International Journal of Biology and Life Sciences
ISSN: 2957-9511 | Vol. 9, No. 3, 2025

Research on Federated Learning Model for Assisted Diagnosis in Cancer Rehabilitation-Based on AMO-XGBoost Model

Zebao Zong [1,*], Qinyu Zhang[1], Li Yuan [2], Zihan Xiao [3], Kenan Wang [1]

[1] College of Science, North China University of Science and Technology, Tangshan Hebei, 063210, China
[2] College of Clinical Medicine, North China University of Science and Technology, Tangshan Hebei, 063210, China
[3] College of mechanical engineering, North China University of Science and Technology, Tangshan Hebei, 063210, China

*** Corresponding author:** Zebao Zong

Abstract: With the application of AI in healthcare, privacy and security issues are a growing concern. Protecting the privacy of patient data and controlling the loss of model accuracy becomes critical when training AI models. Cancer is an increasing threat to human health, and postoperative patients are at high risk of recurrence; predicting the time and location of recurrence is critical. This study uses federated learning combined with localised differential privacy and the AMO-XGBoost model to investigate a privacy-preserving model suitable for cancer recovery patient data. The model is tested on a dataset of 5 types of cancer in rehabilitation patients, aiming to construct a privacy-preserving and accurate model for assisted diagnosis in cancer rehabilitation.

Keywords: Combining Federated Learning; Protect Privacy and Improve Cancer Rehabilitation Diagnostic Model Accuracy; Local Differential Privacy and AMO-XGBoost.

1. Introduction

As AI technology continues to evolve, more and more healthcare data is being collected and used. These data are extremely valuable to healthcare organisations and researchers, not only to help doctors make more accurate diagnoses and decisions, but also for research into the pathogenesis and treatment of diseases. However, the use of medical data also faces significant privacy challenges. Since medical data contains a large amount of sensitive information, such as personal identity, medical history, and disease type, once these data are leaked or misused, it will bring great risks and troubles to patients. Therefore, how to protect the privacy and security of medical data has become a very important issue.

With data security issues raising concerns globally, China implemented the Cybersecurity Law of the People's Republic of China in 2017, which puts forward mandatory requirements for the protection of personal information, and other relevant laws and regulations are also in the pipeline.2021 China published the Personal Information Protection Law of the People's Republic of China, which marks the birth of three important laws in the field of cyberspace security in China that effectively protect the privacy and security of the country and its citizens. Internationally, the European Union started to implement the General Data Protection Regulation (GDPR) on 25 May 2018, which puts forward extremely stringent requirements on the collection and processing of users 'personal data, and is regarded as the most stringent standard in the field of users' personal privacy protection and data security.

With the continuous advancement of medical technology and equipment, the number of treatment methods for tumours has been increasing, but whatever means are used, they will cause a certain degree of damage to the patient's body. The World Health Organisation has clearly pointed out that cancer is a lifestyle disease and its pathogenesis is related to factors such as geographical influence, working environment, living habits and drug use. Therefore, the quality of life of cancer patients should be taken seriously, and rehabilitation and nursing care become more and more important in the fight against cancer. Disease rehabilitation therapy is an emerging cross-discipline combining rehabilitation medicine and oncology. Currently, medical diagnosis mostly relies on the quality of medical images and the interpretation ability of radiologists. However, in the diagnostic interpretation of test results, the diagnosis made by human physicians based on experience has certain limitations and cannot be efficient. With the development of various medical testing equipment and imaging techniques, more and more medical data needs to be assimilated and interpreted by expert physicians. Over the past 20 years, with advances in computer technology, computers have attempted to provide a 'second opinion' for radiologists.In the past two decades, with advances in computer technology, computers have attempted to provide radiologists with a 'second opinion' in order to improve medical imaging. Computer-Aided Diagnosis (CAD) has been used in the screening of many diseases, while machine learning algorithms have been developed to make the diagnosis of diseases more intelligent.

2. Content of the Study

A dataset of cancer rehabilitation patients was constructed, and 51 secondary indicators and 9 primary indicators were collected through patients' physical examination reports. By combining the experience of expert doctors with data analysis and the testing of localised machine learning models, seven experimental input indicators were finally identified to determine the federal learning data access criteria. As shown in the table below, the weight of each module indicates the association between this module and cancer recurrence, and the weight is scored from 0 to 10, with 0 indicating no correlation with tumour recurrence and 10 indicating direct association.

