Biomedical Signal Processing and Control 119 (2026) 109858

Contents lists available at ScienceDirect

Biomedical Signal Processing and Control

journal homepage: www.elsevier.com/locate/bspc

Lung tumor growth prediction of follow-up via Spatio-Temporal Convolutional Transformer

Ning Xiao [a,*], Aoyu Li [b], Yan Qiang [c], Juanjuan Zhao [b], Yan Geng [d], Leqin Li [e]

[a] College of Information, Shanxi University of Finance and Economics, Taiyuan, 030006, China
[b] College of Software, Taiyuan University of Technology, Taiyuan, 030024, China
[c] College of Software, North University of China, Taiyuan, Taiyuan, 030051, China
[d] Clinical Laboratory, Shanxi Provincial People's Hospital, Taiyuan, 030000, China
[e] Department of Computer Science, State University of New York, NY, 10012, USA

ARTICLE INFO

Keywords:
Tumor growth prediction
Convolutional Transformer
Longitudinal study
Lung cancer

ABSTRACT

Precise forecasting of lung tumor growth is essential for devising effective treatment strategies and managing early-stage lung cancer. However, existing methods lack intuitive judgments for tumor growth and evolution patterns, as well as clinical information. To address this problem, this paper introduces a versatile and easy-to-train architecture called the Spatio-Temporal Convolutional Transformer (ST-ConvTransformer) to facilitate the prediction of early-stage tumor growth. The ST-ConvTransformer is composed of two transformer submodules that utilize the same self-attention layers. During the extraction of visual features from medical images, textual features from clinical information are also extracted and integrated. Simultaneously, the visual encoder submodule incorporates the temporal and spatial information of tumors. Extensive experiments are conducted on a dataset with 2800 clinical subjects. The proposed model achieves a Precision of 80.87%, Recall of 89.74%, and Dice Similarity Coefficient of 84.24%. These results demonstrate the potential of the ST-ConvTransformer as a reliable clinical-aided tool for predicting lung tumor growth.

1. Introduction

The precise prediction of future tumor growth trends and structural changes, such as maximum diameter and edge information, holds paramount importance for cancer screening and the development of effective anticancer therapy strategies [1]. In clinical practice, lung tumors exhibit characteristics of unrestrained and continuous proliferation, infiltrating and disrupting surrounding tissues while manifesting exogenous growth. Different types of tumors exhibit distinct growth patterns, as shown in Fig. 1, and the clinical treatment methods for these tumors also vary accordingly. In the study [2], researchers propose that precise tumor growth prediction plays a crucial role in guiding appropriate treatment management and surgical planning. Evaluating the aggressiveness of these tumors at an early stage is essential to ensure that therapeutic toxicity remains within the required limits, thereby minimizing adverse effects on patients. Consequently, there is a need to develop a reliable and patient-specific method for predicting tumor growth, taking into consideration the complexity, heterogeneity, and dynamics of tumors [3].

In recent years, a growing body of research has focused on predicting tumor growth, including mathematical models [4], finite element analysis methods [5], cellular automata [6], diffusion reaction equations, and neural networks. Tumor prediction is crucial for the early development of appropriate treatment. In response to this, Luian et al. [7] proposed a diffusion-reaction coupling system. This system first used registered and segmented images to train the parameters in the coupled tumor growth model, and then used the trained parameters for kidney tumor prediction. Zhang et al. [8] developed a model for predicting pancreatic tumor growth from longitudinal patient data using a convolutional invasion and dilation network fusion model, which can effectively capture and learn cellular invasion dynamics and mass effects in tumor growth prediction. For glioblastoma tumors, which grow expansively, Pei et al. [9] used features such as image intensities, super-pixel gradients, and grayscale histograms to extract tumor features, and then applied a joint label fusion mathematical algorithm to model glioblastoma cell growth. In the study of lung tumor growth prediction, Ghita et al. [10] used exponential, logistic, and Gompertz models to simulate tumor dynamics. With limited data, they developed a growth prediction model for lung cancer and a personalized prediction model. Yonn et al. [11] determined the tumor doubling time of lung adenocarcinoma by extracting edge-related radiomic features of the tumor in CT scans. In addition to these methods, another approach

* Corresponding author.
E-mail address: x20221016@sxufe.edu.cn (N. Xiao).

https://doi.org/10.1016/j.bspc.2026.109858
Received 16 June 2025; Received in revised form 25 January 2026; Accepted 13 February 2026
Available online 17 February 2026

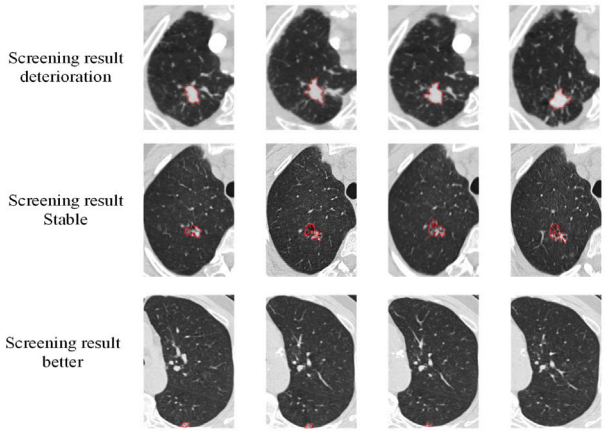

Fig. 1. Examples of changes in pulmonary nodules in different stages. The pulmonary nodules of three patients grow in different forms.

involves predicting tumor growth in the next period by calculating the optical flow change between tumor pixels in images taken at adjacent time points [12]. However, these methods oversimplify tumor growth patterns and simulate tumor growth in a simple linear fashion, whereas most tumors grow in a nonlinear manner.

