GEORGE BROWN PRESS
JOURNALS, BOOKS & PROCEEDINGS

*International Journal of Engineering Advances*

*Review*

# Research on the Application of Integrating Medical Data Intelligence and Machine Learning Algorithms in Cancer Diagnosis

**Xiangtian Hui** [1,*]

[1] School of Professional Studies, New York University, New York, NY, 10012, USA
* Correspondence: Xiangtian Hui, School of Professional Studies, New York University, New York, NY, 10012, USA

**Abstract:** The exponential growth of medical data has made it increasingly important to unlock its latent clinical value. This paper investigates the integration of medical data intelligence and machine learning algorithms in the context of cancer diagnosis. Key challenges and strategies related to multi-source data fusion, feature extraction, model optimization, and system stability are examined. The study reviews the use of imaging, pathology, genomic, and textual data across various machine learning frameworks and proposes a streamlined, clinically applicable approach to intelligent diagnostic support. The aim is to provide an efficient auxiliary diagnostic tool to advance precision medicine and support clinical decision-making in oncology.

**Keywords:** medical data intelligence; cancer diagnosis; machine learning; cancer diagnosis; multimodal fusion; model optimization; clinical decision support; deep learning

Received: 25 October 2025
Revised: 03 November 2025
Accepted: 02 December 2025
Published: 06 December 2025

## 1. Introduction

Cancer remains one of the leading causes of mortality worldwide. Early detection and accurate diagnosis are critical to improving patient survival rates and treatment outcomes. However, traditional diagnostic approaches often rely heavily on physicians' subjective experience and manual interpretation, which can be time-consuming, inconsistent, and prone to human error. With the rapid advancement of medical big data and artificial intelligence, the landscape of cancer diagnosis is shifting toward greater automation and precision. Machine learning-driven analysis of large-scale, heterogeneous medical data offers new opportunities for early cancer detection, risk stratification, and personalized treatment planning. By integrating multi-modal clinical data with intelligent algorithms, it is now possible to build scalable, interpretable, and high-performance diagnostic systems that support evidence-based decision-making in oncology.

## 2. The Concept of Medical Data Intelligence

### *2.1. Types and Characteristics of Medical Data*

Medical data is generated across a wide range of clinical, preventive, and research contexts. It includes information produced during diagnostic and therapeutic procedures, routine health screenings, follow-up care, and biomedical research. Broadly, medical data can be categorized into two main types: structured and unstructured data. Structured data refers to well-organized, standardized information that is easily stored, queried, and analyzed using traditional databases and statistical models. Examples include laboratory test results, medication records, demographic data, and vital signs [1]. In contrast, unstructured data encompasses complex, non-tabular formats such as medical imaging (e.g., CT, MRI), pathological slides, free-text clinical notes, audio recordings, and more.

These data types are often rich in clinical detail but require advanced processing techniques for effective interpretation. Recent advancements in high-throughput biomedical technologies have further expanded the landscape of medical data. In addition to genomics and transcriptomics, emerging omics layers-such as proteomics, metabolomics, and phenomics-are increasingly integrated into clinical research. These data are high-dimensional, heterogeneous, and temporally dynamic, posing significant challenges for integration and analysis. However, they also present new opportunities for intelligent modeling and precision diagnosis using artificial intelligence (AI) technologies.

### *2.2. The Core Connotation and Development of Medical Intelligence*

Medical intelligence refers to the application of advanced information technologies-including AI, big data analytics, and cloud computing-to enhance medical services and decision-making. It focuses on accurate data acquisition, intelligent feedback mechanisms, and predictive modeling to improve the quality, precision, and personalization of healthcare delivery. Core components of medical intelligence include automated pattern recognition, risk stratification, clinical decision support, recommendation of personalized treatment plans. Early systems were primarily rule-based expert systems, relying on predefined logic and heuristics [2]. Over time, the field has evolved into more flexible paradigms grounded in statistical modeling, machine learning, and deep learning. This evolution reflects a shift from knowledge-driven to data-driven approaches, enabling systems to learn complex patterns directly from large-scale datasets rather than relying solely on human-crafted rules. This convergence of data abundance and algorithmic advancement underpins the modern framework of medical intelligence, enabling scalable and adaptive solutions to support evidence-based clinical care [3].