Table 1. Patient data set

Level 1 indicators	Related items (weights)
basic module	systemic diseases (2); family history of cancer (2); nonessentialEssential Dependence (3); Obesity (1); Habitual High Risk (1) Geographic risk (1)
Tumor Module	Size (10); Occupancy (10); Invasive relationship (10). Angiogenesis (10); Pathologic typing (3); CTC value (9) Differentiation (10); Mutant target (1)
Immunization module	CD3+CD4+CD8+/CD45+ (4) ; CD4+CD25+ (1) ; CD3+CD4+/CD45+ (8) ; CD3-CD56+/CD45+ (8) CD3+CD4+/CD45+ (8) ; CD3-CD56+ (5) ; CD4+/CD8+ (10) ; CD3+CD16+CD56+ (5) ; CD4+/CD8+ (10) CD3+CD16+CD56+/CD45+ (6) ; exercise ECG (X ± SD) (2) ; Exercise electrocardiogram (X ± SD) (2)
Nutrition module	Total Nutrition (6); Balanced Nutrition (3); Nutritional Safety Assessment (5); Cancer Cell Proliferation (10); Immune Cell Proliferation (10) Cancer cell proliferation (10); Immune cell proliferation (10); Angiogenesis (8); Amino acid evaluation (5); Amino acid evaluation (5) (Cancer cell proliferation (10); Immune cell proliferation (10); Angiogenesis (8); Amino acid assessment (5); Proteomic assessment (10). Proteomics assessment (10)
Psychological module	Life Events Scale (1); Cornell Medical Index (2). Anxiety Self-Rating Scale (5); Depression Self-Rating Scale (5); Beck Anxiety Scale (5); Beck Depression Inventory (5); Pittsburgh Pittsburgh Sleep Quality Index (4); Texas Social Behavioral Questionnaire (3); Family Functioning Assessment (1); Exercise Electrocardiogram (X±SD) (2); Exercise electrocardiogram (X±SD) (2)
Microenvironmental modules	O2 (3); PH value (4); interstitial pressure (2); inflammatory response (7); vascular permeability (6); CTC values (9). Protein profiling (8)
Motion Module	Aerobic Exercise (4); Social Progression Assignment (3); Dirk Sass Social Behavior Questionnaire (3)

Combined with the cancer rehabilitation patient dataset, the federated learning privacy protection model for cancer rehabilitation assisted diagnosis is constructed to establish the federated learning model, personalised local differential privacy method, and algorithms such as AMO-XGBoost and convolutional neural network that are selected to fit the federated learning framework. Federated Learning (FL) is a distributed machine learning approach that allows multiple participants to train models on local data without sharing the data itself. Each participant trains the model on its local data and sends the updated model parameters to a central server for aggregation and updating. The updated global model parameters are then distributed back to the participants for the next round of training. This approach can effectively protect data privacy and reduce the risk of data leakage, which is especially suitable for the medical field.

Test the constructed federated learning model with AMO-XGBoost as the core algorithm, the federated learning model incorporating personalised local differential privacy and convolutional neural network as the core model inputting multi-party hospital data to protect the privacy data and at the same time substantially increase the amount of model training data. In machine learning, model training requires the use of a large amount of data, which contains a lot of sensitive information. If these data are leaked or misused, it will bring great risks and troubles to individuals. Differential privacy can protect data privacy while allowing model training and data sharing, and thus has important application value in the field of machine learning. Differential privacy provides an information-theoretic security guarantee that makes the output of a function insensitive to any particular record in the dataset. Therefore, in the context of federated learning, localised differential privacy can fit well with the encryption process required by the federated learning framework, as shown in Figure 1 below, preprocessing the data with the idea of localised differential privacy and then adopting the federated learning framework for the subsequent operations can improve the security of the data as well as the user's security.

AMO-XGBoost complete objective function:

$$L_{AMO}^{(t)} = \sum_{i=1}^{n} [g_i f_t(x_i) + \frac{1}{2} h_i f_t^2(x_i)] + \gamma T + \frac{1}{2} \sum_{j=1}^{T} (\lambda_j \omega_j^2 + \alpha \,|\, \omega_j \,|)$$

After the construction of the federated learning privacy-preserving model for cancer rehabilitation assisted diagnosis, the loss of model accuracy is further determined and compared with other models while protecting patient privacy.

3. Federal Learning State of the Art in Privacy Protection Research in Different Industries

Although federated learning has received extensive application attempts since its introduction, most of them still remain in the exploratory stage. In the future practical application, a perfect federated learning system needs to be established. When machine learning is used to identify and make decisions on classified data, federated learning can effectively deal with the problems faced by privacy protection and ensure that sensitive data is adequately protected.