Data-driven deep learning presents a promising solution for extracting features from large-scale tumor data. Importantly, it integrates factors that may influence tumor growth as conditions into personalized models of tumor growth. This approach aims to characterize two fundamental processes: cell invasion and the mass effect of tumor growth. To predict the glioblastoma multiform tumor growth, Kamli et al. [13] designed a tumor growth predictor using an end-to-end convolutional neural network architecture. This model was trained on a public dataset from the Cancer Imaging Archive (TCIA) and further enhanced by incorporating generated synthetic data, thereby broadening the training base and potentially improving predictive performance. Zhang et al. [14] integrated 3D spatial and temporal properties from images with clinical information as input for convolutional long short-term memory units. The network's prediction results were effective in forecasting the image properties of tumors, including cell density and relevant diagnostic information. However, in longitudinal studies of tumors, the challenge of insufficient longitudinal tumor data collection and the extraction of nonlinear representations remains a pressing issue [8]. Addressing this challenge is essential for the practical application of deep learning in tumor growth prediction [15,16].

To bridge these gaps, this work proposes a comprehensive end-to-end tumor prediction model called the Spatio-Temporal ConvTransformer (ST-ConvTransformer). Our underlying assumption is that tumor features at different stages reside on the same high-dimensional manifold, and these features can evolve along a specific direction on the manifold. The image encoder is employed initially to extract local feature maps of the tumor. Simultaneously, a temporal encoder and a spatial encoder were designed to extract image features from follow-up tumor data at multiple time points and the spatial location features of the tumor in the image. Concerning patient clinical information, a text transformer serves the dual purpose of functioning as both a text encoder and a text decoder. Finally, Image-Text Contrastive Learning is applied to integrate all features for tumor growth prediction. This approach aligns visual and textual representations by drawing matching image–text pairs closer in the embedding space while pushing apart mismatched pairs. The loss function primarily combines pixel-level mean square error loss and total variation regularization as the objective function.

The key advantages of ST-ConvTransformer include: (1) ST-Conv Transformer effectively harnesses both image models and language models. To simulate the nonlinear growth mode of tumors, we also incorporated certain clinical information, such as the patient's smoking history, drinking history, and occupation, in tumor growth prediction. (2) To fully exploit the information from tumors at different time points and spatial locations, ST-ConvTransformer incorporates a temporal feature extractor and a spatial feature extractor. This addition enables the comprehensive expression of the continuous temporal and spatial characteristics of tumors. (3) Due to the utilization of unimodal models and a lightweight ConvTransformer, ST-ConvTransformer is more computationally efficient than existing state-of-the-art methods. Furthermore, the inclusion of a total variation regularization term in the objective function effectively mitigates the generation of artifacts.

2. Methods

All procedures in this study were conducted in strict compliance with the guidelines established by the Medical Ethics Committee of Shanxi Provincial People's Hospital, Shanxi University of Finance and Economics. All procedures described in this study were conducted in strict adherence to the relevant guidelines and regulations. The human experiment was conducted in accordance with the protocol approved by the Ethics Committee of Shanxi Provincial People's Hospital (No. 2022-321 v1.0). Prior to conducting imaging screening, all participants or their families signed informed consent forms under the guidance of doctors.

2.1. Overview of proposed method

In this study, we designed the ST-ConvTransformer model to integrate longitudinal CT scans with clinical information for the analysis of tumor growth. The methodological framework consists of four main steps: (1) A feature embedding module performs feature mapping on CT slices from multiple time points. The resulting feature maps are combined with a location map to identify tumor regions. (2) The feature maps with location markers are fed into an encoder to capture long-range sequential dependencies across tumor slices from different time periods. Simultaneously, patient clinical information corresponding to the image is encoded by a text transformer to obtain feature embeddings. (3) Image feature embeddings and clinical feature embeddings are fused using a contrastive learning strategy, which aligns multimodal representations and integrates information from different sources. (4) Finally, the decoder output is projected back to the image space through a stack of convolutional layers, producing the predicted tumor image, as illustrated in Fig. 2.

2.2. Data preprocessing

Image registration is an essential preprocessing step for ensuring the spatial alignment of tumor regions across different time points. It involves aligning the tumor regions from longitudinal CT scans into a standard coordinate system, thus facilitating the accurate extraction of temporal features. This process is the vital for longitudinal tumor analysis, as it minimizes the effect of misalignment due to variations in patient positioning or scanner settings. To address this challenge, this paper initially employs a registration method [17] to align the original CT images. The primary focus is to use a single CT image containing tumors at various time points as the subject of study.

Cropping is applied to focus the model's attention on the tumor region by removing irrelevant background areas. When utilizing spatial and temporal encoders for tumor prediction, it is essential to minimize interference from surrounding lung parenchyma and other background tissues in the original CT image, which may affect the accuracy of tumor analysis. To achieve this, the study incorporates physician-provided annotation information, which labels the tumor regions in the CT images. The regions marked as tumors are then extracted as the input data, with a fixed size of 56×56 pixels, ensuring