## 3. The Application of Machine Learning Algorithms in Cancer Diagnosis

### *3.1. Cancer Image Recognition*

Medical images play a crucial role in the screening and confirmation of cancer. Deep learning neural networks, such as CNN, can automatically identify the characteristics of diseases within images and are widely used in the analysis of images related to breast, lung and brain cancers. For breast X-ray images, they can identify the location and boundaries of calcification and achieve hierarchical detection [4]. Analyzing lung CT scan images with a 3D convolution model can effectively identify the benign and malignant nature of tumors and improve the accuracy of early screening.

The convolutional layer, as the core structure of CNN, its feature extraction process can be expressed by the following formula:

$$O_{\mathrm{i,j}} = \sum_{m=0}^{M-1} \sum_{n=0}^{N-1} I_{i+\mathrm{m,j}+n} \cdot K_{\mathrm{m,n}} \quad (1)$$

Here, $I$ represents the input image matrix, $\mathrm{K}$ is the convolution kernel, $O^{i,j}$ is the convolution result at the corresponding position of the output feature map, and $M$ and $N$ height and width of the kernel. Feature extraction is accomplished through the sliding window method, and the model can capture the changes of local texture, edge shape and spatial structure.

The deep mode can automatically extract features of various shape and texture changes through multi-level feature extraction without the need for manual feature design. It also features focus strategies and multi-scale networks, enhancing the sensitivity to tumors of different sizes and in different locations. This method has been trained through image fusion optimization techniques and transfer learning, etc., to enhance its robustness when working in a small number of samples and a high-noise environment. The AI medical imaging recognition system can be applied to individual examinations in some hospitals, achieving the functions of automatic film reading and marking for initial examination, reducing the workload of physicians and improving work efficiency [5].

### 3.2. Pathological Image Analysis

The panoramic image acquisition ability of Digital secting technology (WSI) makes pathological images an important reference basis for pathological diagnosis. The accuracy of cancer cell region recognition, assessment of pathological changes in tissue structure, tumor grading and tumor boundary recognition based on deep learning models Convolutional neural networks search for the shape of tissue cells, nuclear structure and tissue density on high-definition images, thereby facilitating more accurate classification and local analysis. In the case of HER2 staining scores for breast cancer, this model can accurately identify the staining intensity and classify it, significantly reducing the time for manual assessment. This mode processes the image input by using block segmentation and sliding window methods, and adds an attention mechanism at the same time to ensure the correctness of all the contents in the slice (see Figure 1).

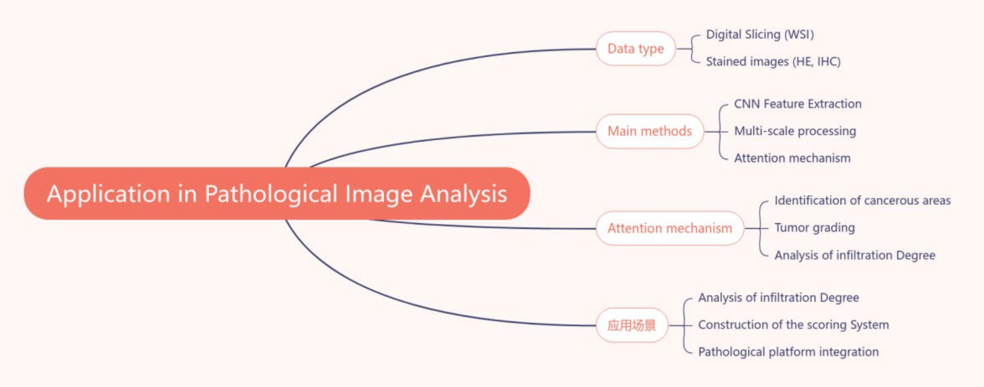
Application in Pathological Image Analysis
Data type
Digital Slicing (WSI)
Stained images (HE, IHC)
Main methods
CNN Feature Extraction
Multi-scale processing
Attention mechanism
Attention mechanism
Identification of cancerous areas
Tumor grading
Analysis of infiltration Degree
应用场景
Analysis of infiltration Degree
Construction of the scoring System
Pathological platform integration

**Figure 1.** Overview of the elements for intelligent analysis of pathological images.

The construction of a model requires factors such as hardware computing power and the reasoning speed of the model. Lightweight models or cloud-based reasoning modes are often chosen.

### 3.3. Gene Data Prediction

The occurrence and development of cancer are closely related to gene mutations and expression disorders in tumors. Accurate prediction and classification of tumors can be carried out by using tumor models based on gene sequences, transcription and methylation levels. Due to the complexity and huge quantity of genetic information, it is extremely important to adopt models with strong feature selection and display capabilities to handle genetic information.