Horizontal federal learning, vertical federal learning, and migration federal learning are three federal learning models that can fully improve the availability and utilisation value of different data in various industries [1].

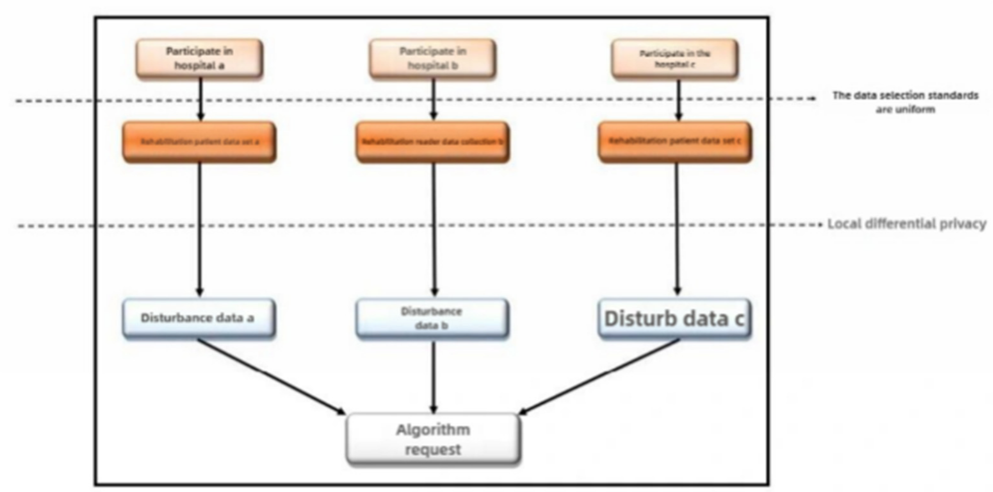

Figure 1. Localised differential privacy client application process

Since the emergence of machine learning, public attention has been focused on the accuracy and time efficiency of the training results, but currently more concerned about privacy protection. The application of federated learning effectively solves the problem of data privacy protection in most cases. In addition, data sharing in federated learning models can significantly expand the data volume and further improve the accuracy of training results under the premise of data security. Currently, at the application level, the most common way of federated learning is to combine artificial intelligence algorithms. The combination process is shown in Figure 2.