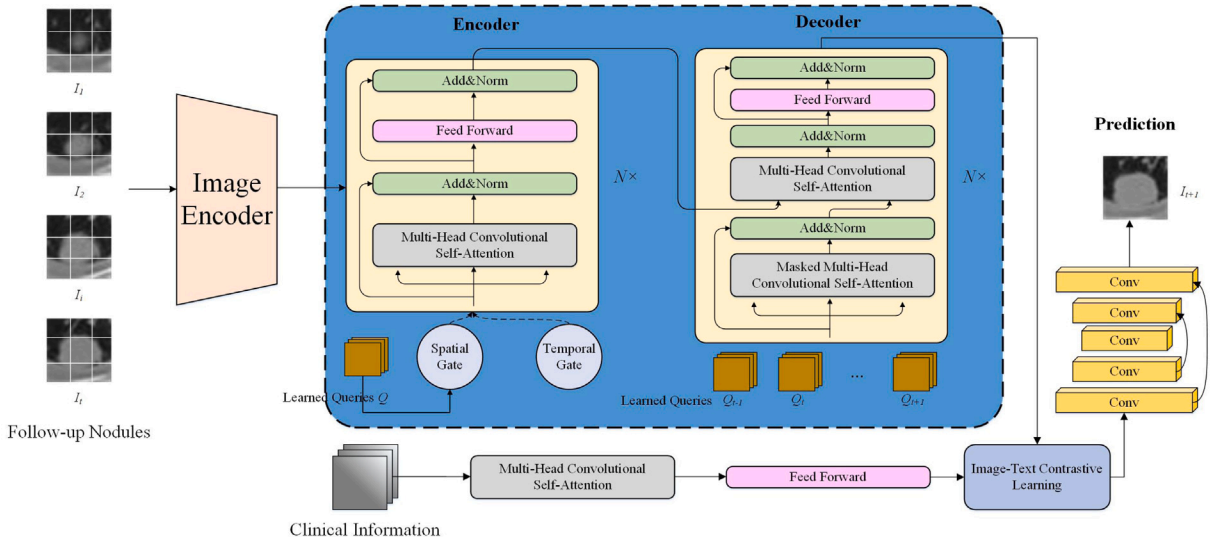

Fig. 2. An overview of the proposed ConvTransformer architecture. The ConvTransformer consists of Feature Embedding, Encoder, Decoder, and Prediction Network.

consistency in the input for subsequent processing and prediction. This approach allows the model to focus specifically on the tumor areas, thereby enhancing the accuracy of predictions while reducing the influence of irrelevant background features.

2.3. Feature embedding and positional encoding

To extract robust feature representations for efficient subsequent learning, tumor features are processed using a 4-layer convolutional network with a Leaky ReLU activation function and a specified hidden dimension of d_{model}. This design ensures that the features are effectively learned while retaining important spatial information. Given a sequence of follow-up tumor images $I_i (i \in [i, n])$ each image is first processed by a shared image encoder to extract low-level spatial features, the embedded feature map f_i serve as the input tokens to the subsequent encoder–decoder architecture, the following equation as shown:

$$f_i = F(I_i), i \in [i, n] \tag{1}$$

To enable the model to capture the order of tumor sequence slices effectively, this paper introduces "positional encoding" at each layer before the encoder and decoder. The positional encoding is designed to have the same dimensionality as the feature map of each slice, allowing for direct element-wise addition with the extracted features. This integration enables the model to recognize and retain the sequential relationships between tumor slices, which is crucial for understanding the tumor's progression and evolution over time. By embedding positional information, the model is better equipped to learn the temporal dynamics inherent in the tumor growth process. In this study, positional encoding is implemented using sine and cosine functions at different frequencies to create distinct positional encodings for each slice position within the sequential imaging data.

$$\begin{aligned} Pos_{p_{(i,j)}(2k)} &= sin\left(\frac{n}{10000^{\frac{2k}{d_{model}}}}\right) \\ Pos_{p_{(i,j)}(2k+1)} &= cos\left(\frac{n}{10000^{\frac{2k}{d_{model}}}}\right) \end{aligned} \tag{2}$$

Where pos is the position marker, (i, j) represents the spatial position of the feature, and the channel dimension is denoted as $2k$. That is, each dimension of the positional encoding corresponds to a sinusoid. The wavelengths form a geometric progression from 2π to $10000 * 2\pi$.

Given embedded feature maps, the feature with location representation can be expressed as the following formula:

$$Q_i = f_i \oplus Pos_{p_{(i,j)}} \tag{3}$$

where $\oplus$ operation represents element-wise addition.

2.4. Encoder and decoder

2.4.1. Encoder

Given the tumor at different time points $I_1, I_2, \ldots, I_t$, the objective is to predict I_{t+1}, the output at the next time(usually time 3). As illustrated in Fig. 2, applying the ConvTransformer directly to the temporal domain enables the exploration of the two-dimensional change process of tumors, facilitating more accurate growth prediction. By leveraging the temporal sequence of tumor images, this approach captures the tumor's evolution over time. Additionally, the spatial consistency between adjacent time points is utilized to establish a correspondence between the image from the previous period and the current image. This relationship enables the calculation of the optical flow of the tumor, facilitating the extraction of precise tumor position information. By tracking the movement and changes in the tumor's location, this method improves the model's ability to predict tumor growth with greater spatial and temporal accuracy.

The encoder consists of N stacked ST-ConvTransformer blocks. Each block follows a residual structure composed of multi-head convolutional self-attention, feed-forward layers, and Add& Norm operations. Specifically, the embedded image features are first fed into a multi-head convolutional self-attention module, which replaces standard linear projections with convolutional kernels to explicitly preserve local spatial structures. This design allows the model to capture tumor morphology, boundary details, and local contextual patterns more effectively than vanilla self-attention.

To further disentangle spatial and temporal dynamics,two gating mechanisms are introduced: a spatial gate f_s and a temporal gate f_t. The temporal gate is employed to extract image features from follow-up tumor data at multiple time points, while the spatial gate is primarily used to extract the location characteristics of the tumor in CT images. The temporal gate is constructed with a stack of two cascaded residual blocks. Each block contains 3D convolution layers operating on spatial dimensions of feature, followed by temporal convolution along the time axis. In contrast, the spatial gate primarily consists of FlowNet [18,19], which is a CNN architecture that directly predicts dense optical flow

from two input frames. FlowNet consists of three stages: (i) convolutional layers for dimensionality reduction and hierarchical feature extraction from each input image, (ii) a correlation layer that computes dense correspondence via element-wise multiplication and summation across spatial locations, and (iii) an upsampling module that restores spatial resolution.

2.4.2. Text transformer

The text transformer used for clinical information encoding is implemented as a lightweight transformer encoder. Clinical variables (e.g.smoking situation, working, family cancer) are first mapped to discrete tokens using predefined vocabularies. Each category is represented by a learnable embedding vector of dimension d. All variable embeddings are then concatenated to form a clinical token sequence, optionally augmented with a learnable token to capture global clinical context. The resulting clinical embeddings are processed by the text transformer to model interactions among different clinical factors. The transformer output is subsequently aligned with imaging features through the image–text contrastive learning module, encouraging semantic consistency between clinical context and imaging-derived tumor representations.

