the probability density of observable data, and we can regard the data generation process of Deep Generative Models can be regarded as the process of transforming the sample points of a priori distribution into the sample points of a data distribution. In practical applications, Chatterjee et al. proposed a computer-aided method for diagnosing oral pre-cancer/cancer using oral exfoliative cytology, which combined with forest classifier provide numbers randomly to achieve the accuracy of 94.58% at most [13].

2.2.4. Other cancers. In addition to the three cancers mentioned above, people also suffer from brain cancer, lung cancer, skin cancer and other cancers in their daily lives, and according to the cancer statistics of 2019 [14], we can find that in the America, every ten adults have one adult that are diagnosed having cancer, and about 16.67 percent of deaths worldwide are due to cancer. In the cases we have observed, convolutional neural networks outperform other machine learning methods [9], and their accuracy increases as the convolutional maturity increases. In the following part, we will represent the use of machine learning to diagnose and predict brain and lung cancer.

For identifying and predicting lung cancer, Shen Wen et al. constructed a multivariate convolutional neural network to solve the problem of the changing nodule size, it reached this purpose through using multi-crop pooling layer rather than maximum pooling layer to generate multi-scale features and it comes out with 87.14% of accuracy and 0.93% of specificity which are pretty good in perform [9]. Our brain is the most elaborate and important part in our body, regardless of what kinds of tumor existing, once any part of the brain is compressed in our brain, the tumor will certainly do harm to our bodies in different ways, so to diagnosis cancer in our brains and the separation of cancerous areas from healthy parts of the brain is a great challenge. Sérgio Pereira et al. investigated a CNN-based method for automatic segmentation of magnetic resonance images, after which the in Liya Zhao and Kebin Jia established fully connected convolutional neural networks (FCN) and conditional random fields (CRF) for brain cancer segmentation [9]. Through the combined use of multiple machine learning methods, the difficulty of brain cancer treatment, diagnosis, and judgment has been greatly reduced.

3. Conclusion

As we have analyzed, machine learning has made a huge contribution to cancer prediction, diagnosis, and by using different machine learning methods, the diagnostic accuracy of different types of cancers can be dramatically improved. Although machine learning has a great role in cancer diagnosis and its development prospect is very optimistic, however, we still face many problems. One of the first problems is the lack of available datasets, due to the lack of viable management tools, patient privacy can be compromised, so there are few publicly available datasets for researchers to train on, which leads to uncertainty and instability in using machine learning to do cancer diagnosis. Secondly, the imbalance of the training set can also cause the diagnosis accuracy to fluctuate, as the patient profiles in the training dataset are not equal, resulting in a certain tendency of the trained model to misdiagnose the cancer situation. Finally, due to the presence of a great deal of noisy data towards the real situation, the processing towards the dataset, the model training process will cause a great challenge, which may lead to a large error in the results.

DOI: 10.54254/2753-8818/44/20240415

For the above mentioned problems, there have been some researchers using the method of transfer learning, and the use of pre-trained models in large datasets and then migrated to a small dataset for training, and achieved a certain degree of success [14]. Nevertheless, we still hope to be able to have a sufficiently large training set to complete the training of the machine learning model, so as to achieve the best results. In addition, we hope to work out a general machine learning model, so that the model can complete the diagnosis of various types of cancer, reduce the complexity of the model, so that the cancer diagnosis can be popularized, universal, so as to achieve the original purpose of using machine learning means to diagnose cancer.

The prospect of diagnosing cancer by means of machine learning is one we can already foresee, but I think there is more to machine learning than just the end of the road. I think that in the future, we may be able to carry out real-time diagnosis through using machine learning means combined with the advantages of the Internet, through the use of certain wearable medical testing equipment, uploading the data to the cloud, and analyzing the data in the cloud through the machine learning model, and delivering the user's health status to the user's cell phone in real time, so that we can reduce the occurrence of the aggravation of diseases or even life-threatening events due to the delay in diagnosis, and so that we can reduce the number of cases in the future. At the same time, the complete transfer of disease information to the hands of professional doctors can bring timely and accurate treatment for patients, reducing medical costs and unnecessary waste of human and material resources.

References

[1] World Health Organization. (2022) Cancer. https://www.who.int/health-topics/cancer#

[2] Cruz, JA., Wishart, DS. (2007) Applications of machine learning in cancer prediction and prognosis. Cancer Inform. Commun., 2:59-77.

[3] Xiangchun, X., Yangon, K., Yuncheol, B., Dae Wong, R., Soo-Hong, K. (2005) Analysis of breast cancer using data mining & statistical techniques. In: Sixth International Conference on Software Engineering, Artificial Intelligence, Networking and Parallel/Distributed Computing and First ACIS International Workshop on Self-Assembling Wireless Network, Towson, MD, USA. pp. 82-87.

[4] McDonald, J.T., Farnworth, M. & Liu, Z. (2017) Cancer and the healthy immigrant effect: a statistical analysis of cancer diagnosis using a linked Census-cancer registry administrative database. BMC Public Health. Commun., 17(1):296.

[5] Tom M, M. (1997) Machine Learning. McGraw Hill publishing. Maidenhead.

[6] Munir, K., Elahi, H., Ayub, A., Frezza, F., Rizzi. (2019) A. Cancer Diagnosis Using Deep Learning: A Bibliographic Review. Cancers (Basel). Commun.,11(9):1235.

[7] Perwej, Y., Akhtar, N., Pant, H., Dwivedi, A., Jain, V. (2023) A Breast Cancer Diagnosis Framework Based on Machine Learning. International Journal of Scientific Research in Science, Engineering and Technology. Commun., 10: 118-132.

[8] Rania R., K., Mohammed Y., K. (2023) Comparison of machine learning models for breast cancer diagnosis. IAES International Journal of Artificial Intelligence. Commun., 12(1):415-421

[9] Munir, K., Elahi, H., Ayub, A., Frezza, F., Rizzi, A. (2019) Cancer Diagnosis Using Deep Learning: A Bibliographic Review. Cancers. Commun., 11(9):1235.

[10] Liu, YR., Jiang, YZ., Xu, XE. et al.(2016) Comprehensive transcriptome analysis identifies novel molecular subtypes and subtype-specific RNAs of triple-negative breast cancer. Breast Cancer Res. Commun., 18: 33.

[11] Gao F, Wang W, Tan M, Zhu L, Zhang Y, Fessler E, et al. (2019) DeepCC: a novel deep learning-based framework for cancer molecular subtype classification. Oncogenesis.Commun., 8:44.

[12] Nematollahi, H., Mosleh,i M., Aminolroayaei, F., Maleki, M., Shahbazi-Gahrouei, D. (2023) Diagnostic Performance Evaluation of Multiparametric Magnetic Resonance Imaging in the

DOI: 10.54254/2753-8818/44/20240415

Detection of Prostate Cancer with Supervised Machine Learning Methods. Diagnostics. Commun., 13(4):806.

[13] Dixit, S., Kumar, A., Srinivasan, K. (2023) A Current Review of Machine Learning and Deep Learning Models in Oral Cancer Diagnosis: Recent Technologies, Open Challenges, and Future Research Directions. Diagnostics. Commun., 13(7):1353.

[14] Zhu, W., Xie, L., Han, J., Guo, X. (2020) The Application of Deep Learning in Cancer Prognosis Prediction. Cancers. Commun., 12(3):603.