2025 International Conference on Intelligent Computing
July 26-29, Ningbo, China
https://www.ic-icc.cn/2025/index.php

Prediction Research of TACE Treatment Response Based on Multimodal Data Fusion

Lin Tong[1], Zhengkui Chen[2], Jun Luo[3], Qingli Zhou[4], Yuqiang Shen[4] and Jijun Tong[2(✉)]

[1] School of computer science and technology, Zhejiang Sci-Tech University, Hangzhou 310018, China
[2] School of Information Science and Engineering, Zhejiang Sci-Tech University, Hangzhou 310018, China
[3] Zhejiang Cancer Hospital, Hangzhou Institute of Medicine (HIM), Chinese Academy of Sciences, Hangzhou, Zhejiang 310022, China
[4] Department of Information Technology, the Fourth Affiliated Hospital of School of Medicine, and International School of Medicine, International Institutes of Medicine, Zhejiang University, Yiwu 322000, China

Abstract. Transcatheter arterial chemoembolization (TACE) is the preferred non-surgical treatment for HCC patients, but up to 60% of HCC patients do not benefit from TACE treatment. Therefore, accurately and efficiently predicting the treatment response after TACE in HCC patients is of great significance for treatment planning. To address this, a predictive model based on the integration of clinical data and preoperative CT imaging was proposed. This model first used a convolutional neural network to extract microscopic structural features from CT images, then inputed the extracted features into a Long Short-Term Memory network to obtain the global feature vector for each CT slice. Next, a deep neural network was used to extract features from the clinical data. Finally, the features were fused using an asymmetric cross-attention mechanism, followed by classification using a feedforward neural network. A retrospective study was conducted on 181 HCC patients who underwent TACE treatment at a hospital in Zhejiang Province from January 2018 to April 2022. The AUC, precision, accuracy, and recall of the prediction model are 0.85, 0.86, 0.88, and 0.87, respectively. The experimental results demonstrate that the model (Cnn-Lstm-Dnn-Cross-Attention, CLDCA) outperforms the comparison models, providing an effective solution for predicting the post-TACE treatment response in HCC patients.

Keywords: Multimodal fusion, Transcatheter arterial chemoembolization, Cross-attention mechanism, Convolutional neural networks.

1 Introduction

Hepatocellular carcinoma (HCC), as a malignant tumor, has become a major global health concern due to its persistently high incidence and mortality rates, posing a serious threat to human health. According to statistics from the World Health Organization

(WHO), liver cancer is the sixth most common cancer worldwide and the third leading cause of cancer-related deaths [1]. In its early stages, liver cancer often presents no specific symptoms. Approximately 80% of patients are diagnosed at an intermediate or advanced stage when surgical resection is no longer feasible, thereby missing the optimal window for surgery. This leads to increased treatment difficulty and a poor prognosis [2]. According to the National Cancer Center of China, there were approximately 367,700 new liver cancer cases in 2022, making it the fourth most commonly diagnosed cancer in the country. Liver cancer also caused 316,500 deaths, ranking second among all cancer-related causes of death [3]. The high incidence and mortality rates of liver cancer are closely associated with several major risk factors, including infection with hepatitis B virus (HBV), hepatitis C virus (HCV), alcoholic liver disease, and non-alcoholic fatty liver disease (NAFLD) [4]. With an aging population, changes in lifestyle, and the rising prevalence of metabolic disorders, the disease burden of liver tumors is expected to increase further. According to clinical guidelines for the diagnosis and treatment of primary liver cancer, current therapeutic approaches include hepatic resection, liver transplantation, ablation therapy, transcatheter arterial chemoembolization (TACE), radiotherapy, and systemic anti-tumor therapies. Selecting the appropriate treatment based on cancer staging can help maximize therapeutic efficacy [5]. According to the Barcelona Clinic Liver Cancer (BCLC) staging system, TACE is currently the main treatment modality for patients with intermediate-stage HCC [6-7]. TACE works by directly delivering a combination of chemotherapeutic agents and embolic materials into the arteries supplying the tumor, thereby blocking its blood supply and locally releasing high concentrations of chemotherapy to inhibit tumor growth. The advantages of TACE include its minimally invasive nature, repeatability, and relatively limited impact on liver function, making it particularly suitable for patients with unresectable intermediate or advanced liver cancer. However, the therapeutic response to TACE varies significantly among individuals; some patients may experience tumor recurrence or metastasis after treatment, and complications such as hepatic function deterioration may also occur [8]. Therefore, optimizing TACE treatment strategies and accurately predicting patient outcomes and prognosis have become key research priorities in the field of liver cancer treatment. Studies have shown that therapeutic response following TACE is an independent prognostic factor for clinical outcomes in HCC patients, and it provides critical guidance for the development of individualized treatment strategies [9-10].

Against the backdrop of rapid advancements in imaging and artificial intelligence technologies, researchers are increasingly focusing on integrating multi-source data to improve the accuracy of prognosis prediction. Kim et al. [11] demonstrated that prognostic models combining radiomic features from preoperative CT scans with clinical variables outperform models based on a single data source in predicting survival outcomes. Morshid et al. [12] developed a machine learning algorithm that significantly improved the prediction accuracy of TACE treatment response by integrating quantitative CT features with clinical factors. Guo et al. [13] conducted a systematic review of multimodal representation learning, offering an in-depth analysis of fusion strategies across various data modalities and their recent developments.In the integration of pathology and imaging studies, Saillard et al. [14] used deep learning to analyze whole-

