# Design a Workflow for the Application of Machine Learning in Diagnosis of Cancer Metastasis

Baichuan Peng[1]

[1]Chengdu Hongwen School, China

## ABSTRACT

Benign tumors can turn into malignant tumors if they are metastatic. Pathological analysis of cancer tissue is the major method for the diagnosis of cancer malignancy and has been widely used for decades. The analysis is based on the specific features of the malignant tissues with metastatic cancer cells. Nevertheless, the diagnosis fully relies on experts' efforts, which is a time-consuming process. Also, the decision based on experts' experiences could be empirical. In this mini review work, we introduced the definition and the diagnosis of cancer malignancy and pointed out the disadvantages of the traditional diagnosis. To improve the efficiency and accuracy of the diagnosis, we proposed a novel workflow for the diagnosis of cancer malignancy. In this workflow, we integrated the in vitro primary cancer cell culture and the machine learning algorithms. After training with big data that consists of images with known features of malignancy status, the machine learning algorithms can recognize the cancer malignancy to perform the diagnosis for cancer patients.

## Introduction

### Cancer metastasis and the current diagnostic methods

Metastasis, the dissemination of cancer cells from the primary tumor to a distant organ, is the leading cause of death for cancer patients. [1] In the process of metastasis, the cancer cells spread to other parts of the body. In other words, a carcinoid, or benign tumor, can turn into a malignant tumor if it is metastatic. The metastatic status of cancers is divided into three categories T, N, and M. T category describes the extent of the primary tumor, considering either size, depth of invasion or invasion of adjacent tissues; the N category indicates the extent of regional lymph nodes metastasis; and the M category indicates the presence of metastasis to distant organs. [2] Currently, pathological analysis of cancer tissue is the major method for the diagnosis of cancer cell metastasis and has been widely used for decades. The analysis is based on the specific features of the malignant tissues with metastatic cancer cells. Briefly, pathological diagnosis includes the following steps. First, acquire biopsies, i.e., cancer tissue samples removed from patients; Second, process the cancer tissue to do histological staining with antibodies of specific metastatic markers; Finally, the malignancy of the tumor is determined through pathological analysis. Pathologists determine whether the tissue is metastatic based on the features of the metastatic cancer cells. [3]

Pathologists make decisions based on two major features. The first is the expression level of metastasis - related genes [4]. There are specific genes that are identified as biomarkers for the diagnosis of the malignancy of tumors. Compared to benign tumors or healthy tissue, these genes could be overexpressed, or suppressed, or could be mutated to express in different formats. Certain antibodies are designed to detect the expression or the mutation of these genes to provide references for Pathologists. The second feature is the different morphology between the cancer tissue and the normal tissue. For example, the malignant metastatic cancer cells are polarized, characterized with amoeboid morphology and invade into surrounding tissues [5]. In this sense, the polarization and spreading of tumor cells shown in the histological slides indicate that the tumor is malignant. The structure and directions of extracellular matrix fibers are also different between malignant tumor and benign tumor. The extracellular matrix fibers usually align into a specific direction in the malignant case, instead of the homogenous distribution in normal or benign cases.

There are several disadvantages in traditional pathological diagnosis. The diagnosis fully relies on experts' efforts that could take a long time to finish the diagnosis process and requires experts' experiences in making decisions. Different criteria are considered during analyzing the tumor tissue, so the experts need to repeatedly observe the tumor slide in order to get an informative report and accurate diagnosis. Therefore, the diagnosis would be a time-consuming process. Also, the diagnosis could be empirical because the results rely on experts' experiences. The accuracy of the diagnosis from different pathologists might vary. New technologies are developed to solve the problems in traditional cancer diagnosis. Machine learning (ML) technology can provide a standard process to make predictions or decisions effectively and accurately. Recently, machine learning as an emerging technology is employed in cancer diagnosis.

### The algorithms of machine learning (ML) and applications in cancer diagnosis

As a part of artificial intelligence, machine learning (ML) algorithms that are trained with existing big data with known features can make predictions or decisions based on new data without any known features. The recent achievements of fundamental ML theory provide a solid basis for improving the accuracy of the predictions or decisions made by ML algorithms. The gradient descent iterative algorithm is the most important and indispensable theory for ML. In a gradient descent algorithm there will be a function that alters its parameter after each iteration. The parameters are updated from a cost function and a hypothesis function. The hypothesis function provides a model to distinguish sets of data with different features. Then, the equation of the hypothesis function with all possibilities will be calculated to find the most suitable regression equation that is coherent. To find the most cogent line of the regression equation, a cost function is used to calculate the deviation of the hypothesis function. This is to say that the cost function with the minimum cost gives the parameter of the hypothesis function that expresses the line of best fit. The gradient descent technique as a general foundation is leveraged to develop ML models with higher accuracy. [11]

