

The Advancements of Machine Learning Applications in Cancer Research and Treatment

Xinze Li^a

Hefei NO.8 High School, Hefei, China

Keywords: Cancer, Machine Learning, Neural Networks, Machine Deep Learning.

Abstract: The cancer, also known as the malignant tumor, is a disease caused by the abnormal cell proliferation. It ranked sixth in the top 10 causes of death published by the World Trade Organization in 2019, and seriously threatens human life and health. The mortality rate is 93% and 19.3 million people developed cancer in 2022 alone. This paper discusses some well-established machine learning-based cancer treatment research methods, including the application of simple models like random forests and some complex models such as neural networks in breast cancer, lung cancer and thyroid cancer. In addition, most of the models in this article are deep learning models because their application scope and value are generally higher than traditional models. For each cancer mentioned, two or three models are presented, along with their basic information and their results. This review can provide some references for the application of machine learning in cancer treatment research.


1 INTRODUCTION

Cancer, also called as malignant tumor, is a disease caused by abnormal cell proliferation, ranked sixth in the top 10 causes of death published by the WHO in 2019, and seriously threatens human life and health. According to IARC data (WHO, 2024), in 2022 alone, 19.3 million people developed cancer, and the mortality rate is as high as 93%. Although the traditional diagnostic methods, such as imaging examination and pathological section analysis are effective, they rely on the doctor's personal experience and subjective judgment, and they have problems of high misdiagnosis rate, low efficiency and high cost. Therefore, it is necessary to introduce some new advanced technologies as auxiliary. Artificial intelligence technology has strong learning, accurate feature extraction ability and feature prediction ability, which can provide doctors with objective and comprehensive diagnostic basis, effectively assist doctors to judge patients' conditions, so as to improve efficiency and reduce costs in cancer research.

In the current cancer-related machine learning (ML), ensemble learning and transfer learning, as the two main core strategies, have significant advantages in improving prediction accuracy and accelerating

research process. Ensemble learning, with its strong generalization ability, has become the key to improving the accuracy of cancer prediction by using integrated algorithms such as random forest and gradient elevator. With the flexibility of cross-domain application, transfer learning can transfer models trained on common cancer types to rare cancer types, reduce the dependence on large amounts of labeled data, significantly shorten the development cycle of new models, and open up a new path for the study of rare cancer types.

For example, Downs et al. used a fuzzy ARTMAP neural network model to extract the features of breast cancer to support network prediction and validation (Downs, 1996). GVani et al. used an Extreme Learning Machine (ELM) to distinguish between benign and malignant breast tumors, which is significantly more efficient than other algorithms (Vani, 2011). HRH et al. used Monte Carlo methods to build an auxiliary diagnostic system for lung cancer with a diagnostic accuracy of 99.15% for benign and 98.70% for malignant tumors (Al-Absi, 2014). Olatunji et al. used different methods, including random forest and neural network, to predict thyroid cancer on the data set of Saudi Arabia, and the accuracy of the random forest model even reached 90.91% (Olatunji, 2021). Li et al. (Li, 2023)

^a <https://orcid.org/0009-0000-6880-3211>

used least absolute shrinkage and selection operator (LASSO) and Support Vector Machine-Recursive Feature Elimination (SVM-RFE), two machine learning algorithms, to identify ATC feature genes and build a prediction model for anaplastic thyroid cancer.

Due to the importance of this field and the fact that AI has proven its superiority for many tasks, it is necessary to give a comprehensive overview of this aspect. The rest of this article is arranged as follows: Part 2 will explain ML and present existing methods on several common cancers. These sections will be further discussed in part 3 and summarized in Part 4.

2 METHOD

There are many types of cancer, and each cancer has its own detection and treatment methods. In order to show the research of machine learning in the field of cancer more succinctly and intuitively, the paper selected three cancers with high incidence and strong representation as the review object.

2.1 Introduction of the Machine Learning

There are two types of ML, “supervised learning” and “unsupervised learning”. The “supervised learning” is to Provide training data consisting of input-output pairs and learn mapping; and the “unsupervised learning” is to find patterns in the data without the help of labels (outputs) (Mitchell, 2003). In order for the machine to accurately identify the desired cancer, the model is usually trained through a few steps shown in Figure 1:

At present, the more common supervised learning models include random forest, decision tree, etc., which are generally suitable for common cancers with many cases. The common unsupervised models are neural networks, clustering algorithms, etc., which are suitable for very rare cancers. Before a model can be confirmed as viable, it needs to be trained on data to ensure that its accuracy is up to the required standard. Not only the supervised learning but also unsupervised learning needs to be trained in these steps.