Subsequently, a set of learnable query embeddings is generated as input to the text Transformer, and these embeddings are inserted at every other transformation block. Within these blocks, the queries first interact through self-attention layers, which enhance their contextual understanding by capturing intricate relationships among themselves. Simultaneously, they engage with corresponding image features via cross-attention layers, enabling a seamless fusion of textual and visual data that is critical for effective multimodal feature integration.

Specifically, the text transformer consists of L stacked transformer encoder layers (in our implementation, $L = 2$), each composed of multi-head self-attention followed by a position-wise feed-forward network with residual connections and layer normalization. The attention module employs H attention heads ($H = 4$), and the embedding dimension is set to ($d = 256$).

2.4.3. Decoder

After the tumor image passes through the encoder, this paper combines the time-series feature map of the encoded tumor, the spatial feature map, and the position identification to generate the fused feature map. The fused feature map is subsequently fed into the decoder for further processing. The decoder mirrors the encoder with N stacked blocks but introduces masked multi-head convolutional self-attention to ensure causal temporal modeling. At each decoding step, the model only attends to previous time points, preventing information leakage from future scans.

Learned temporal queries $Q_{t-1}, Q_t, \ldots, Q_{t+1}$ are used to guide the decoding process, enabling the model to align historical tumor states with the prediction target. The decoder integrates encoder outputs through attention mechanisms and refines the representations via feed-forward layers and residual normalization.

Self-attention is utilized in both the encoder and decoder, as it proves highly effective for handling image data. This mechanism enables the model to concentrate on the most relevant regions of the input, enhancing feature extraction and representation. In self-attention, the query Q, key K, and value V all come from the same set of input features. The attention formula, which is central to this mechanism, is given by:

$$Attention(Q, K, V) = softmax(\frac{QK^T}{\sqrt{d}})V \tag{4}$$

where d represents the dimension of query, K^T represents the transposition of K. This attention mechanism improves the model's ability to capture both local and global relationships within the tumor's spatial and temporal features.

2.4.4. Prediction

After the encoding and decoding calculations, it is necessary to consider that the previous series of operations may not fit the complex process well. To address this, the model's predictive capability is further enhanced by introducing a Fully Convolutional Network (FCN). Specifically, during the prediction phase, the fully convolutional network is constructed based on a 5-layer U-Net architecture. It consists of an encoder–decoder architecture with skip connections, allowing the model to retain fine-grained spatial information while processing the data through multiple layers. By leveraging the U-Net, the model becomes more adept at learning complex patterns, thereby improving the overall performance of tumor prediction tasks.

2.5. Image-Text Matching and fusion

Image-Text Matching aims to learn the alignment between medical images and clinical information feature embeddings. This study utilizes bidirectional self-attention, enabling all queries and text to interact with each other using masks. Bidirectional self-attention means that in the Transformer's self-attention mechanism, every token can attend to all other tokens in the sequence (both left and right). The resulting output query embeddings capture multimodal information. Each output query embedding is then fed into two linear classifiers to obtain logits, and the logits of all queries are aggregated to form the output matching score [20].

In Image-Text data, the text corresponds to the clinical information associated with each image, and each matched pair of images and text description is assigned a distinct label. During text-image fusion, image feature embeddings and text feature embeddings are projected into a shared embedding space via learned MLPs. During fine-tuning, the concatenated image-text embeddings are passed through a fully connected layer to predict tumor growth probability.

2.6. Training loss

In this study, the pixel-level mean square error (MSE) loss is chosen as the loss function for the ConvTransformer. Additionally, the paper incorporates the Total Variation (TV) term [21]. This term enables the image to retain its resolution without losing boundary information, preventing the generation of apparent staircase effects.

$$L = \frac{1}{N}\sum_{i=1}^{N}(I_i - Y_i)^2 + \int \sqrt{I_i^2 + Y_i^2}dxdy \tag{5}$$

where I_i represent the original tumor image and Y_i represent the predicted tumor image, (x, y) is pixel in images.

3. Experiment and result

3.1. Research objects and implementation details

Part of the experimental data in this paper is sourced from the NLST, while the other part is obtained from a cooperative hospital. This article categorizes the images of these lung cancer patients into up to five groups based on different disease stages. Therefore, this article uses a unique heat vector composed of five elements to represent the period of each lung cancer image during training. The final dataset comprises 2800 patients and 8400 images. This article utilizes cancer detection algorithms to crop and calibrate cancer regions, thereby enhancing the effectiveness of training.

The algorithm implementation is based on the PyTorch deep learning framework and the Python programming language. The training is conducted on an Intel Xeon system, with 64 GB of memory, and the GPU utilized is the NVIDIA GeForce GTX 3090 Ti.

Table 1
Comparison of image acquisition parameters between public and private datasets.

Parameter	Scanner manufacturer	Tube voltage (kVp)	Slice thickness (mm)	Radiation dose (mSv)
NLST	GE, Siemens, Philips, Toshiba	120	1.0–2.5	1–2
Cooperative hospital	GE scanner	120	1.0	1

3.1.1. NLST

The NLST dataset was collected by the American Institute of Radiological Imaging Network and the Lung Cancer Screening Research Group [22,23]. This study selected 2058 participants aged between 55 and 77 who underwent three consecutive annual follow-up CT scans. The male to female male-to-female ratio was 7:6. Cases with missing slices, poor image quality, or inconsistent metadata were excluded. All scans followed standardized acquisition protocols across multiple participating centres, ensuring high consistency in image quality and metadata. The effective radiation dose was also maintained between one mSv and two mSv. The directions of CT scanning include three types: axial, coronal, and sagittal.