ML algorithms have been used in a variety of applications including the field of cancer diagnosis. [6] For example, ML is applied to improve the genomic characterization of tumors, accelerating the discovery and therapeutic efficacy of cancer drugs. ML models were also designed to provide suggestions for the therapeutic schedule of patients. Moreover, it was proved to achieve accuracy of over 98% in deciding an adequate schedule [7]. Another example is the ML algorithms trained with the MRI, CT or X-rays images. It has shown that the application of ML algorithms in the image data of MRI, CT or X-rays improved cancer detection and quickly identified the degree of cancer malignancy [9][10].

## Design a workflow for diagnosis of cancer metastasis via integrating primary tumor cell culture and ML Technology

Here, we designed a new workflow of integrating primary tumor cell culture and ML technology for the diagnosis of cancer metastasis (Fig. 1). The ML programs are trained with big data that consist of images with known features of malignancy status (Fig. 1). [13] In the traditional diagnosis, tumor resection samples obtained from patients are processed for pathological diagnosis. In our workflow, we propose to expand and culture the primary cancer cells collected from cancer patients. By culturing and characterizing the primary tumor cells, high-throughput imaging microscopes could efficiently take images of tumor cells. Then, we analyze the features of the in-vitro cells, instead of the tissue resection samples. [12] There are no limits to expanding the patents' primary cancer cell, providing resources for diagnosis. In addition, an advantage of our workflow is that the primary tumor cells could be leveraged for personalized drug screening for the patients' specific type of tumor. We describe the details of each step involved in our workflow (Fig. 1) in the following context.

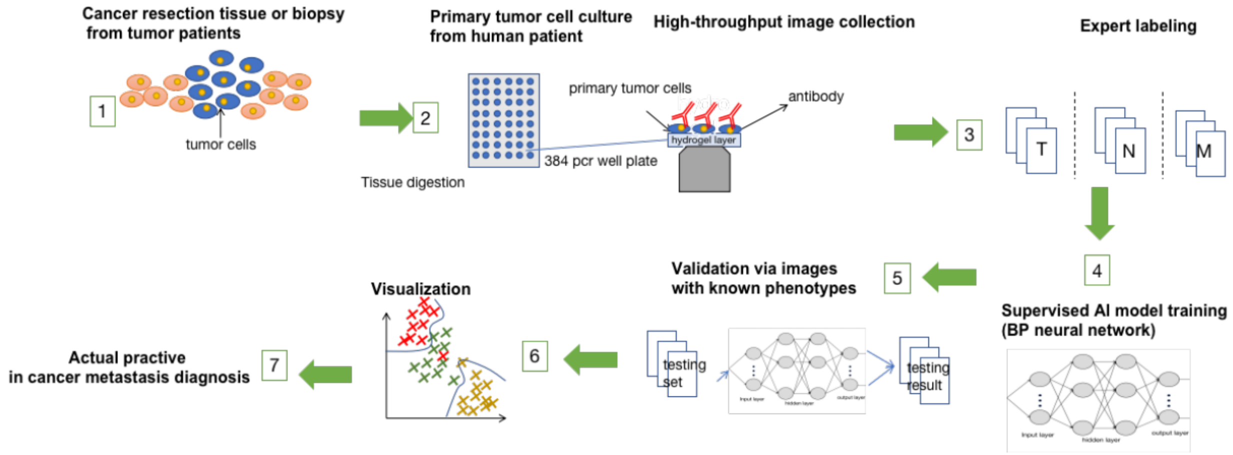
Cancer resection tissue or biopsy from tumor patients
1
tumor cells
2
Tissue digestion
Primary tumor cell culture from human patient
384 pcr well plate
High-throughput image collection
primary tumor cells
antibody
hydrogel layer
3
Expert labeling
T
N
M
4
Supervised AI model training (BP neural network)
input layer
hidden layer
output layer
5
Validation via images with known phenotypes
testing set
testing result
6
Visualization
7
Actual practive in cancer metastasis diagnosis

**Figure 1. Application of ML technology in the diagnosis of cancer metastasis.** The workflow consists of seven steps. (1) Biopsy or resection samples of tumor tissue and adjacent tissue are obtained from the patients. (2) The samples are digested to isolate primary cancer cells that are cultured on the multi-well plates through the standard primary cell culture process. After the cell culture process, cells are stained using antibodies for immunofluorescence staining. The images of primary tumor cells are obtained via high-throughput imaging microscope. (3) The images of the primary tumor cells are labeled with their metastatic status by experts. (4) The images with the known features are divided into a test set and training set for training the neural network. (5) The trained neural network is tested to verify the accuracy in recognizing the tumor metastatic status. (6) The predictions of the ML model could be visualized using non-linear regression in Cartesian coordinates. (7) After the neural network is fully tested and validated, it can bring several positive impacts including predicting metastatic status and accelerating the discovery of new drugs.