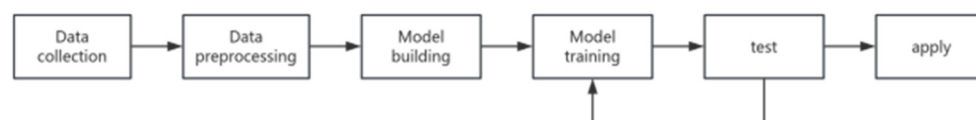


Figure 1. The workflow of the machine learning (Photo/Picture credit: Original).

2.2 Cancer-Based Prediction

2.2.1 Breast Cancer

HOF-ATR-MIR Spectroscopy with Deep Learning in Breast Cancer

Shang et al. (Shang, 2024) recently proposed a three-way classification model based on HOF-ATR-MIR spectroscopy to predict different breast tissues, which is significantly more accurate for breast cancer diagnosis than other common models. They added their self-developed hollow on top of baseline correction (BC) and data-enhanced the one-dimensional convolutional neural network (1D-CNN) model optical fiber attenuated total reflection (HOF-ATR). To obtain a more correct model, they coupled the model with Fourier transform infrared spectroscopy (FTIR), and its accuracy rate is as high as 95.09%.

Deep Neural Network

Vaka et al (Vaka, 2020) comparing various methods in the past, and found that most of them have certain limitations. Therefore, they introduced Deep Neural Network with Support Value (DNNS) to repair the previous images to obtain more accurate images, thus improving the accuracy of breast cancer diagnosis, and its performance, efficiency and image quality have been greatly improved compared with the previous models.

2.2.2 Lung Cancer

Cancer death in men is dominated by lung cancer, which is second only to breast cancer in women. Numerous efforts have been made to develop a realistic and accurate prediction model due to its extremely high mortality rate.

Predictive Model of Radiation Pneumonia

Choi SH et al. (Choi, 2024) successfully developed a predictive model of radiation pneumonia after studying samples from 59 volunteers who had received radiation therapy for primary lung cancer.

They used three models, including logistic regression, to predict different genomes, and three models, including random forests, to evaluate the performance of each functional set. Finally, the importance of single nucleotide polymorphisms (SNPs) for the prediction of radiation pneumonia was confirmed and the feasibility of its model was confirmed.

Shepard Convolutional Neural Network

A new model for accurate segmentation and classification of lung cancer using CT images was proposed by Shetty et al. (Shetty, 2022) after synthesizing a variety of models. The method is divided into several steps. Median filtering and Bayesian fuzzy clustering were used to segment the lung lobes. Then the lung cancer was segmented using a deformation model based on Water Cycle Sea Lion Optimization (WSLNO). They used data enhancement methods to improve the classification accuracy, that is, increasing the size of the segmentation area. Finally, using WSLNO algorithm to train the Shepard Convolutional neural network (ShCNN) to effectively classify lung cancer. The proposed algorithm combines the water cycle algorithm (WCA) and the Sea lion optimization algorithm (SLNO), which improves the accuracy, sensitivity, specificity and average segmentation accuracy compared with the previous models.

Computer-aided Classification via Deep Learning Technique

Hua et al. (Hua, 2015) proposed to simplify the traditional CAD image analysis process by using deep learning technology to distinguish pulmonary nodules on CT images. Two deep learning models, Deep belief Network (DBN) and CNN, were introduced by the authors because lung nodes are an important factor in lung cancer diagnosis. They compared them to two baseline methods that involve feature computation steps. LIDC datasets were utilized to classify the malignancy of pulmonary nodules, preventing the necessity of calculating morphological and textural features. According to the result, the proposed deep learning framework is obviously outperforming conventional hand-crafted feature computing CAD frameworks.

2.2.3 Thyroid Cancer

Worldwide, thyroid cancer is becoming more common as it is one of the most common malignant tumors of the head and neck. Its trigger factors are very common, so it also troubles many high-risk people.

Deep Artificial Neural Network Model for Prediction

Barfejani et al. (Barfejani, 2024) intends to create a deep neural network that can predict mortality in patients with differentiated thyroid cancer. They used data from the SEER database to develop the ThyDAMP (Deep Artificial Neural Network Model for Prediction) to predict mortality in DTC patients. Demographic, histological, and staging information is included in this dataset. After standardization and feature coding, they segment the data into three subsets, including training, testing, and validation subsets, and adjust the model hyperparameters by cross-validation. The predictive potential of ThyDAMP for differentiated thyroid cancer was demonstrated after testing on a separate dataset.