3.1.2. Cooperative hospital

The data from the cooperative hospital spans from January 2016 to December 2019. Over the four years, the cohort comprised 742 participants, including 582 individuals under investigation for suspected lung cancer and 160 histologically confirmed lung cancer cases. The age range of these participants is between 56 and 78 years old, with a male-to-female ratio of approximately 5:3

For the cooperative hospital dataset, CT scans were obtained from GE scanner models under routine clinical conditions using spiral scanning. All experiments involving human subjects were conducted with prior approval from the relevant institutions of Shanxi Provincial People's Hospital, and informed consent was obtained from all participants. All methods described in this study were performed in accordance with the relevant guidelines and regulations. The human experiment was conducted in accordance with the human protocol approved by the Ethics Committee of Shanxi Provincial People's Hospital (2022-321 v1.0).

The NLST dataset comprises scans acquired from multiple institutions with heterogeneous protocols, while the private dataset was collected from our hospital using a single scanner with standardized acquisition settings. To better clarify the imaging differences between the public and private datasets, Table 1 summarize the key acquisition parameters, including scanner manufacturer, slice thickness, tube voltage, and radiation dose.

3.2. Evaluation criterion

Tumor progression involves temporal changes, and static overlap measures may not accurately capture these dynamics Therefore, this paper selected Recall, Precision, Dice Similarity Coefficient (DSC), Root Mean Squared Error (RMSE) for Intracellular Volume Fraction and diff.HU (difference of average HU values) for the CT value between predicted and ground truth future segmentations. These metrics more directly quantify the model's ability to predict growth patterns over time. Additionally, the Frechet Inception Distance score (FID) [24] is employed as a criterion for validating image quality assessment.

$$Recall = \frac{TP}{TP + FN} \tag{6}$$

$$Precision = \frac{TP}{TP + FP} \tag{7}$$

$$DSC = \frac{2TP}{FP + FN + 2TP} \tag{8}$$

$$RMSE = \sqrt{\frac{\sum(\frac{ICVF_{pred} - ICVF_{GT}}{ICVF_{GT}})^2}{TP}} \tag{9}$$

$$diff.HU = \frac{HU_{pred} - HU_{gt}}{HU_{gt}} \tag{10}$$

In the formula provided above, TP(True Positive), FP (False Positive), and FN (False Negative) denote the true positive, false positive, and false negative values, respectively, indicating the correspondence between the predicted image and the Ground Truth (GT). These metrics provide insight into the accuracy and reliability of the model's predictions in comparison to the actual data. Additionally, $ICVF_{pred}$ and $ICVF_{GT}$ represent the predicted value and ground truth of the intracellular volume fraction, respectively. Intracellular Volume Fraction (ICVF) images, which represent cell density, are normalized within the range [0, 100]. (More details about ICVF calculation can be referred to [25]). HU represents the average Hounsfield units within a volume. Both RMSE and diff.HU are evaluated within the TP.

Furthermore, this study employs the Frechet Inception Distance (FID) score as a quantitative measure to assess the feature-level similarity between real and generated images. The FID score quantifies the feature vector of disparity between a real image x and a generated image $\hat{x}$. It assesses the similarity between two image sets by analyzing the statistical resemblance of computer vision features in the original images. FID measures the distance between their distributions using the mean and covariance matrix. A lower FID score indicates a higher degree of similarity, implying that the generated images closely resemble real ones in terms of structure and content. This paper also employs the FID as an evaluation metric to assess the predictive performance of the method.

$$FID(x, \hat{x}) = (\mu_x - \mu_{\hat{x}})^2 + Tr(\sum_x + \sum_{\hat{x}} - 2\sqrt{\sum_x \sum_{\hat{x}}}) \tag{11}$$

μ_x and $\mu_{\hat{x}}$ represent the mean values of the real image and generated image, respectively. Tr is the rank of the image.

3.3. Performance analysis

To validate the effectiveness of the proposed method, this paper conducts a comparative analysis with the Spatial Transformation (ST) [26], ConvLSTM coordinated longitudinal transformer (LCT-former) [27], Growth Prediction Generative Adversarial Networks (GP-GAN) [25], Conditional Recurrent Variational Autoencoder (CRVAE) [28], Spatio-Temporal Convolution Long Short-Term Memory Network (ST-ConvLSTM) [14], and 3D Contrast-Enhanced Convolutional Long Short-Term Memory network (CE-ConvLSTM) [29]. This study retrained these methods using the collected dataset and tested them on the same test set. Based on tumor images from the first two periods, generate predicted tumor growth results using different methods and compare them with tumors from the third period and real tumors. The results are presented in Table 2. This table provides a quantitative assessment of each method's ability to accurately predict tumor progression, offering insights into their respective performance in forecasting tumor growth patterns.

Based on the results presented in Table 2, the proposed method demonstrates superior performance compared to other approaches, except for a slightly lower recall rate of 79.13% compared to the 80.87% achieved by ST-ConvLSTM. Despite this minor difference, our method excels in other key evaluation metrics, including precision (89.74%), DSC (84.24%), and FID (16.00), indicating a higher degree of accuracy and image similarity. Although the recall rate is not as good as that of ST-ConvLSTM, it is more suitable for screening some high-risk patients to avoid missed diagnosis. Additionally, to ensure

Table 2
Comparison results of different methods on the test set.