## In-vitro Culture and imaging of primary cancer cells

The primary cancer cells could be isolated from the resection tumor tissues from cancer patients. Usually, the tumor tissues are digested via enzymatic methods in sterile environments, and the cancer cells are cultured in tissue culture plates [15, 16]. These primary cells grown in in vitro conditions replicate the genetic signatures of individual patients. The cell migration assays and the quantification of metastasis-related gene expression could be used to characterize these tumor cells. The tumor cells exhibit different features based on their type. In other words, tumors can be categorized according to the genes that mutate in the patients.

With respect to diagnosing cancer metastasis, cells are stained with specific biomarkers of metastasis. [17] In this way, the gene expression of mutated cells would become visible under fluorescent microscope. Immunofluorescence (IF) is an important immunochemical technique that can paint gene expression levels. [18] immunofluorescence staining normally includes the following steps. First, cells are usually fixed and cell membrane is permeabilized. Second, cells are stained with primary antibodies to bind specific proteins. Third, a secondary antibody with fluorophore is used to bind the primary antibody fixation. At the end, the fluorophore is imaged using a fluorescent microscope. [18] To improve the efficiency, a high-throughput imaging process is designed to collect a large number of images.

## Label the image data of primary cancer cells with metastatic status

ML algorithms are trained with big data with known features to recognize the features of new data. To establish big data for training ML, the images of the primary tumor cells are further processed through expert labeling. The images of benign tumors are marked and categorized to a data set, and other images of malignant tumors are categorized to different sets. Therefore, each data set is marked with either malignant or benign. Expert labeling is usually carried out by experts who specialize in examining tumor malignancy. The larger the quantity, the trained ML model would be able to analyze the malignancy of the tumor more accurately. Moreover, in ML programs, those images with known features are processed and modified to make sure the recognition of a wider range of input data. In other words, the program may modify the image mathematically using a statistical method to promote the accuracy of recognition. In addition, a small portion of both malignant tumor set, and benign tumor set are extracted from the original set and stored as a test set. The extracted sets are prepared for the verification of the accuracy of the ML model in the later stage.

## Back Propagation Neural network (BPNN) of ML model

After training the ML model with the expert-labeled images, the ML model is supposed to recognize the tumor malignancy based on the tumor cell features including cell shape, size, textures, and pixel intensity, which is a complex process. Back Propagation neural networks (BPNN) are suggested to perform this complex analysis.

The BPNN or ML model that has many different interconnected layers is typically known as a broad family of Artificial Neural networks (Fig. 2). In BP NN, the Deepest-Descent technique [20] (gradient descent) is used as introduced above. This algorithm is implemented in each hidden unit that is responsible for minimizing the errors of non-linear functions. [20] The standard BP is composed of input units, hidden units and output units (Fig. 2). The input unit's sensors scan images of primary tumor cells, and hidden units are the internal units that generate an output according to the input data using minimum cost methods. The output units make predictions and decisions on the phenotypes of the tumor cells. The network system establishes relationships between these units. The connections are determined by a weight function, indicating the strength of the connection. [21] This connection is first initialized with a unique and small value of weight [22]. The BPNN model can adjust itself if a set of training data is given, and this weight would adjust constantly and update through gradient technique [23]. The BP ANN program will choose a weight distribution that gives the most adequate prediction that approaches the target value.

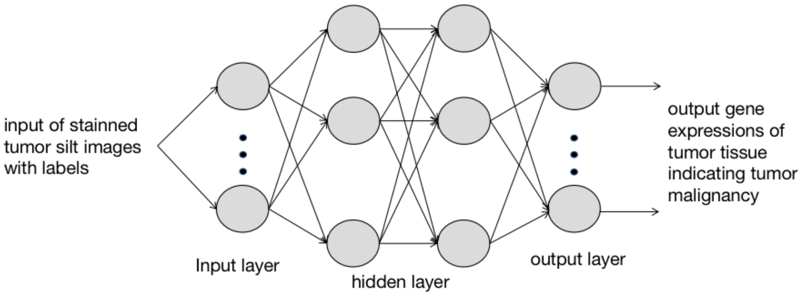

**Figure 2. The internal structure of a back propagation neural network (BPNN).** The network has three layers which are input, output and hidden layers. The input layer of a neural network has several elements. Each element in the input layer receives a section of the tumor cell image in the form of a pixel vector component. The hidden layers receive inputs from input layer neurons and process the data by adjusting the weight and bias. And the output layer will be responsible for the final output, which is the final level of tumor malignancy.