Deep Learning Techniques

Shah et al. (Shah, 2024) used deep learning technologies such as Long Short-Term Memory (LSTM), Gated Recurrent Units (GRUs), and Bi-directional LSTM (Bi-LSTM) to make timely prediction of thyroid cancer. Shah et al. (Shah, 2024) used deep learning technologies such as Long Short-Term Memory (LSTM), Gated Recurrent Units (GRUs), and Bi-directional LSTM (Bi-LSTM) to make timely prediction of thyroid cancer. The model was trained on datasets from asia.ensembl.org and IntOGen.org, information was extracted using a variety of matrix-based methods, including the Hahn moment, and the model was tested using three methods, one of them being SCT. In tests with independent data sets, the model achieved 96 percent accuracy.

3 DISCUSSION

3.1 Contrast Between Traditional ML and Deep Neural Networks

ML models are often used to automate the construction of analytical models, extracting patterns from past learning and databases to help people make reliable and repeatable decisions. This is why these conventional models need to invest a lot of data in training and constantly modify the model to have a certain accuracy. ML has proved to be useful in tasks that require high-dimensional data, such as classification, regression, and clustering (Janiesch, 2021; Shinde, 2018). However, the large number of data sources makes it difficult for conventional models to predict some rare symptoms. Neural

networks, especially deep ones, make up for this shortcoming on the basis of traditional ML. Some conventional ML can also be classified as shallow neural networks. Deep neural networks are able to automatically discover a representation needed for the corresponding learning task because they are organized in deeply nested network architectures. This is why deep neural networks outperform shallow ML algorithms in most situations, such as processing text, images, video, speech, and audio data (LeCun, 2015). However, shallow ML can still yield better results with low data input or limited data samples (Zhang, 2018).

3.2 Limitations and Challenges

Although machine learning and its various models have shown great promise in cancer prediction and research, and have many mature applications, it still has many problems. Because neural network simulates the working mode of human brain, its operation process is in a "black box" state (Shinde, 2018; Shehab, 2022), so doctors cannot know what kind of reasons the model gives diagnosis based on. In this case, it is difficult to assign responsibility if a model diagnosis error leads to a bad result.

Similarly, due to the difference in the training set data, the model will also have applicability problems. There are always subtle differences between different cancers in different regions and different people. Forcing a model that works for one population to another does not guarantee the original accuracy; The development of multiple models corresponds to a huge cost of money and time. All these reasons increase the difficulty of developing and applying predictive models

Access to data sources may also be denied for a number of reasons, making it impossible to develop the most suitable model. Because the leakage of these patient data can lead to serious consequences, such as being used by terrorists to develop targeted viruses, or reverse the rollout of other patient data, researchers need to ensure the security of their data sources and fully consider the privacy of patients.

3.3 Future Prospects

Although the algorithms mentioned above all have some problems such as "black box", there are still some methods that can be used to try to solve the problem. Among them, expert system is a very potential scheme. By adding an expert system to the model, the expert system can sort out the information in the model and put forward explanations like real people, which can alleviate the problem of "black box" to a certain extent (Liao, 2005). SHAP

Explanations are another potential solution. SHAP Explanations are a feature-attribution mechanism that uses game theory concepts to discuss individual features in a model to explain the model as a whole (Broeck, 2022). However, because this concept has not been put forward for too long, the academic community cannot confirm its accuracy and reliability for the time being.

For the adaptability of the model, it is possible to consider trying to solve the problem through transfer learning and domain adaptation security. Transfer learning is especially suitable for situations where data sets are limited or difficult to obtain. It can reduce model training costs by extracting similar features between data sets and optimizing information from the source data set to the new data set (Weiss, 2016). Domain adaptation security is to obtain a new database by labeling different data, matching and comparing the labels one by one. The database contains both old and new data and is able to extract similarities between different libraries (Singhal, 2023).

4 CONCLUSIONS

This paper discusses some well-established machine learning-based cancer treatment research methods, including the application of simple models such as random forests and complex models such as neural networks in breast, lung, and thyroid cancers. For each cancer mentioned, two or three models are given, along with their basic information and results. This review can provide some references for the application of machine learning in cancer treatment research.

But there are still some shortcomings in this paper. Although some traditional machine learning algorithms such as random forest are mentioned in this article, most of the content is introduced to models based on neural networks. Of course, this is not that traditional machine learning algorithms are not good enough, but that neural networks are more widely used. In the follow-up research, the further study can still try the traditional algorithm, or combine the traditional algorithm with the new algorithm to achieve better results.

REFERENCES

- Al-Absi, H. R. H., Belhaouari Samir, B., & Sulaiman, S. 2014. A computer aided diagnosis system for lung cancer based on statistical and machine learning techniques. International Symposium on VLSI Design. IEEE.