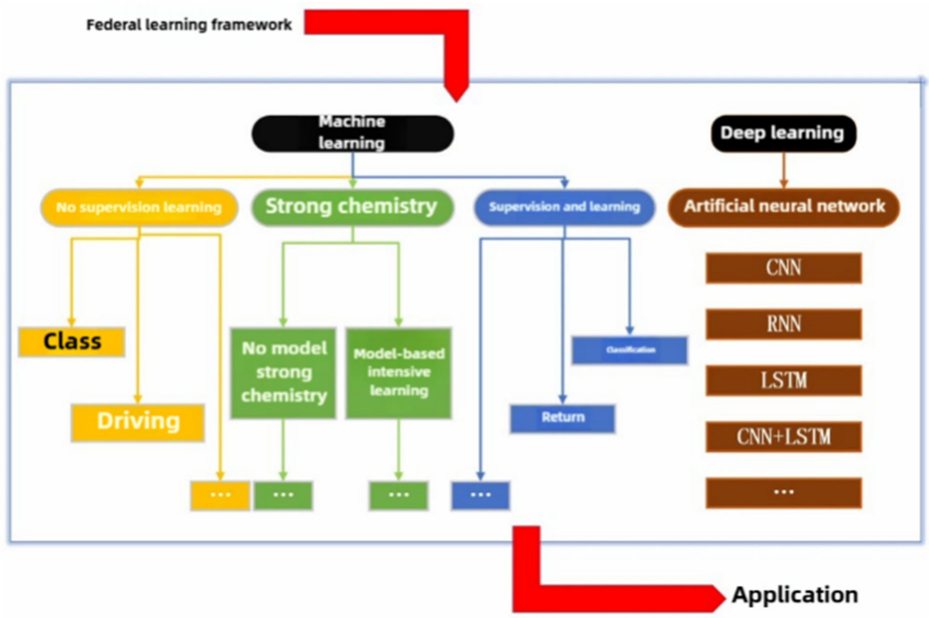

Figure 2. Federal Learning Framework

From Figure 2, it can be seen that the federated learning framework can combine multiple algorithm types, including aggregation algorithms, differential privacy algorithms, deep learning algorithms, meta-learning algorithms, and reinforcement learning algorithms. These algorithms can be adapted and combined according to specific application scenarios and requirements to achieve effective model training and data sharing within the federated learning framework. Through continuous improvement and innovation, Federated Learning will provide feasible solutions to the problems of data sharing and model training in distributed environments. Since federated learning was proposed, the healthcare industry has been exploring its applications in parallel. In recent years, the rapid development of artificial intelligence technology has made the development and application of medical expert systems and artificial neural networks in the healthcare industry a reality, and significant breakthroughs have been achieved.[7-8] However, despite our joy about this, we inevitably face some problems, such as the lack of model training accuracy, the limited amount of patient data in individual hospitals, and the challenge of sharing medical data [2-4].Kaissis G A et al. introduced the application of federated learning in medical imaging, and also pointed out the potential safety hazards faced, and other other methods applied in medical imaging were presented. In addition, Yan Z et al. proposed a Variable Awareness Federated Learning Frame (VAFL) and applied it to the medical industry. This framework has improved stability compared to the basic federated learning. Federated learning also shows excellent results in medical image privacy protection. Zhang C et al. proposed a method to enhance the weights during medical image training, adding privacy protection and encryption methods through different weights, effectively preventing the problem of excessive model visibility. In addition to medical imaging, the application of federated learning in the pharmaceutical field has also attracted much attention. Chen S et al. combined federated learning with quantitative constitutive

Chen S et al. combined federated learning with quantitative structure-effect relationship (QSAR) analysis to break the phenomenon of 'data silos' in pharmaceutical companies, and realised data sharing under the premise of protecting the privacy of enterprises. In MRI, Li X et al. showed that multi-site data without data sharing can improve the performance of neuroimaging analysis and find reliable disease-related biomarkers, and Stripelis D and Linardos A et al. used a federated learning framework to extend neuroscience and cardiovascular disease research, effectively overcoming the problem of unshared medical data.

Since the 20th century, with the improvement of information storage capacity and information processing speed, it has promoted the rapid development of data storage industry and big data information technology, and injected a huge impetus for the rise and development of emerging industries. Currently, the amount of data generated in the medical field is growing exponentially. Through effective data resource storage and transmission management technology, combined with big data mining technology, the utilisation efficiency and intelligence of data in the medical field have been significantly improved [5-6], injecting a new impetus for the long-term development of the medical field.

As the material standard of living of human society continues to improve, the living environment and lifestyle of human beings have also changed accordingly, resulting in cancer becoming one of the most important factors threatening human health. According to the Global Cancer Incidence and Mortality Report 2021 (GLOBOCAN 2018) compiled by the International Agency for Research on Cancer (IARC), it is estimated that there will be 19.29 million new cancer cases and 9.96 million cancer deaths worldwide. Since the concept of 'oncological rehabilitation' was first proposed in the United States in 1971, the approach to cancer rehabilitation has been largely dependent on the nature and stage of development of the tumour [7]. Early detection of cancer and rehabilitation have a significant impact on the survival rate and quality of life of patients. However, the complexity of factors affecting postoperative recurrence in cancer patients makes accurate prediction of postoperative conditions and trends challenging.

In this regard, many scholars have carried out a lot of work, including the prediction of benign and malignant classification of tumours, the prediction of postoperative recurrence time as well as the study of tumour types and development trends based on the data of cancer patients' condition assessment [8-10]. The continuous development in the field of machine learning and deep learning has provided an important impetus for assisted cancer diagnosis and treatment [11-12]. Statistical theory and a variety of machine learning techniques, such as support vector machines, artificial neural networks, decision trees, and deep learning frameworks, have been widely used in cancer diagnosis. In the field of assisted cancer diagnosis and treatment, many scholars have used data mining techniques to study the prediction of cancer prognostic regression trends and to mine relevant patterns from complex cancer risk factors. Doctors use data mining techniques to analyse and predict the relevant conditions of patients, which is crucial for their rehabilitation. This technology helps doctors to better understand patients' health status, detect signs of disease in advance, develop personalised treatment plans, and improve the effectiveness of treatment and speed of recovery. In addition, data mining technology can also help doctors to prevent diseases and improve patients' quality of life.

In addition, data mining technology can also help doctors to prevent diseases and improve patients' quality of life.

4. Significance of the Study

Federated learning can share and utilize data while protecting data privacy, and its machine learning algorithms are not limited to linear models, tree models, neural network models, etc., and can be used to solve prediction problems such as classification and regression.