### Validation of accuracy of ML predictions and visualization

At this stage, the machine learning program with trained BPNN can determine the malignancy of the tumor with reliable accuracy. In case the predictions and determinations are not accurate, we need to verify the accuracy using the image set with known phenotypes prepared in the expert labeling step as described above. Learning rate during training in the ML model is critical for accuracy. If the learning rate is too small, the program would take a very long time to complete the training period. Also, the number of iterations would have a great influence on the result. Theoretically, the more the number of iterations, the more precise the prediction. In fact, however, an adequate number of iterations is important. Much more or much less iterations would reduce the accuracy of the prediction. Ideally, the training model must be adjusted through a long period of time till it is able to recognize the malignant tumor with an accuracy above 95 percent.

In the final stage, the predictions of the ML program for the cancer malignancy are visualized and scored using non-linear regression in Cartesian coordinates. [24] By using this method, the scores are constructed according to features of the primary cancer cell images. For example, we could assign the position (0, 1) for the feature of health cells, the position (1, 0) for the feature of benign tumor cells, and the position (1, 1) for the feature of malignant cells. The features of the new data could be analyzed and fitted to find the corresponding position in this two-dimensional space. The predictions are scored based on the distance of this corresponding position to the other known positions representing different phenotypes. For example, if it is close to the position (1, 1), this means the cells tend to be malignant. Note, we also could screen cancer drugs using this workflow. The drugs can be administered to the in-vitro tumor cell cultures. After performing the same process, we could filter out the drugs that can reverse the malignant feature towards a benign tumor.

## Discussion

Traditional diagnosis of cancer malignancy is based on pathological analysis of cancer tissue, which has been widely used for decades. Pathological analysis requires experts with experiences in making decisions based on the observation of tumor tissue. Meanwhile, different aspects are needed to be considered and balanced when analyzing the tumor tissue. Thus, the diagnosis would be a time-consuming process, and the decision could be empirical.

To overcome the problems in the traditional diagnosis for cancer malignancy, we proposed a new workflow that integrates primary tumor cell culture and BPNN of the ML model. Briefly, the cells collected from patients' tumor tissues are cultured and imaged to establish big data. The big data is labeled via experts to identify the healthy state, benign state, and the malignant state. The BPNN is trained with big data to be able to recognize and quantify the features of these three states via quantified scores. [2] After the training process, the accuracy of BPNN identifying the healthy, benign and malignant states is verified via the images with known state. Usually, the accuracy is expected to reach at least 90%. With the establishment of the ML workflow, we can efficiently and accurately diagnose cancer malignancy for cancer patients. Moreover, this automated diagnosis process avoids the needs of experts in identifying the phenotypes, eliminating the empirical decisions made by experts.

We proposed to leverage the images of primary tumor cells instead of the pathological tissue slides for cancer diagnosis. The tissue slides can only provide local information with several micron thickness, and the phenotypes of a tumor tissue are usually complicated and not easy to quantify. Compared to the tissue slides, the primary tumor cells can provide more stable, more concise and more robust phenotypes that help promote the accuracy and efficiency of the diagnosis process. In addition, the primary tumor cells in culture can respond to the administration of cancer drugs. Since the primary tumor cells reserved the gene features of the cancer patients, the drug response of the primary tumor cells might be used to predict the response of the drug treatment in patients. Therefore, we can collect the images of primary tumor cells with the administration of cancer drugs and perform feature analysis using our ML workflow to predict the drug efficacy in treating the cancer patients. In summary, we proposed a novel workflow integrating BPNN of ML model and in vitro primary tumor cell culture to perform automated, efficient, and accurate diagnosis of malignancy for cancer patients.

## Acknowledgements

Primarily, I would like to thank Dr. Xie for guiding me during the research project. I would like to express my special thanks to Dr. Xie was able to guide me in the design of the framework and work flow of this paper despite the presence of difficulties in communications due to the 12-hour time zone between mainland China and the U.S. EST. Then, I would like to thank my mentor Ms. Anlynn Liu, for providing me a valuable opportunity to attend this scientific research project and work on this paper during my junior high school year and the kind assistance throughout the project.

Finally, I would like to thank my parents and friends who have always helped me with their valuable suggestions and guidance, supported and encouraged me in various stages of project completion.