slide digitized tissue images, achieving superior performance in predicting postoperative survival in HCC patients compared to conventional methods. Liu et al. [15] employed machine learning to extract CT imaging features and integrated them with a multi-task deep learning algorithm to jointly predict microvascular invasion, tumor grading, and long-term survival. Han et al. [16] developed a prognostic model based on large-scale multicenter data that enables risk stratification based on patient characteristics and treatment response, providing valuable guidance for personalized therapy. The study by Peng et al. [17] further confirmed the advantage of combining deep learning with radiomics in predicting initial response to TACE. From a technological innovation perspective, Wang Yuqi et al. [18] developed a Vision Transformer-based model that achieved excellent performance in predicting post-TACE recurrence risk by integrating preoperative clinical and imaging data. The HCC dataset established by Moawad et al. [19] has provided crucial support for developing algorithms for treatment response prediction and tumor segmentation. Chang et al. [20] showed that models based on multiphasic CT imaging significantly outperformed traditional methods in predicting response to TACE. Sun et al. [21] proposed the DLOPCombin model, which extracts imaging features using an improved residual network and integrates clinical variables to successfully predict overall survival in TACE patients. Zhou et al. [22] developed the IRENE model, which enables unified processing of medical imaging, textual, and structured data, demonstrating outstanding performance in disease identification and prognosis prediction. Pino et al. [23] introduced the TwinLiverNet model, which combines 3D convolutional networks with capsule networks and achieved an 82% prediction accuracy in multiphasic CT image analysis. Xi Zihan et al. [24] highlighted that combining functional MRI imaging with multimodal data significantly enhances the predictive performance for TACE treatment efficacy. These studies provide essential technological support for the precision treatment of HCC.

Building on previous research, this paper proposed a multimodal data fusion model (Cnn-Lstm-Dnn-Cross-Attention, CLDCA) for predicting TACE treatment response. This predictive model enhanced the ability to forecast TACE treatment outcomes for HCC patients by integrating the sequential features of CT images with clinical data.

2 Database and Methods

2.1 Database

This study conducted a retrospective analysis of 181 HCC patients who underwent TACE treatment at a hospital in Zhejiang Province from January 2018 to April 2022. The inclusion criteria were: (1) HCC diagnosed by preoperative biopsy or surgical pathology and clinical diagnostic standards; (2) patients with HCC who underwent TACE; (3) availability of mRECIST [25] post-TACE evaluation data. The exclusion criteria were: (1) patients with concomitant other malignant tumors; (2) patients with missing or incomplete clinical data; (3) patients who received other treatments such as radiotherapy, chemotherapy, or liver transplantation before or during follow-up. The flowchart for the screening of the study population is detailed as **Fig. 1**.

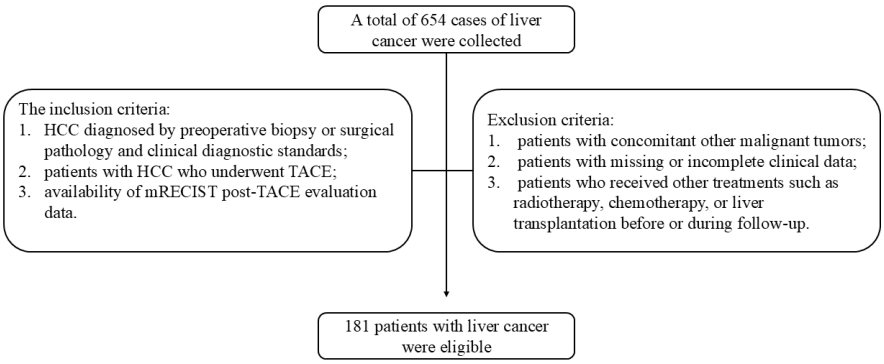

Fig. 1. Flowchart for the screening of research subjects.

Collection of detailed clinical information included past liver disease history (HBV, HCV, cirrhosis, fatty liver), family history of liver cancer, history of liver transplantation, ascites, total bilirubin, albumin, prothrombin time, BCLC staging, number of liver tumors, largest liver lesion diameter, presence of vascular invasion, presence of extrahepatic metastasis, presence of portal vein tumor thrombus, white blood cell count, hemoglobin, platelet count, alanine aminotransferase (ALT), aspartate aminotransferase (AST), alkaline phosphatase, total bilirubin, indirect bilirubin, direct bilirubin, urea, creatinine, albumin, triglycerides, cholesterol, low-density lipoprotein, potassium, sodium, prothrombin time, international normalized ratio, partial thromboplastin time, fibrinogen, height, weight, heart rate, respiratory rate, high and low blood pressure values, pulse rate, temperature, and 43 other laboratory test results.

All imaging examinations of liver-enhanced CT in this study were performed in the imaging department of the hospital, and patients were examined preoperatively on a CT scanner (Somatom Sensation 64, Simens Healthcare, Erlangen, Germany). The scanning parameters were as follows: tube voltage: 120 KV, tube current: 308 mA, layer thickness: 5 mm, and layer spacing: 3.5 mm. 80 to 90 mL of iodinated contrast agent (lopromide, Ultravist 370; Bayer, Germany) was injected from the back of the hand or elbow via a power syringe at a rate of 2.5 mL/sec. Unenhanced CT images were obtained first, followed by contrast injection, and images were acquired during the arterial phase for 30 seconds, the portal venous phase for 60 seconds, and the delayed phase for 90 seconds. Each image was resized to 600×450 pixels at a dpi of 96. Image pre-processing was performed prior to the experiment, including normalisation of the image pixel values to the range of [0,1] and resizing of the image using nearest interpolation.

All patients were assessed by mRECIST within 6-9 weeks after TACE treatment and were categorised into complete remission (CR), partial remission (PR), stable disease (SD) and progressive disease (PD). In this paper, the disease control rate (CR+PR+SD) [26] was used to decide whether the patients should continue TACE treatment subsequently according to the actual situation in the hospital.