Federated Learning can effectively address the issue of privacy protection when sharing data across organizations, and since organizations can share their data, the technique can improve the accuracy of the models and deliver better results. The implementation of Federated Learning technology can benefit the organizations for better collaboration and growth. As federated learning continues to evolve, the model is being used in a variety of fields. Meanwhile, diagnosing whether a cancer is an in-situ recurrence and predicting the time of recurrence are crucial to the patient's next step in rehabilitation.

According to statistics, about 90% of cancer deaths are caused by failure to prepare for cancer recurrence and failure to determine whether cancer cells have metastasized in time. Therefore, it is of great significance for cancer clinical diagnosis and treatment to accurately predict the location of cancer recurrence and the time of recurrence under the premise of guaranteeing data security.

5. Feasibility Analysis

We applied a variety of basic machine learning algorithms including Linear Regression (LR), Bayesian Regression (BR), Support Vector Regression (SVR), Gradient Boosting Tree (GBDT), Extreme Gradient Boosting (XGBoost) and Multi-Layer Perceptual Machine Neural Networks (MLPR) to initially model the dataset. This phase not only validates the validity of the dataset, but also provides baseline performance metrics for the construction of subsequent federated learning models. Through parameter optimization, we ensured that these models could perform optimally, demonstrating the feasibility of the dataset for constructing cancer-assisted diagnosis models. By comparing the prediction performance of the underlying models, we find that deep learning models such as MLPR exhibit higher prediction accuracy, demonstrating that complex model structures have significant advantages when dealing with high-dimensional data such as medical images.

Theoretical feasibility

In the context of increasingly sophisticated machine learning, where automatic identification and intelligent decision-making have been realized, federated learning has emerged as a highly promising solution to the data privacy protection problem. In the process of federated learning, the training data remains local to the participants, which both realizes the sharing of training data across participants and ensures the protection of each participant's privacy. The basic workflow consists of participants downloading the initialized global model from the cloud server, training the model using the local dataset, and generating the latest local model updates (i.e., model parameters). Subsequently, the cloud server collects the individual local update parameters and updates the global model via a model averaging algorithm. Federated learning has the unique advantage of enabling the training of a unified machine learning model from the local data of multiple participants while protecting data privacy. The main innovation is to provide a distributed machine learning framework with privacy-preserving features that enables iterative training of a particular machine learning model with thousands of participants in a distributed manner.

Feasibility of technical means

Artificial intelligence has made remarkable achievements in the fields of natural language processing, computer vision and speech recognition. Meanwhile, in the medical field, a large amount of cancer data has been collected and made available for AI research. Compared to the traditional way of cancer diagnosis, AI does not rely on explicit instructions, but rather identifies and recognizes specific patterns from complex data by recognizing patterns and reasoning to effectively predict cancer. However, the prediction of cancer recurrence remains one of the most challenging and intriguing tasks in the field of AI. In this study, we applied AI algorithms to the field of intelligent healthcare and constructed a cancer recovery recurrence prediction model using multiple impact indicators of patients. The model uses basic indicators, immune indicators, tumor indicators, microenvironmental indicators, psychological indicators, nutritional indicators, aerobic exercise indicators, and advanced homework indicators as attributes, and cancer recurrence time and life value assessment as predictive labels, to assist doctors in predicting the cancer recurrence problem.

Feasibility of data support

The data on cancer patients' recovery comes from a horizontal project, provided by Beijing Stairway Health Technology Co. A large amount of cancer patient data has been collected, and the weights are determined based on the experience of physician experts. The dataset is based on long-term medical consensus, the ASC Cancer Factors Study Report, and cancer assessment data from TIES.IO. Due to the wide age range of cancer patients undergoing surgery, model predictions under 25 years of age are significantly biased, whereas other systemic diseases may be present in patients over 65 years of age affecting the data. Therefore, only cancer patients between the ages of 25 and 65 years were considered for the study.

6. Conclusion

The constructed federated learning-based AMO-XGBoost model has an average prediction accuracy of 89.65% on the five cancer rehabilitation datasets, which significantly outperforms the local XGBoost model of 83.21%, as well as the traditional logistic regression and random forest models, which only have accuracies of 78.95% and 81.34%, respectively. The introduction of personalized local differential privacy results in a slight decrease in model accuracy of 88.43%, but the loss of accuracy is only 1.22% compared to the federated learning model without privacy protection.