In the application of machine learning methods, classic methods such as vector machines, random forests, and logistic regression, as well as the relatively popular ones in recent years like deep neural networks, autoencoders, and graph neural networks, have been widely applied. In RNA expression profile analysis, the Autoencoder is used for nonlinear dimensionality reduction, and its basic structure can be expressed as:

$$z = f_{enc}(x)\ \hat{x} = f_{dec}(z) \tag{2}$$

Among them, is the original high-dimensional gene expression vector of $x$ , $z$ is the latent feature after compression, and $f_{enc}$ and $f_{dec}$ are the encoder and decoder functions respectively. This method can remove redundant information and extract expression patterns with biological significance.

### 3.4. Analysis of Medical Record Texts

Electronic medical records, including disease course, physical examination, tests, examinations, image descriptions, medication usage, and follow-up evaluation records, etc., are the core information source for the diagnosis and treatment of tumors. These unstructured texts have characteristics such as diverse words and variable structures. Traditional statistical methods or fixed patterns cannot describe their meanings in depth. With the aid of natural language processing, machine learning methods can be adopted to describe entity extraction, event extraction, causal inference and time construction in medical record data. Especially in deep learning language models (such as BERT, BioBERT, ClinicalBERT), they have demonstrated excellent semantic understanding capabilities in medical Settings and can be applied to the prediction of tumor types, treatment plans, and recurrence situations, etc. In the follow-up management of tumors, symptoms, subjective emotions and therapeutic effects are obtained through texts, and the risk of recurrence is monitored and adjusted in real time.

## 4. The Integration of Medical Data Intelligence and Machine Learning Algorithms

### *4.1. Integrate Multi-Source Data*

To further improve the accuracy of cancer diagnosis, it is necessary to conduct a joint analysis of various data types (such as imaging, histology, genomics and electronic health records, etc.). Each data type provides relevant information in its own unique form. The fusion methods can be divided into three parts: integration at the data level, integration at the feature level, and integration at the decision level (see Table 1).

**Table 1.** Comparison of Different Data Fusion Methods.

| Fusion mode | Description | Advantages | Limitations | Applicable scenarios |
|---|---|---|---|---|
| Data layer fusion | Unify the data format before input and concatenate it into a single input | The implementation is simple and the original information is retained | It requires a high degree of data synchronization and is vulnerable to missing values | Structured data collected synchronously |
| Feature layer fusion | The features of each mode are independently extracted and then spliced or weighted fused | It is highly flexible and adaptable to heterogeneous data | The differences in the feature space need to be unified, and the fusion method has a significant impact | Multimodal tasks such as image + text and image + gene |
| Decision-making level integration | Each mode is independently modeled and the prediction results are fused | The model has strong independence and is suitable for situations with missing modalities | The model collaboration mechanism is complex and the computational overhead is relatively high | Multi-center data or incomplete modal scenarios |

Integrating the data layer is the most direct approach. Before entering the network, data from different sources are transformed and spliced in the same way through corresponding algorithms. It has high requirements for the consistency of the time points and dimensions of various types of data and is often used for synchronously collected data, such as image and feature data of corresponding markers. More commonly, it is the fusion of feature layers. Corresponding models are used to extract features for each type of data respectively. For example, CNN is used to extract image features, BiLSTM is used to process case records of time series, Transformer is used to analyze text, and the feature vectors of the intermediate layers are concatenated or fused through the attention layer.

Input them together into the diagnostic network. Its advantage is to maintain the characteristics of each modality itself while being able to capture information across modalities. The integration of the decision-making level adopts the method of model parallelism. Corresponding models are developed for different modalities respectively. When combining the prediction results of each modality, methods such as weighted average, voting or superimposed integration are used. This method is very suitable for situations where the collected data is inconsistent or certain modalities are missing. For practical operation, it is necessary to carry out standardized preprocessing work for data of different modalities to make them comparable.