2.2 Methods

The model designed primarily consists of three modules (as **Fig. 2**): the feature extraction module, multimodal fusion module, and classification module. The feature extraction module mainly extracted features from different modules, including image features and clinical data features. In the multimodal fusion module, an asymmetric cross-attention mechanism was used to fuse and interact the feature representations of the two modalities, outputting the fused multimodal features. These fused multimodal features were not directly used for the classification task but were input into a self-attention mechanism for deep feature extraction. This aggregated multimodal information allowed the model to learn the deep relationships between different modalities, leading to more reliable classification.

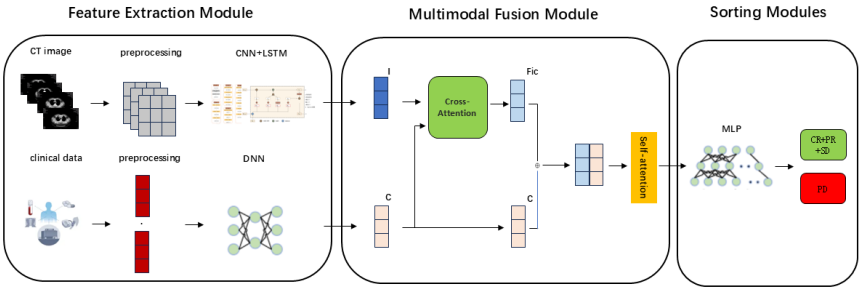

Fig. 2. CLDCA model framework.

CT image feature extraction. To address the dynamic variation in the number of CT slices in medical imaging data, this study proposes an innovative multi-input channel network architecture. The core design philosophy of this module is to transform three-dimensional CT data into a hybrid processing paradigm that integrates two-dimensional convolution and temporal sequence modeling. Prior to feeding the CT images into the ResNet50 network for feature extraction, we performed preprocessing on the raw CT images. As CT images typically have high resolution and are susceptible to noise, we applied normalization to map pixel values from their original range to the [0, 1] interval, mitigating the impact of scale differences on model training.Since the ResNet50 architecture requires input images of fixed dimensions, all CT images were resized to 224×224 pixels. This size was experimentally validated to preserve essential image details while avoiding excessive computational overhead. During the feature extraction stage, a modified ResNet-50 architecture was employed as the backbone encoder, as **Fig. 3**.

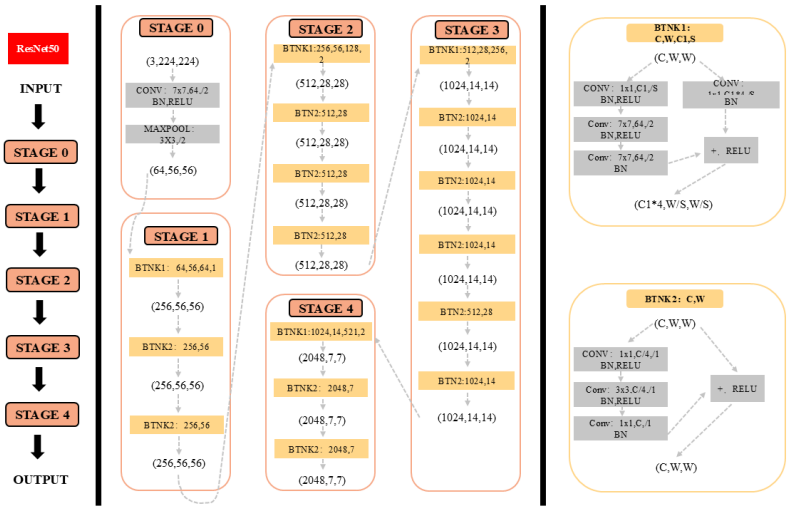

Fig. 3. ResNet50 module structure diagram.

ResNet50 is composed of multiple residual blocks stacked together, containing a total of 50 convolutional layers and 5 stages. Stage 0 is a 7×7 convolutional layer with 64 filters and a stride of 2, followed by a 3×3 max pooling layer, which is responsible for the initial feature extraction and reduction of the feature map size. The remaining four stages each contain multiple residual blocks. Each residual block consists of three convolutional layers, with the first and third convolutional layers having a kernel size of 1×1, and the second convolutional layer having a kernel size of 3×3. This design reduces the computational complexity of the network while maintaining a high level of feature representation capability. Stage 1 contains 4 residual blocks (input channels: 64, output channels: 256), Stage 2 contains 4 residual blocks (input channels: 128, output channels: 512), Stage 3 contains 6 residual blocks (input channels: 256, output channels: 1024), and Stage 4 contains 3 residual blocks (input channels: 512, output channels: 2048).

After feature extraction, the output layer outputs tensors with shapes (num-slices, 224,224,3). The TimeDistributed wrapper is applied to the GlobalAveragePooling2D layer to obtain the global feature vector for each CT slice. The processed output layer outputs tensors with shapes(num-slices, 2048), representing the deep features of each CT slice. Finally, an LSTM layer is applied to the pooled feature vectors to capture the temporal information of image features across the slice sequence. The internal structure of the LSTM is shown as **Fig. 4**.