Flexible control of privacy is realized by adjusting the privacy parameter ε of local differential privacy. When ε is set to 1.0, the accuracy of the model is 87.56%, and as ε is increased to 2.5, the accuracy rises to 88.92%, but the strength of privacy protection decreases at this time. In practical applications, the value of ε can be adjusted according to the specific privacy requirements to balance the accuracy and privacy protection.

Table 2. Quantitative indicators of the model

Mould	Accuracy (%)	Precision (%)	Recall rate (%)
Federated Learning AMO-XGBoost	89.65	90.23	89.01
Local XGBoost	83.21	84.56	82.11
Logistic regression	78.95	79.87	78.22
Random forest	81.34	82.45	80.67

Table 3. Effect of privacy parameters on the model

Privacy parameters ε	Accuracy (%)	Privacy risk
0.5	86.21	Lower
1.0	87.56	Moderate
1.5	88.12	Moderately high
2.0	88.56	High
2.5	88.92	Higher

Acknowledgments

This work was supported by 2024 North China University of Science and Technology Undergraduate Innovation and Entrepreneurship Training Program Project(X2024173).

References

[1] Tie Zhang, Linli Xu, Yuan Zhou. Clustered federated learning based on prototype networks[J]. Small Microcomputer Systems, 2024, 45(1):1-8.

[2] Ma, J.; Chen, L.; Xu, J.; Yuan, Y. FedCrow: Federated-Learning-Based Data Privacy Pre servation in Crowd Sensing. Appl. Sci. 2024, 14, 4788. https://doi.org/10. 3390/ app 1411 4788.

[3] Zhao, R.; Lu, J.; Liu, Z.; Wang, T.; Guo, W.; Lan, T.; Hu, C. A Federated-Learning Al gorithm Based on Client Sampling and Gradient Projection for the Smart Grid. Electronics 2024, 13, 2023. https://doi.org/10.3390/electronics13112023.

[4] GUO Fanghong, LIU Shishuo, WU Xiang,et al. Fault diagnosis of power transformer containing unbalanced sample data based on federated learning[J]. Power System Automation, 2023, 47(10):145-152.

[5] Al-Karawi, D.; Al-Zaidi, S.; Helael, K.A.; Obeidat, N.; Mouhsen, A.M.; Ajam, T.; Alsha labi, B.A.; Salman, M.; Ahmed, M.H. A Review of Artificial Intelligence in Breast Imaging.Tomography 2024, 10, 705-726. https://doi.org/ 10.3390/ tomography10050055.

[6] ORUJOV F, MASKELIŪNAS R, DAMAŠEVIČIUS R, et al. Fuzzy based image edge detection algorithm for blood vessel detection in retinal images[J]. Applied Soft Computing, 2020, 94:106452.

[7] FU J B, MOLINARES D M, MORISHITA S, et al. Retrospective analysis of acute rehabilitation outcomes of cancer inpatients with leptomeningeal disease[J]. Pm&r, 2020, 12 (3):263-270.

[8] MONTAZERI M, MONTAZERI M, MONTAZERI M, et al. Machine learning models in breast cancer survival prediction [J]. Technology and Health Care, 2016, 24(1):31-42.

[9] Xu, X.; Li, J.; Zhu, Z.; Zhao, L.; Wang, H.; Song, C.; Chen, Y.; Zhao, Q.; Yang, J.; P ei, Y. A Comprehensive Review on Synergy of Multi-Modal Data and AI Technologies in Medical Diagnosis. Bioengineering 2024, 11, 219. https://doi.org/ 10. 3390/ bioengineering11030 219.

[10] Kalatzis, D.; Spyratou, E.; Karnachoriti, M.; Kouri, M.A.; Stathopoulos, I.; Danias, N.; Arkadopoulos, N.; Orfanoudakis, S.; Seimenis, I.; Kontos, A.G.; et al. Extended Analysis o f Raman Spectra Using Artificial Intelligence Techniques for Colorectal Abnormality Classifi cation. J. Imaging 2023, 9, 261. https://doi.org/10.3390/jimaging9120261.

[11] ZHU Xiaoteng, GONG Shixia, XU Xiaoqing, et al. Exploration and reflection on the new model of cancer care based on 'Internet+' [J]. Chinese Health Care, 2023, 41(9):15-18.

[12] YANG A M, HAN Y, LIU C S, et al. D-TSVR recurrence prediction driven by medic al big data in cancer[J]. IEEE Transactions on Industrial Informatics, 2020, 17(5):3508-3517.