	Recall (%)	Precision (%)	DSC (%)	RMSE	FID	diff.HU (%)
ST	57.18 ± 0.11	52.22 ± 0.32	75.40 ± 0.67	36.64	27.67	10.61
CE-ConvLSTM	67.33 ± 0.21	56.51 ± 0.71	66.04 ± 0.42	30.29	26.84	11.06
GP-GAN	65.26 ± 0.58	57.16 ± 0.27	61.27 ± 0.37	33.85	26.68	11.35
ST-ConvLSTM	80.87 ± 0.81	75.46 ± 0.33	78.36 ± 0.21	22.21	25.34	10.21
CRVAE	79.89 ± 0.18	82.22 ± 0.39	82.49 ± 0.27	27.01	35.94	20.35
CE-LCTformer	74.51 ± 0.33	60.90 ± 0.59	80.12 ± 0.43	57.11	21.13	10.33
Our method	79.13 ± 0.51	89.74 ± 0.32	84.24 ± 0.65	18.32	16.00	10.05

the robustness of the results, statistical significance testing was conducted across key performance metrics. Statistical significance was assessed using paired two-sided t-tests over five independent runs on per-case metrics; comparisons were limited to the proposed method versus selected baselines, and no multiple-comparison correction was applied. The results indicate that the differences between the proposed model and the compared models are statistically significant ($p < 0.05$) for all evaluated metrics, confirming the reliability and validity of the observed improvements. The results indicate that the differences between the proposed model and the compared models are statistically significant ($p < 0.05$) for all evaluated metrics, confirming the reliability and validity of the observed improvements.

Upon analyzing the model and experimental outcomes, the ST method suffers from weak performance due to the absence of temporal and structural modeling. GAN-based approaches enhance visual quality but often introduce artifacts, while VAE achieves high recall and precision at the cost of result ambiguity. Transformer-based methods benefit from global dependency modeling and yield high precision, but their excessively high RMSE indicates insufficient pixel-level restoration. In contrast, the ConvLSTM series effectively leverages temporal dependencies to capture tumor dynamics, significantly improving prediction accuracy. Our proposed method further combines convolution for local feature extraction with Transformer for global modeling, achieving the best overall balance in structural fidelity, detail preservation, and pixel consistency. These results suggest that the proposed model effectively captures tumor progression patterns, producing predictions that exhibit minimal deviation from the ground truth.

3.4. Qualitative results

The visualization results of tumor growth prediction are depicted in Fig. 3. In this figure, columns (A), (B), and (C) correspond to tumor images of patients at three different time points, capturing the progression of tumor growth over time. Column (D) showcases the predicted tumor images generated by the ST-ConvTransformer based on the first two time points (A) and (B). It is evident from Fig. 3 that the proposed method effectively predicts tumor growth, with a minimal the disparity between the predictions and the ground truth.

Fig. 3 illustrates qualitative examples where tumor size and shape visibly change across time points (columns A–C). These cases exhibit apparent progression, defined as an increase in segmented tumor volume over time. The predicted tumor images exhibit remarkable similarity to the actual tumor states, indicating the method's accuracy in forecasting tumor growth patterns. This visual comparison underscores the reliability and predictive power of the proposed approach, showcasing its potential utility in clinical settings for tumor prognosis and treatment planning.

By comparing the predictions with actual outcomes, it was demonstrated that the tumor growth prediction method proposed in this study achieves results that meet clinical expectations. It confirms the significant clinical value of tumor growth prediction, as it enables physicians to accurately assess tumor progression speed and malignant potential, thereby guiding clinical decision-making effectively. It can help doctors accurately evaluate the progression rate and malignant potential of tumors, thereby guiding clinical decision-making. Through quantitative analysis of tumor dynamics, clinicians can more effectively identify high-risk patients who require closer monitoring or timely intervention, thereby avoiding both overtreatment and delays in care. This predictive information supports the development of personalized follow-up and treatment plans that consider the patient's overall health status and medical history, enhancing the precision and safety of clinical management.

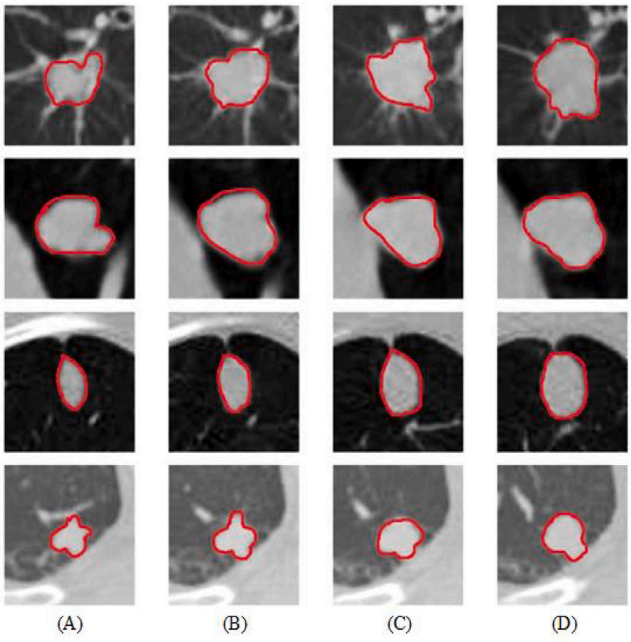

Fig. 3. Pulmonary nodules future prediction results. The first three columns display tumor images at the 1st, 2nd and 3rd (ground truth) time points for four patients. The A column represents the initial tumor image, and the B column represents tumor image at the second stage. The C column is the actual tumor at the third stage. The D column represents the predicted tumor images according to the first two stages (A) and (B) using our methods.

3.5. Ablation study

To validate the effectiveness of the proposed architecture, this study conducted ablation experiments on the collected dataset. These experiments were designed to systematically evaluate the contribution of different model components by selectively removing or modifying specific modules and assessing their impact on the final predictive performance. By analyzing the changes in key evaluation metrics, we can determine the significance of each module in enhancing tumor growth prediction accuracy.

The experimental results, as shown in Table 3, indicate that the model performs exceptionally well in predicting tumor growth. This enhancement underscores the importance of each module within the model and demonstrates that the inclusion of all modules contributes to

Table 3
The impact of different modules on the prediction results.