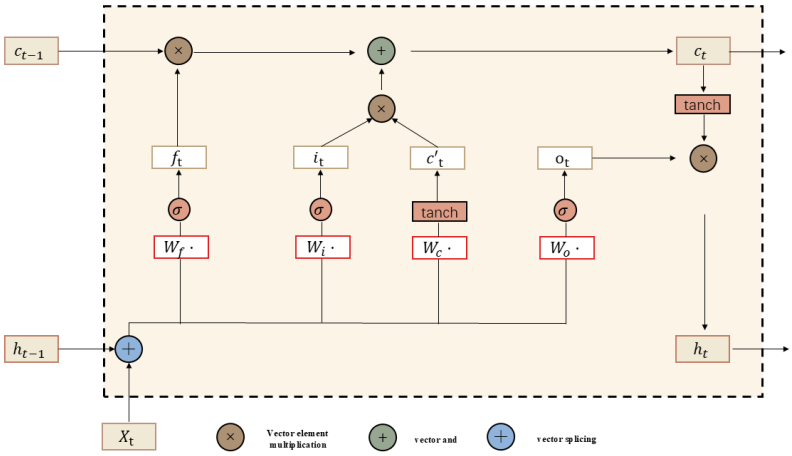

Fig. 4. LSTM module structure diagram

In the figure, f_t represents the forget gate, i_t represents the input gate, and o_t represents the output gate. c' is the candidate cell state. c_{t-1} represents the memory information from the previous time step, c_t represents the current memory state, h_t is the output of the LSTM unit, and h_{t-1} is the output from the previous time step. The formulas for LSTM are as follows:

$$f_t = \sigma(W_f[h_{t-1}, x_t] + b_f) \tag{1}$$

$$c'_t = tanh\,(W_c[h_{t-1}, x_t] + b_f) \tag{1}$$

$$i_t = \sigma(W_i[h_{t-1}, x_t] + b_i) \tag{3}$$

$$c_t = f_t \odot c_{t-1} + i_t \odot c'_t \tag{4}$$

$$o_t = \sigma(W_o + b_o) \tag{5}$$

$$h_t = o_t \odot tanh\,(c_t) \tag{6}$$

In the equation, W_f, W_c, W_i, W_o are the weight matrices, b_f, b_f, b_i, b_o are the bias vectors, and $[h_{t-1}, x_t]$ is the vector that connects the previous hidden state h_{t-1} and the current input x_t. σ is the Sigmoid function, which compresses the output values into the range $[0,1]$The forget gate controls which information should be forgotten from the cell state, while the candidate memory cell state c'_t represents the information that can be added to the cell state at the current time step. The input gate controls which parts of the candidate memory cell state c'_t will be added to the current cell state c'_t . The forget and input gates are used to update the cell state. $f_t \odot c_{t-1}$ represents the information retained in the old cell state by the forget gate, while $i_t \odot c'_t$ represents the new information added to the cell state. The output gate determines the hidden state h_t at

the current time step, which is based on the current cell state and decides which parts will be output.

Clinical Data Extraction. Clinical data contains both discrete and continuous data. Discrete data, usually exist in the form of categorical variables. In this paper, One-Hot Encoding is used to preprocess such data. One-Hot Encoding maps each category to a unique binary vector, thus avoiding the influence of numerical magnitude relationships between categories on the model. For example, in gender data, males and females are coded as [1,0] and [0,1], respectively, and this coding allows the model to correctly handle categorical variables without introducing unwanted bias. Continuous data usually have a large range of values and are not uniformly distributed. In order for the model to better handle these data, Normalisation [27] is used in this paper. Normalisation scales the data into the interval [0,1] with the formula:

$$x_{norm} = \frac{x - min\,(x)}{max(x) - min\,(x)} \tag{7}$$

In this study, a deep neural network (DNN) is employed to extract features from the encoded clinical data. As **Fig. 5**, the network consists of an input layer, multiple hidden layers, and an output layer. The input layer contains 43 neurons, each corresponding to one laboratory feature. The three hidden layers comprise 128, 256, and 128 neurons, respectively, with each layer employing the ReLU activation function for non-linear transformation. The input layer contains 43 neurons, each corresponding to one laboratory feature. The three hidden layers comprise 128, 256, and 128 neurons, respectively, with each layer employing the ReLU activation function for non-linear transformation. The output layer consists of 128 neurons and generates a 128-dimensional vector, representing the high-dimensional feature representation of the clinical data, aligned in dimensionality with the CT image features. By projecting clinical and imaging features into the same dimensional space, the model is encouraged to learn similar representations across modalities, facilitating more effective understanding and processing of the heterogeneous data. This approach enhances the efficiency of multimodal data fusion and joint learning.

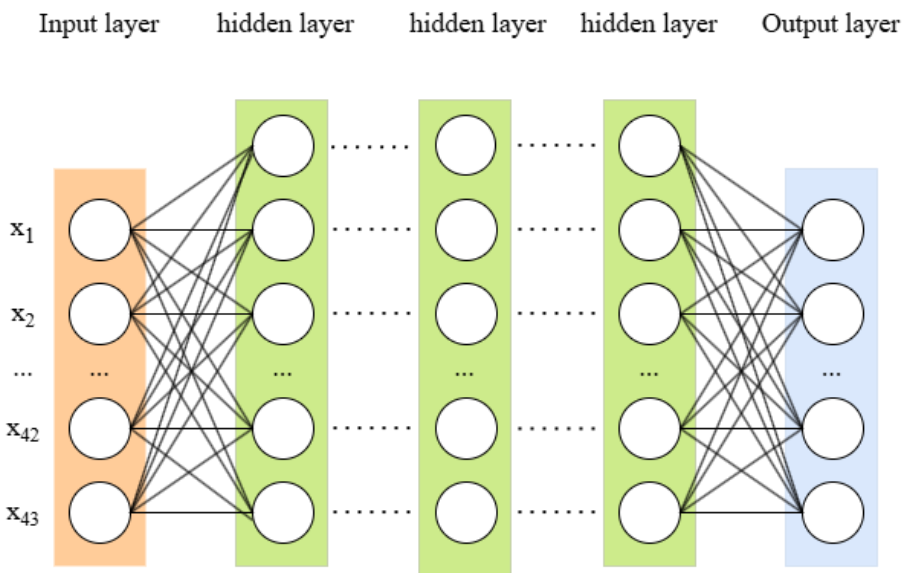

Fig. 5. DNN module structure diagram.