## References

1. Matz, M., et al., *The histology of ovarian cancer: worldwide distribution and implications for international survival comparisons (CONCORD-2).* Gynecol Oncol, 2017. **144**(2): p. 405-413.
2. Brierley, J., M. Gospodarowicz, and B. O'Sullivan, *The principles of cancer staging.* Ecancermedicalscience, 2016. **10**: p. ed61.
3. Kim, H., et al., *Rapid histologic diagnosis using quick fluorescence staining and tissue confocal microscopy.* Microsc Res Tech, 2019. **82**(6): p. 892-897.
4. Peng, X.-H., et al., *Real-time Detection of Gene Expression in Cancer Cells Using Molecular Beacon Imaging: New Strategies for Cancer Research.* Cancer Research, 2005. **65**(5): p. 1909-1917.
5. Shackleton, M., *Normal stem cells and cancer stem cells: similar and different.* Seminars in Cancer Biology, 2010. **20**(2): p. 85-92.
6. Gupta, S., A. Gupta, and Y. Kumar, *Artificial intelligence techniques in Cancer research: Opportunities and challenges*. 2021. 411-416.
7. 涛, 徐.姜.段.华.孙., *Watson for Oncology 在乳腺癌治疗中的应用与思考.* 中国研究型医院, 2018. **5**(3): p. 19-24.
8. Siegel, R.L., et al., *Cancer Statistics, 2021.* CA: A Cancer Journal for Clinicians, 2021. **71**(1): p. 7-33.
9. DeGrave, A.J., J.D. Janizek, and S.I. Lee, *AI for radiographic COVID-19 detection selects shortcuts over signal.* medRxiv, 2020.
10. Oren, O., B.J. Gersh, and D.L. Bhatt, *Artificial intelligence in medical imaging: switching from radiographic pathological data to clinically meaningful endpoints.* Lancet Digit Health, 2020. **2**(9): p. e486-e488.
11. Schneider, A., G. Hommel, and M. Blettner, *Linear regression analysis: part 14 of a series on evaluation of scientific publications.* Deutsches Arzteblatt international, 2010. **107 44**: p. 776-82.
12. Zhou, Y., *Understanding the cancer/tumor biology from 2D to 3D.* J Thorac Dis, 2016. **8**(11): p. E1484-e1486.
13. Dlamini, Z., et al., *Artificial intelligence (AI) and big data in cancer and precision oncology.* Computational and Structural Biotechnology Journal, 2020. **18**: p. 2300 - 2311.
14. Romero-Garcia, S., et al., *Tumor cell metabolism: an integral view.* Cancer Biol Ther, 2011. **12**(11): p. 939-48.
15. Thakor, J., et al., *Engineered hydrogels for brain tumor culture and therapy.* Bio-Design and Manufacturing, 2020. **3**(3): p. 203-226.

16. Ding, Z.Z., et al., *Simulation of ECM with silk and chitosan nanocomposite materials.* Journal of Materials Chemistry B, 2017. **5**(24): p. 4789-4796.
17. Rindi, G., et al., *ECL cell tumor and poorly differentiated endocrine carcinoma of the stomach: Prognostic evaluation by pathological analysis.* Gastroenterology, 1999. **116**(3): p. 532-542.
18. Im, K., et al., *An Introduction to Performing Immunofluorescence Staining*, in *Biobanking: Methods and Protocols*, W.H. Yong, Editor. 2019, Springer New York: New York, NY. p. 299-311.
19. Levi, I., et al., *Characterization of tumor infiltrating natural killer cell subset.* Oncotarget, 2015. **6**(15): p. 13835-43.
20. Buscema, M., *Back Propagation Neural Networks.* Substance Use & Misuse, 1998. **33**(2): p. 233-270.
21. Cao, W., et al., *A review on neural networks with random weights.* Neurocomputing, 2018. **275**: p. 278-287.
22. De Wilde, P., *Backpropagation*, in *Neural Network Models: Theory and Projects*, P. De Wilde, Editor. 1997, Springer London: London. p. 33-52.
23. S, A. and Y. Zhang, *A Review on Back-Propagation Neural Networks in the Application of Remote Sensing Image Classification.* Journal of Earth Science and Engineering, 2015. **5**.
24. Zhao, J., *Visualization of BP neural network using parallel coordinates.* 2010.
25. Bi, W.L., et al., *Artificial intelligence in cancer imaging: Clinical challenges and applications.* CA Cancer J Clin, 2019. **69**(2): p. 127-157.
26. Benque, D., et al., *Bio Model Analyzer: Visual Tool for Modeling and Analysis of Biological Networks.* 2012. p. 686-692.