	Recall (%)	Precision (%)	Dice (%)	dif.HU (%)
ConvTransformer w/o clinical information	63.79	56.93	67.61	14.32
S-ConvTransformer w/o clinical information	71.22	80.65	74.19	12.34
T-ConvTransformer w/o clinical information	75.13	87.74	76.24	13.24
ST-ConvTransformer w/o clinical information	76.10	88.01	79.34	12.11
ST-ConvTransformer with clinical information	79.13	89.74	84.24	10.05

Table 4
The impact of loss function on the prediction results.

	Recall (%)	Precision (%)	Dice (%)
MAE w/o total variation	63.88	76.93	72.76
MAE & total variation	74.34	88.34	79.41
MSE w/o total variation	67.33	78.37	73.41
MSE & total variation	79.13	89.74	84.24

more accurate and reliable tumor predictions. The overall findings confirm the efficacy of the ST-ConvTransformer as a robust and effective tool for predicting tumor growth.

Compared to the spatial gate, the temporal gate demonstrates superior prediction accuracy. It can be attributed to the fact that during the tumor growth process, the characteristics of the tumor change significantly at each time point. At the same time, the spatial position remains relatively stable. As a result, the temporal features, which capture the dynamic changes in tumor characteristics over time, prove to be more effective in predicting tumor growth. In contrast, spatial features are less informative for tumor progression as they primarily focus on the tumor's location rather than its temporal evolution. Additionally, the integration of clinical information can further enhance the model's ability to predict tumor growth, providing valuable context that helps improve overall predictive performance. This additional data enables the model to capture patient-specific risk factors and disease progression patterns that are not discernible from imaging alone. Specifically, in the ablation study, deleting clinical information resulted in a decrease in prediction accuracy from 79.10% to 76.10%, indicating that imaging features alone are insufficient to fully simulate tumor growth dynamics. On the contrary, integrating clinical variables significantly improves accuracy, confirming that the combination of imaging and clinical data enables the model to distinguish between invasive and indolent nodules better, ultimately enhancing its predictive reliability.

This article primarily employs pixel-level mean squared error loss for tumor prediction, while also incorporating a total variation term to encourage smoothness in the predicted tumor images. To assess the effectiveness of this tumor prediction loss function, this study conducted a comparison by using an alternative loss function, namely the mean absolute error, for tumor prediction. The results of this comparison are summarized in Table 4, where we evaluate the performance of the model using both MSE loss and MAE loss in terms of key metrics, such as recall, precision, and Dice similarity coefficient. This comparison enables us to examine the effect of selecting a loss function on the accuracy and quality of tumor predictions.

Adding the total variation term to the original loss function results in significant improvements in the Recall, Precision, and Dice similarity coefficient (DSC). It demonstrates the effectiveness of the total variation term in enhancing the model's ability to predict tumor growth. The total variation term helps reduce noise and preserves the structural integrity of tumor images, leading to more accurate predictions and better alignment with ground truth data.

Table 5
Prediction results of different types of lesions.

	Recall (%)	Precision (%)	Dice (%)
Benign Nodules	80.16 ± 0.12	73.25 ± 0.21	85.41 ± 0.37
Malignant Tumor	64.43 ± 0.35	87.25 ± 0.20	70.88 ± 0.10

Furthermore, when comparing the two loss functions, MSE is found to be more effective than the MAE in predicting tumor growth. The MSE loss function, particularly when combined with the total variation term, yields superior performance across all metrics. It suggests that MSE is better suited for capturing the underlying patterns of tumor progression, especially when pixel-level accuracy is crucial for precise tumor prediction.

3.6. Subgroup experiments

Given the significant differences in growth rates and imaging manifestations between benign pulmonary nodules and malignant tumors, this study differentiated between the two types of lung lesions within the experimental data. Specifically, the dataset includes 2147 images of benign nodules and 653 images of malignant tumors. Both sets of image data were fed into the proposed method for training and verification. The model's predictive performance on these distinct types of lung lesions is evaluated, and the results of the predictions are summarized in Table 4. This differentiation allows for a more nuanced analysis of the model's ability to accurately predict tumor growth in both benign and malignant cases. The results presented in Table 5 demonstrate the method's effectiveness across various lesion types.

From Table 5, it is clear that the recall rate, accuracy, and Dice coefficient for predicting benign nodules reached 80.16%, 93.25%, and 85.41%, respectively. In contrast, the results for malignant tumors were 64.43%, 87.25%, and 70.88%, respectively. These results indicate that the prediction performance for benign nodules is significantly better than for malignant tumors across all metrics.

The primary reason for this discrepancy lies in the growth patterns of the two types of lesions. Most benign nodules tend to grow slowly and exhibit relatively uniform characteristics, making them easier to predict with higher accuracy. On the other hand, malignant tumors grow rapidly and often display more irregular features, such as burrs and lobulation, which complicate the prediction process. The presence of these features increases the variability of tumor appearance, making it more challenging for the model to accurately capture the growth dynamics of malignant tumors, which results in decreased accuracy and overall performance.

To address the relatively weaker prediction performance on malignant tumors, future work will focus on capturing tumor heterogeneity through multimodal imaging and employing malignancy-sensitive loss functions further to enhance the robustness and accuracy of malignant case prediction.

3.7. The relationship between tumor growth and clinical information

In this study, a text transformer is utilized to extract and incorporate clinical information into the tumor prediction model. This clinical data includes factors such as patient smoking history, family disease history, alcohol consumption frequency, and occupational status. By integrating this information, the model can not only learn general tumor growth patterns but also consider personalized patient characteristics, which may significantly impact the progression of lung cancer.