Multimodal Feature Fusion. In multimodal learning that combines medical image analysis and clinical data, effectively integrating features from different modalities is one of the key factors in improving model performance. Effectively integrating these heterogeneous features can leverage the complementarity of both to enhance the model's ability to comprehensively assess the patient's condition. Given that clinical data, as structured data, has strong heterogeneity compared to medical imaging data, we believe that the fusion process should focus more on the relationships between data with significant differences. Therefore, we introduce an asymmetric cross-attention mechanism [28] to capture the cross-modal information interaction between clinical data and imaging data, as shown as **Fig. 6**. The cross-modal cross-attention mechanism maps features from different modalities into Query, Key, and Value vectors through positional encoding.Then, by calculating the similarity between the Query and Key, the relationships between different positions are determined. Finally, by performing a weighted sum of the Value vectors, the representation vector for each position is obtained, which contains the attention information between different modalities.

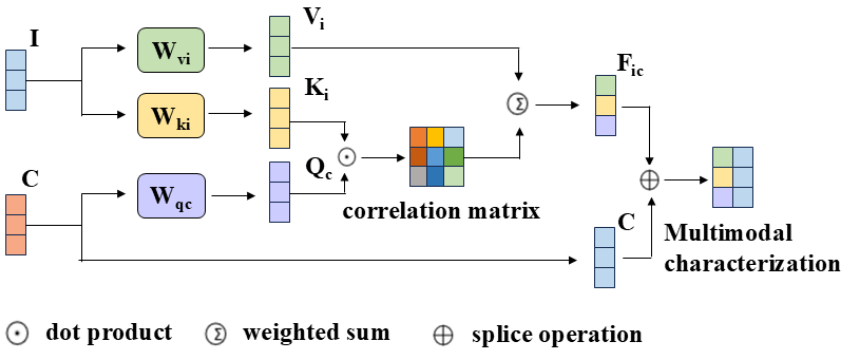

Fig. 6. Asymmetric cross-attention mechanis.

In the fusion of the clinical data modality and the CT image modality, clinical data features are used as Query vectors, as shown in Formula 8, while CT features are mapped into Key and Value vectors, as shown in Formula 9. For each Query vector, the dot product between it and all Key vectors is calculated, then divided by the square root of its dimension, and the Softmax function is applied to convert the result into a probability vector, obtaining the correlation matrix between features at different positions. Finally, the correlation matrix is used with the Value vectors to obtain the image representation processed by the attention mechanism, as shown in Formula 10.

$$Q_c = W_{qc}C \tag{8}$$

$$K_i = W_{ki}I,\ V_i = W_{vi}I \tag{9}$$

$$F_{ic} = CMA(Q_c, K_i, V_i) = Softmax(\frac{Q_c K_i^T}{\sqrt{d_k}})V_i \tag{10}$$

C represents the clinical data features output from the clinical data feature extraction module, I represents the CT image features, with both C and I having a feature dimension of 128. W is an iteratively optimized transformation matrix used for linear transformation of the features. Q is the query vector obtained by transforming the clinical data features, while the CT image features are transformed into key vectors K_i and V_i. F_{ic} represents the fused features of the CT image features and the clinical data features. The final output of the fusion module is as shown in Formula 11.

$$concat(F_{ic}, C) \tag{11}$$

This fusion method using asymmetric cross-attention allows the model to focus on learning the alignment relationship between medical images and clinical data, thus effectively understanding the association between the imaging blocks and clinical data. The fused features will be fed to the self-attention mechanism [29] module.

In the attention mechanism, the multimodal features are mapped into three vectors: Query, key and value. The similarity is calculated by taking the dot product of the query and the key, which generates the attention weights. The similarity is then normalized using the softmax function, and the value vector V is weighted and summed according to the attention weights to form the feature representation. The specific formula is as follows:

$$Q = XW_Q, K = XW_K, V = XW_V \tag{12}$$

$$Output = softmax(\frac{QK^T}{\sqrt{d_k}})V \tag{13}$$

W_Q, W_K, and W_V are learnable weight matrices, and d_k is the dimension of the key vector. Through these steps, the model can compute the attention weights of each element with respect to the other elements in the sequence, generate the output accordingly, and obtain the final fused features.

Classification model. The final fused features will be fed to the classification module. The classification module consists of a two-layer multi-layer perceptron(MLP) . Specifically, the deep multimodal features are first linearly transformed by a fully connected layer, followed by the introduction of nonlinearity using the ReLU activation function. Finally, the module outputs two classification results: effective control (CR+PR+SD) and ineffective control (PD).

3 Results

This section introduces the description ofthe evaluation metrics used. It then provides a detailed introduction to the hyperparameters and experimental setup used in the model. Finally, the section presents the experimental results.

3.1 Experimental Environment and Performance Indicators

The multimodal fusion model proposed in this paper utilizes the TensorFlow deep learning framework for parameter training and is implemented using Python programming. The hardware environment for this experiment includes the Ubuntu 18.04 operating system, an AMD 5800X CPU, an RTX 3090 GPU, and 64GB of RAM. Hyperparameters play a crucial role in determining the training effectiveness and speed of deep learning model. After several rounds of experimentation and adjustments, we selected the parameters listed as **Table 1** for the experiments conducted in this chapter.

Table 1. Hyperparameter setting.