- Barfejani, A. H., Rahimi, M., Safdari, H., Gholizadeh, S., Borzooei, S., Roshanaei, G., ... & Tarokhian, A. 2024. Thy-DAMP: Deep artificial neural network model for prediction of thyroid cancer mortality. *European Archives of Oto-Rhino-Laryngology*, 1-7.
- Choi, S. H., Kim, E., Heo, S. J., Seol, M. Y., Chung, Y., & Yoon, H. I. 2024. Integrative prediction model for radiation pneumonitis incorporating genetic and clinical-pathological factors using machine learning. *Clinical and Translational Radiation Oncology*, 48, 100819.
- Downs, J., Harrison, R. F., Kennedy, R. L., & Cross, S. S. 1996. Application of the fuzzy artmap neural network model to medical pattern classification tasks. *Artificial Intelligence in Medicine*, 8(4), 403.
- Hua, K. L., Hsu, C. H., Hidayati, S. C., Cheng, W. H., & Chen, Y. J. 2015. Computer-aided classification of lung nodules on computed tomography images via deep learning technique. *OncoTargets and therapy*, 2015-2022.
- Janiesch, C., Zschech, P., & Heinrich, K. 2021. Machine learning and deep learning. *Electronic Markets*, 31(3), 685-695.
- LeCun, Y., Bengio, Y., & Hinton, G. 2015. Deep learning. *Nature*, 521(7553), 436-444.
- Li, C., Dong, X., Yuan, Q., Xu, G., Di, Z., & Yang, Y., et al. 2023. Identification of novel characteristic biomarkers and immune infiltration profile for the anaplastic thyroid cancer via machine learning algorithms. *Journal of endocrinological investigation*.
- Liao, S. H. 2005. Expert system methodologies and applications—a decade review from 1995 to 2004. *Expert Systems with Applications*, 28(1), 93-103.
- Mitchell, T. M. 2003. *Machine Learning*. McGraw-Hill.
- Olatunji, S. O., Alotaibi, S., Almutairi, E., Alrabac, Z., & Alhiyafi, J. 2021. Early diagnosis of thyroid cancer diseases using computational intelligence techniques: a case study of a Saudi Arabian dataset. *Computers in Biology and Medicine*, 131(4), 104267.
- Shang, H., Wu, Q., Wu, J., Zhou, S., Wang, Z., Wang, H., & Yin, J. 2024. Study on breast cancerization and isolated diagnosis in situ by HOF-ATR-MIR spectroscopy with deep learning. *Spectrochimica Acta Part A: Molecular and Biomolecular Spectroscopy*, 124546.
- Shah, A. A., Daud, A., Bukhari, A., Alshemaimri, B., Ahsan, M., & Younis, R. 2024. DEL-Thyroid: Deep ensemble learning framework for detection of thyroid cancer progression through genomic mutation. *BMC Medical Informatics and Decision Making*, 24(1), 198.
- Shetty, M. V., & Tunga, S. 2022. Optimized deformable model-based segmentation and deep learning for lung cancer classification. *The Journal of Medical Investigation*, 69(3.4), 244-255.
- Shehab, M., Abualigah, L., Shambour, Q., Abu-Hashem, M. A., Shambour, M. K. Y., Alslibi, A. I., & Gandomi, A. H. 2022. Machine learning in medical applications: A review of state-of-the-art methods. *Computers in Biology and Medicine*, 145, 105458.
- Shinde, P. P., & Shah, S. 2018. A review of machine learning and deep learning applications. In *2018 Fourth international conference on computing communication control and automation (ICCUBEA)*, 1-6. IEEE.
- Singhal, P., Walambe, R., Ramanna, S., & Kotecha, K. 2023. Domain adaptation: Challenges, methods, datasets, and applications. *IEEE Access*, 11, 6973-7020.
- Van den Broeck, G., Lykov, A., Schleich, M., & Suci, D. 2022. On the tractability of SHAP explanations. *Journal of Artificial Intelligence Research*, 74, 851-886.
- Vani, G., Savitha, R., & Sundararajan, N. 2011. Classification of abnormalities in digitized mammograms using Extreme Learning Machine. *International Conference on Control Automation Robotics & Vision*. IEEE.
- Vaka, A. R., Soni, B., & Reddy, S. 2020. Breast cancer detection by leveraging Machine Learning. *ICT Express*, 6(4), 320-324.
- Weiss, K., Khoshgoftaar, T. M., & Wang, D. 2016. A survey of transfer learning. *Journal of Big Data*, 3, 1-40.
- WHO, IARC. 2024. Retrieved from <https://www.iarc.who.int/>
- Zhang, Y., & Ling, C. 2018. A strategy to apply machine learning to small datasets in materials science. *NPJ Computational Materials*, 4(1), 25.