### *4.2. Extract Key Features*

Efficient feature extraction is the key guarantee for the high efficiency of intelligent diagnostic models. In order to obtain effective features, efficient feature selection and representation learning techniques are adopted to obtain the features with the greatest ability to identify cancer and reduce the accumulation of irrelevant noise information. The main features can be selected from the structured information (test report or population information) through the filtering method (information gain or chi-square test); Or the most important variable selection can be adaptively made through the wrapping method (recursive feature elimination) or the insertion method (L1 regularization). For unstructured data such as image data and text, the semantic feature information can be automatically learned through the depth of learning. For image data, high-level features such as local texture, boundary, and tumor morphology of the region can be mined by constructing a multi-level convolutional neural network system. For pathological images, multi-scale convolution and focus strategy can be introduced to locate the precise lesion areas. For genetic data, the dimensionality reduction methods of PCA and autoencoder can be utilized to reduce the data of thousands of gene expression levels into a small number of latent elements, so as to avoid the phenomenon of dimensionality disaster.

The problem of feature space differences existing in multimodal information can be attempted to complete the transformation from multimodal features to a unified vector space by using methods such as shared expression learning or cross-modal attention interaction. These methods can also be used to dynamically identify important features. It is also possible to use some methods such as SHAP and LIME to evaluate the importance of features during training to understand and interpret how the model relies on certain variables, thereby exploring some new features that may have medical value.

Feature extraction is not only for improving the model performance, but also for the model's interpretability and clinical applicability (see Table 2).

**Table 2.** Common Feature Extraction Methods for Medical Data.

| Data type | Feature extraction method | Technical Description | Advantage |
|---|---|---|---|
| Structured data | Information gain, L1 regularization, tree model feature scoring | Evaluate the importance of variables through statistical or model mechanisms | Easy to implement and with high computational efficiency |
| Medical imaging | CNN, multi-scale convolution, attention mechanism | Automatically extract features such as spatial texture, edge and shape | No need for manual design of features and adaptation to complex structures |
| Pathological image | Sliding window, local feature enhancement | Extract the cancerous areas in high- | Precisely identify tiny lesions |

| | | | |
|---|---|---|---|
| | | resolution images in blocks | |
| Genetic data | PCA, autoencoder, variational autoencoder (VAE) | Dimensionality reduction is used to extract latent variables and eliminate redundancy and noise | Suitable for high-dimensional sparse data and improving model performance |
| Text/Medical record | BERT, BiLSTM, Word Embedding (Word2Vec) | Express semantics and word order, and extract medical entities and relationships | Capture the context and enhance comprehension ability |

Through the combination of reasonable feature engineering and end-to-end learning mechanisms, the predictive ability and interpretability of the model can be effectively balanced, providing a reliable data basis for intelligent diagnosis.

### *4.3. Optimize Model Combination*

The combined model can achieve better predictive performance and system robustness in cancer diagnosis. A single model often has difficulty coping with the calculation of high-dimensional, multimodal, and complex multimodal patient medical data. Constructing a collaborative model is a further improvement in the optimization direction of the combined strategy.

The understanding of the model set can be approached from three aspects: integrated modeling, architecture transformation, and parameter debugging. Integrated modeling builds a group decision-making model by aggregating several basic models, thereby enhancing the overall efficiency. Bagging is a method that creates a large number of independent models based on resampling technology, and then summarizes the combination of prediction results through the mean method or voting to enhance the stability of the system. Boosting is a process of training and error correction layer by layer, which can gradually improve the model and is usually used in the work of accuracy. Stacking is a method that integrates various models with different architectures into a single entity and adds a meta-learner to process the combination of these output results, further enhancing the accuracy of predictions.

In structural optimization, different subnet structures are selected based on the data type. For image data, convolutional networks should be adopted to extract spatial features; for text data, Transformer models should be used to extract semantic features; and for time series data, recurrent networks can be employed to construct temporal correlations. For multimodal problems, multiple branches can be used to process input information of different forms, and then the fusion of information and the decision-making of the same output can be achieved through the fusion layer. The use of residual links and standardized layers can increase the training speed and reduce the risk of vanishing gradients. At the same time, the function of processing multiple tasks simultaneously can also be achieved through multi-task learning, such as cancer detection, staging, and future prediction. The adaptability of the model to the target task is changed by altering the structural parameters of the model and training hyperparameters. Automated parameter adjustments such as grid search and Bayesian optimization can search the hyperparameter space within the system scope to explore the best configuration. If the training resources are limited, the pre-trained model can be used for transfer learning, which can save training time and achieve better performance with fewer training samples.

### *4.4. Strengthen Explanation and Security*

Interpretability and the privacy and security of data are the problems that must be solved before the intelligent cancer diagnosis system enters the clinical field. Physicians and patients need to be able to explain the judgment criteria of the model, and regulatory agencies need to verify that the operation process of the model can be tracked and explained. Considering data privacy and security, patients' privacy is not leaked and the model operation is not attacked.