Parameter name	Parameter value
learning_rate	1e-4
batch_size	32
dropout	0.1
optmizer	Adam

Based on the actual situation in the hospital, this study determines whether the patient should continue TACE treatment based on the disease control rate (CR+PR+SD). The post-TACE outcomes are categorized as either continuation or termination of TACE treatment. This study evaluates the performance of the predictive model through receiver operating characteristic (ROC) curve analysis, obtaining the area under the curve (AUC), precision , accuracy, and recall. The specific calculation methods are shown in equations (14) to (18):

$$FPR = \frac{FP}{FP+TN} \tag{14}$$

$$TPR = \frac{TP}{TP+FN} \tag{15}$$

$$Accuracy = \frac{TP+TN}{TP+TN+FP+FN} \tag{16}$$

$$Precision = \frac{TP}{TP+FP} \tag{17}$$

$$Recall = \frac{TP}{TP+FN} \tag{18}$$

The Area Under the Curve (AUC) is the area under the ROC curve and is used to measure the overall performance of the model. The AUC ranges from 0 to 1, with a higher AUC indicating better classification performance of the model.An AUC of 0.5 indicates that the model has no classification ability, which is equivalent to random guessing.

3.2 Performance

The dataset used in this experiment is derived from the dataset mentioned in Chapter 2. We performed experiments using 5-fold cross-validation (K=5) and took the average of the evaluation metrics across the k folds as the final evaluation metric. We had chosen to compare the multimodal fusion models of recent years with the multimodal model of this paper. The selected comparison models are briefly described below:

Mode1 [30]: J Venugopalan et al. stitched together the extracted imaging and non-imaging features and then directly used a random forest classifier for the classification task.

Mode2 [31]: J Zhang et al. used a cross-attention mechanism to fuse imaging and non-imaging features, the fused imaging and non-imaging features were downscaled, and the network was optimised using cross-entropy loss and modal alignment loss.

Mode3 [32]: H Chen et al. input the imaging features and non-imaging features successively at the same time through the channel attention and spatial attention for modal fusion, and the fused features are fed to the convolution module for further feature extraction, and finally sent to the classification.

Mode4 [33]: M Golovanevsky et al. pass three different modal features through a multi-head self-attention module, the self-attention processed features are then cross-attended two by two, and finally the fused features are spliced together for the classification task.

The AUC, Precision, Accuracy, and Recall for each model in the dataset are shown as **Table 2**.

Table 2. Comparative experiments.

Models	AUC	Accuracy	Precision	Recall
Mode1	0.70	0.73	0.75	0.74
Mode2	0.79	0.84	0.81	0.84
Mode3	0.78	0.79	0.83	0.85
Mode4	0.82	0.83	0.84	0.82
CLDCL(OURS)	0.85	0.86	0.88	0.87

Analysis of the tabular data shows that the CLDCA model achieved an AUC value of 0.85, an accuracy of 0.86, a Precision of 0.88, and a Recall of 0.87 on the dataset, surpassing all the comparison models. This indicates that the CLDCA model has superior performance in predicting post-TACE outcomes for liver cancer patients.

To demonstrate the effectiveness of each substructure of the model, this section compared the performance of the proposed model with three ablation models on the same dataset for the classification task.The ablation results are shown as **Table 3**:

Mode5: The proposed model removed the cross-attention module and directly concatenated the imaging features and non-imaging features before inputting them into the self-attention mechanism module.

Mode6: The proposed model concatenated the features output by the asymmetric cross-attention fusion module, masked the self-attention mechanism module, and directly performed the classification task.

Mode7: The proposed model masked the LSTM module, while keeping the other modules unchanged.

Table 3. Ablation experiments.

Models	AUC	Accuracy	Precision	Recall
CLDCL(OURS)	0.85	0.86	0.88	0.87
Mode5	0.75	0.74	0.76	0.73
Mode6	0.80	0.82	0.79	0.81
Mode7	0.77	0.80	0.75	0.79

The ablation experiment results showed that after masking the cross-attention mechanism, the AUC value decreased by 0.1, indicating that the asymmetric cross-attention module helped the model better understand the relationship between different modalities, thus improving classification performance. After masking the self-attention mechanism, the AUC decreased by 0.05, indicating that deep feature extraction through the self-attention mechanism is necessary for the fused features. After masking the LSTM module, the AUC decreased by 0.08, suggesting that LSTM effectively captured the temporal relationships between CT slices and integrated information to extract global features.

Table 4. Comparison experiment between multimodal and unimodal.

Datasets	AUC	Accuracy	Precision	Recall
CT+clinic	0.85	0.86	0.88	0.87
Only CT	0.70	0.72	0.69	0.68
Only clinic	0.65	0.68	0.69	0.71

By comparing the performance of multimodal and unimodal prediction models on the dataset, the significant advantage of the multimodal approach in this prediction task was validated. The experimental results show that the multimodal model outperforms the unimodal model in terms of AUC, precision, accuracy, and recall.

4 Discussion

In the experimental evaluation, the CLDCA model demonstrated significant superiority over traditional unimodal models, especially excelling in key metrics such as AUC and accuracy. This advantage arises from the multimodal model's ability to simultaneously integrate various perceptual modalities (e.g., images and text), thereby obtaining information from different perspectives. This multimodal integration not only enriches the dimensions of the data but also provides a more comprehensive and accurate representation of information, significantly improving the overall performance of the model. Through ablation experiments, we validated the contribution of each component in the CLDCA model. The multimodal fusion module, self-attention mechanism module, and LSTM module all significantly contribute to the model's overall performance. These modules work in synergy, resulting in the CLDCA model excelling in key metrics such as AUC and accuracy. Therefore, each component is an indispensable part of the CLDCA model, providing strong support for the model's high performance.

Moreover, compared to the contrast models, the attention mechanism can effectively improve the model's classification performance.Model 1 proposed by J. Venugopalan et al. only performs a simple concatenation of features from different modalities, which is insufficient to capture the underlying relationships between modalities. Models 3 and 4 by MADDi and H. Chen, as well as the CLDCA model in this paper, use the attention mechanism to process features from different modalities. Compared to the results of Model 1, the AUC was improved by 0.08, 0.12, and 0.15, respectively, indicating that the use of attention mechanisms for modality fusion helps the model understand the relationships between multimodal data.J. Zhang et al. also recognized the strong heterogeneity between imaging and non-imaging features. In Model 2, they first concatenate the imaging features and then apply symmetric cross-attention processing between the concatenated imaging features and non-imaging features. When we replaced the asymmetric cross-attention in this paper's model with the same approach, the classification performance decreased. This demonstrates that the asymmetric cross-attention fusion mechanism performs better on this dataset. Through the asymmetric cross-attention mechanism, the model can flexibly allocate attention weights, thus better capturing complementary information between different modalities, which enhances prediction ability.

Future work will focus on two areas. First, more imaging and clinical data will be collected to further improve the model's robustness and generalizability. Integrating multicenter data can effectively reduce the model's dependence on specific datasets and enhance its applicability in different clinical environments. Second, data augmentation techniques will be considered to expand the training set, thereby improving the model's performance in small sample scenarios and reducing the occurrence of overfitting.

5 Conclusions

This study developed a multimodal data fusion-based model for predicting TACE treatment response, combining CT imaging features with clinical data, significantly improving the prediction accuracy of postoperative treatment response in liver cancer patients. The introduction of the cross-attention mechanism, by focusing on the interaction between the two features, helps the model extract visual information from CT images while incorporating the correlations from clinical data, thus achieving a more comprehensive disease representation and enhancing the depth and precision of the analysis. The use of the self-attention mechanism enables the model to dynamically adjust the weights of different modalities based on the importance of the features, which not only increases the model's attention to key features but also effectively suppresses the impact of noise, preventing redundant information from interfering with the prediction results. Through multimodal fusion, the model demonstrates greater flexibility and adaptability in handling complex data, thereby enhancing the accuracy of the predictions. The model performs exceptionally well on the dataset with AUC, precision, accuracy and recall of 0.85, 0.86, 0.88 and 0.87 respectively.

Acknowledgments. This study is supported by the Key Research and Development Program of Zhejiang Province (2025C01135).

Disclosure of Interests. The authors have no competing interests to declare that are relevant to the content of this article.

References

1. Siegel, R.L., Miller, K.D., Wagle, N.S., Jemal, A.: Cancer statistics, 2023. CA: a cancer journal for clinicians **73**(1), 17–48 (2023).
2. Gbolahan, O.B., Schacht, M.A., Beckley, E.W., LaRoche, T.P., O'Neil, B.H., Pyko,M.: Locoregional and systemic therapy for hepatocellular carcinoma. Journal of gastrointestinal oncology **8**(2), 215 (2017).
3. Zheng, R., Chen, R., Han, B., Wang, S., Li, L., Sun, K., Zeng, H., Wei, W., He, J.:Cancer incidence and mortality in china, 2022. Zhonghua zhong liu za zhi [Chinese journal of oncology] **46**, 221–231 (2024).
4. Xin, H., Q, X.: Management of adverse reactions to systemic antineoplastic therapy in advanced hepatocellular liver cancer. Lin chuang nei ke za zhi **39**(12), 806–810(2022).
5. Guidelines for the diagnosis and treatment of primary liver cancer2022. Zhe jiang shi yong yi xue **27**(06), 528–536 (2022). https://doi.org/10.16794/j.cnki.cn331207/r.2022.06.014