Interpretability can be characterized both globally and locally. Global interpretation can utilize the importance of features (such as SHAP values) to determine the main feature influence weights that dominate the diagnostic results. Locally, the main input features of the prediction basis can be explained through LIME, Anchor, etc. For image tasks, visual explanations such as Grad-CAM and Score-CAM are used to understand the main recognition areas of the model and how to identify tumors. For text tasks, it can be used to highlight the key words and important sentences that the model is concerned about, helping doctors understand the process of the model.

For the issue of data security, a federated learning structure can be adopted, enabling the model training of each medical institution to be carried out independently in its own environment. Joint modeling can be achieved without sharing the original data, that is, privacy is well guaranteed. Technologies such as encrypted computing and differential privacy can also enhance the security of data. For possible malicious behaviors and heterogeneous inputs, adversarial training and robustness evaluation systems can be adopted to enhance their robustness and security.

Conclusion: Medical data intelligence and machine learning are comprehensively promoting the development of cancer diagnosis technology. By combining multiple types of data, optimizing feature extraction and model optimization, and enhancing the transparency, stability and robustness of the system, the accuracy and speed of cancer identification have been significantly improved. While ensuring personal information security and model stability, The artificial intelligence diagnosis and treatment system has great practical value. In the near future, it is expected to achieve full-process services from initial screening to personalized diagnosis and treatment, leading cancer diagnosis and treatment into the era of precision.

## 5. Conclusion

The integration of medical data intelligence and machine learning is fundamentally reshaping the landscape of cancer diagnosis. By leveraging multi-source data-including imaging, pathology, genomic profiles, and clinical text-alongside advanced algorithms for feature extraction and model optimization, diagnostic systems have achieved significant gains in both accuracy and efficiency. These intelligent frameworks not only enhance early detection and risk stratification but also enable more personalized and timely decision-making. Key contributions of this approach include robust feature engineering across modalities, the use of ensemble and multi-task learning for performance improvement, and the incorporation of explainability and data security into clinical workflows. Tools such as SHAP, LIME, Grad-CAM, and federated learning architectures ensure interpretability and safeguard patient privacy-critical prerequisites for real-world deployment in healthcare settings. As artificial intelligence systems become more transparent, stable, and adaptable, their clinical value continues to grow. Looking ahead, intelligent diagnostic platforms are expected to support the full continuum of care-from initial screening through diagnosis to individualized treatment planning-thus propelling oncology into the next era of precision medicine.

## References

1. M. J. Iqbal, Z. Javed, H. Sadia, I. A. Qureshi, A. Irshad, R. Ahmed, and J. Sharifi-Rad, "Clinical applications of artificial intelligence and machine learning in cancer diagnosis: looking into the future," *Cancer cell international*, vol. 21, no. 1, p. 270, 2021. doi: 10.1186/s12935-021-01981-1
2. E. V. Varlamova, M. A. Butakova, V. V. Semyonova, S. A. Soldatov, A. V. Poltavskiy, O. I. Kit, and A. V. Soldatov, "Machine learning meets cancer," *Cancers*, vol. 16, no. 6, p. 1100, 2024. doi: 10.3390/cancers16061100
3. P. N. Kamalapathy, M. R. Gonzalez, T. M. de Groot, D. Ramkumar, K. A. Raskin, S. Ashkani-Esfahani, and S. A. Lozano-Calderón, "Prediction of 5year survival in soft tissue leiomyosarcoma using a machine learning model algorithm," *Journal of Surgical Oncology*, vol. 129, no. 3, pp. 531-536, 2024. doi: 10.1002/jso.27514
4. S. Bendale, A. Parai, S. Deshpande, A. Iyer, and A. Kumbhare, "Predictive Brain Cancer Detection and Treatment Using Machine Learning and Artificial Intelligence," *International Journal of Chemical Separation Technology*, vol. 9, no. 1, pp. 28-39, 2023.
5. M. Farsi, "Filter-Based Feature Selection and Machine-Learning Classification of Cancer Data," *Intelligent Automation & Soft Computing*, vol. 28, no. 1, 2021. doi: 10.32604/iasc.2021.015460

**Disclaimer/Publisher’s Note:** The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of GBP and/or the editor(s). GBP and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.