6. Galle, P.R., Forner, A., Llovet, J.M., Mazzaferro, V., Piscaglia, F., Raoul, J.L.,Schirmacher, P., Vilgrain, V.: Easl clinical practice guidelines: management of hepatocellular carcinoma. Journal of hepatology **69**(1), 182–236 (2018)
7. Balogh, J., Victor III, D., Asham, E.H., Burroughs, S.G., Boktour, M., Saharia,A., Li, X., Ghobrial, R.M., Monsour Jr, H.P.: Hepatocellular carcinoma: a review.Journal of hepatocellular carcinoma pp. 41–53 (2016)
8. Bi, W.L., Hosny, A., Schabath, M.B., Giger, M.L., Birkbak, N.J., Mehrtash, A.,Allison, T., Arnaout, O., Abbosh, C., Dunn, I.F., et al.: Artificial intelligence in cancer imaging: clinical challenges and applications. CA: a cancer journal for clinicians **69**(2), 127–157 (2019)
9. Bera, K., Braman, N., Gupta, A., Velcheti, V., Madabhushi, A.: Predicting cancer outcomes with radiomics and artificial intelligence in radiology. Nature reviews Clinical oncology **19**(2), 132–146 (2022)
10. Nam, D., Chapiro, J., Paradis, V., Seraphin, T.P., Kather, J.N.: Artificial intelligence in liver diseases: Improving diagnostics, prognostics and response prediction. Jhep Reports **4**(4), 100443 (2022)
11. Kim, J., Choi, S.J., Lee, S.H., Lee, H.Y., Park, H.: Predicting survival using pretreatment ct for patients with hepatocellular carcinoma treated with transarterial chemoembolization: comparison of models using radiomics. American Journal of Roentgenology pp. 1026–1034 (2018)
12. Morshid, A., Elsayes, K.M., Khalaf, A.M., Elmohr, M.M., Yu, J., Kaseb, A.O.,Hassan, M., Mahvash, A., Wang, Z., Hazle, J.D., et al.: A machine learning model to predict hepatocellular carcinoma response to transcatheter arterial chemoembolization. Radiology: Artificial Intelligence **1**(5), e180021 (2019)
13. Guo, W., Wang, J., Wang, S.: Deep multimodal representation learning: A survey. Ieee Access **7**, 63373–63394 (2019)
14. Saillard, C., Schmauch, B., Laifa, O., Moarii, M., Toldo, S., Zaslavskiy, M., Pronier, E., Laurent, A., Amaddeo, G., Regnault, H., et al.: Predicting survival after hepatocellular carcinoma resection using deep learning on histological slides. Hepatology **72**(6), 2000–2013 (2020)
15. Liu, Q.P., Xu, X., Zhu, F.P., Zhang, Y.D., Liu, X.S.: Prediction of prognostic risk factors in hepatocellular carcinoma with transarterial chemoembolization using multi-modal multi-task deep learning. EClinicalMedicine **23** (2020)
16. Han, G., Berhane, S., Toyoda, H., Bettinger, D., Elshaarawy, O., Chan, A.W., Kirstein, M., Mosconi, C., Hucke, F., Palmer, D., et al.: Prediction of survival among patients receiving transarterial chemoembolization for hepatocellular carcinoma: a response-based approach. Hepatology **72**(1), 198–212 (2020)
17. Peng, J., Huang, J., Huang, G., Zhang, J.: Predicting the initial treatment response to transarterial chemoembolization in intermediate-stage hepatocellular carcinoma by the integration of radiomics and deep learning. Frontiers in Oncology **11**, 730282 (2021)
18. Wang, Q., Liu, Q., Su, M.: Application analysis of a deep learning-based recurrence prediction model after transarterial chemoembolisation for hepatocellular carcinoma. Yi xue yin xiang xue za zhi **34**(10), 71–74 (2024)
19. Moawad, A.W., Morshid, A., Khalaf, A.M., Elmohr, M.M., Hazle, J.D., Fuentes, D., Badawy, M., Kaseb, A.O., Hassan, M., Mahvash, A., et al.: Multimodality annotated hepatocellular carcinoma data set including pre-and post-tace with imaging segmentation. Scientific data **10**(1), 33 (2023)
20. Chang, S.C., Liu, C.F., Misztal, M., Hsieh, Y.H., Lin, C.: A multi-phase deep learning approach for predicting hcc response to tace using complete computed tomography. In: 2023

IEEE EMBS Special Topic Conference on Data Science and Engineering in Healthcare, Medicine and Biology. pp. 99–100. IEEE (2023)
21. Sun, Z., Li, X., Liang, H., Shi, Z., Ren, H.: A deep learning model combining multimodal factors to predict the overall survival of transarterial chemoembolization.Journal of Hepatocellular Carcinoma pp. 385–397 (2024)
22. Zhou, H.Y., Yu, Y., Wang, C., Zhang, S., Gao, Y., Pan, J., Shao, J., Lu, G., Zhang, K., Li, W.: A transformer-based representation-learning model with unified processing of multimodal input for clinical diagnostics. Nature Biomedical Engineering **7**(6), 743–755 (2023)
23. Pino, C., Vecchio, G., Fronda, M., Calandri, M., Aldinucci, M., Spampinato, C.:Twinliver-net: predicting tace treatment outcome from ct scans for hepatocellular carcinoma using deep capsule networks. In: 2021 43rd Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC). pp.3039–3043. IEEE (2021)
24. Xi, Z., Shao, G.: Progress of mri imaging-based histology for predicting efficacy after transcatheter arterial chemoembolisation for hepatocellular carcinoma. Jie rufang she xue za zhi **33**(01), 90–94 (2024)
25. Schwartz, L.H., Litière, S., De Vries, E., Ford, R., Gwyther, S., Mandrekar, S., Shankar, L., Bogaerts, J., Chen, A., Dancey, J., et al.: Recist 1.1—update and clarification: From the recist committee. European journal of cancer **62**, 132–137(2016)
26. Jiang, B., Lu, D., Dai, J., Li, K., Du, Q., Xie, B., Xie, J., Zhu, X., Xie, X.: A simple prognostic scoring system for hepatocellular carcinoma treated with debtace. Journal of Hepatocellular Carcinoma pp. 1403–1414 (2024)
27. Jeong, S., Lee, M., Yoo, S.: Machine learning-based stroke risk prediction using public big data. Journal of Advanced Navigation Technology **25**(1), 96–101 (2021)
28. Li, H., Wu, X.J.: Crossfuse: A novel cross attention mechanism based infrared and visible image fusion approach. Information Fusion **103**, 102147 (2024)
29. Pan, S., Yang, B., Wang, S., Guo, Z., Wang, L., Liu, J., Wu, S.: Oil well production prediction based on cnn-lstm model with self-attention mechanism. Energy **284**,128701 (2023)
30. Venugopalan, J., Tong, L., Hassanzadeh, H.R., Wang, M.D.: Multimodal deep learning models for early detection of alzheimer's disease stage. Scientific reports **11**(1), 3254 (2021)
31. Zhang, J., He, X., Liu, Y., Cai, Q., Chen, H., Qing, L.: Multi-modal cross-attention network for alzheimer's disease diagnosis with multi-modality data. Computers in Biology and Medicine **162**, 107050 (2023)
32. Chen, H., Guo, H., Xing, L., Chen, D., Yuan, T., Zhang, Y., Zhang, X.: Multimodal predictive classification of alzheimer's disease based on attention-combined fusion network: Integrated neuroimaging modalities and medical examination data. IET Image Processing **17**(11), 3153–3164 (2023)
33. Golovanevsky, M., Eickhoff, C., Singh, R.: Multimodal attention-based deep learning for alzheimer's disease diagnosis. Journal of the American Medical Informatics Association **29**(12), 2014–2022 (2022)