Fig. 4 illustrates the relationship between these four key influencing factors and the model's predictions for tumor growth. This visualization provides a clear representation of how each factor contributes to the tumor growth predictions, offering a more comprehensive understanding of individual risk profiles. By factoring in these clinical elements, the model can tailor predictions based on a patient's unique characteristics,

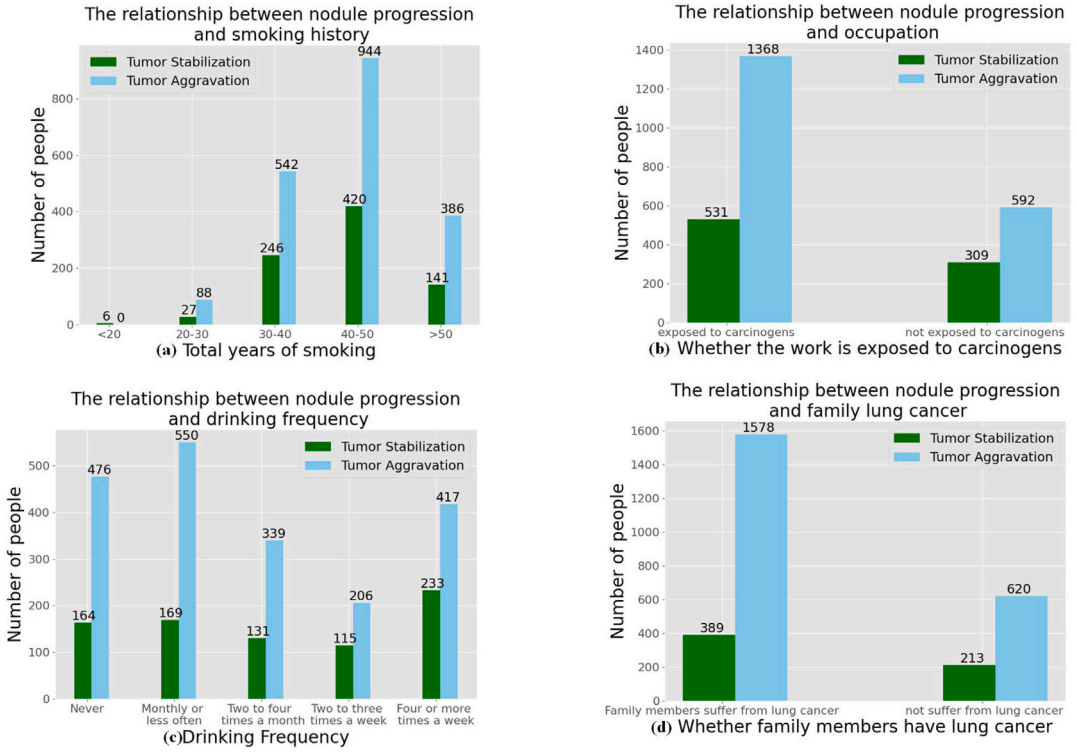

Fig. 4. THE relationship between the risk factors of patients and nodule progression.

providing more accurate and personalized insights into lung cancer progression.

Fig. 4 demonstrates the relationship between various clinical factors and tumor growth, showing that as smoking history and alcohol consumption increase, the tumor growth rate also rises. In contrast, the work environment is shown to have a significant impact on tumor progression, suggesting that occupational exposure plays a crucial role in lung cancer development.

The Pearson correlation coefficients provide further insight into the relationships between these factors and tumor outcomes. Smoking history (C = 0.6531), occupational exposure (C = 0.6400), and family cancer history (C = 0.5519) are all strongly correlated with worse outcomes, indicating that these factors contribute substantially to more aggressive tumor growth. On the other hand, alcohol frequency (C = −0.0969) shows a negative correlation with tumor growth, but its influence is relatively weak compared to smoking and occupational factors. It suggests that while alcohol consumption might have some effect, it is less impactful on tumor progression than smoking or work-related exposures.

Age is not discussed in detail in this study, as the research focuses specifically on patients over 55 years old, where age-related factors are presumed to have a more uniform impact on tumor progression. Therefore, the study does not differentiate age as a separate variable within the patient population.

4. Discussion

Prior studies on tumor progression prediction have primarily relied on radiomics-based models or deep learning approaches such as CNNs. Radiomics models [30] demonstrated the potential of handcrafted features in characterizing tumor morphology, but they were often limited by their sensitivity to feature selection and imaging protocols. Deep learning methods, particularly CNN- and LSTM-based architectures [8,14], have provided improvements by modeling spatial and temporal information; however, they tend to emphasize either local spatial representations or short-range temporal dynamics, which limits their capacity to capture the whole trajectory of tumor evolution. More recently, transformer-based frameworks [16,27] have been introduced to model long-range dependencies, but most of these approaches were designed for unimodal imaging data and did not fully incorporate clinical information.

In contrast, the proposed ST-ConvTransformer integrates longitudinal CT scans with clinical data through a multimodal contrastive learning strategy, enabling a holistic representation of tumor evolution, as shown in Table 3. Our comparative results show that ST-ConvTransformer achieves consistently higher performance across multiple metrics, including Dice similarity for morphology, MAE for volumetric changes, and calibration for progression risk. These findings suggest that the joint modeling of sequential imaging features and clinical characteristics enables a more precise and robust characterization of tumor dynamics. These consistent gains highlight the model's ability to capture both structural evolution and patient-specific factors, thereby offering more precise and clinically meaningful predictions.

The proposed architecture integrates multiple modules, each serving a distinct purpose while working synergistically to enhance prediction accuracy. The spatial gate is designed to filter irrelevant image regions, ensuring that the model focuses on tumor-related features, while the temporal gate captures sequential dependencies across multiple CT scans to model longitudinal tumor progression. The text transformer encodes clinical information into a structured embedding, providing complementary patient-specific context beyond imaging data. Contrastive learning is introduced to align and fuse the image and text embeddings, thereby enhancing cross-modal representation learning. Finally, the U-Net-based fully convolutional network decodes the fused features into fine-grained morphological predictions of tumor evolution. In this framework, spatial and temporal dynamics, clinical information, contrast learning, and U-Net interact to provide more holistic and precise predictions for tumor progression.

Furthermore, this study leverages a multimodal framework to jointly model longitudinal CT scans and clinical data, thereby providing a more robust and individualized prediction of tumor growth trajectories. While imaging data provide detailed morphological and structural

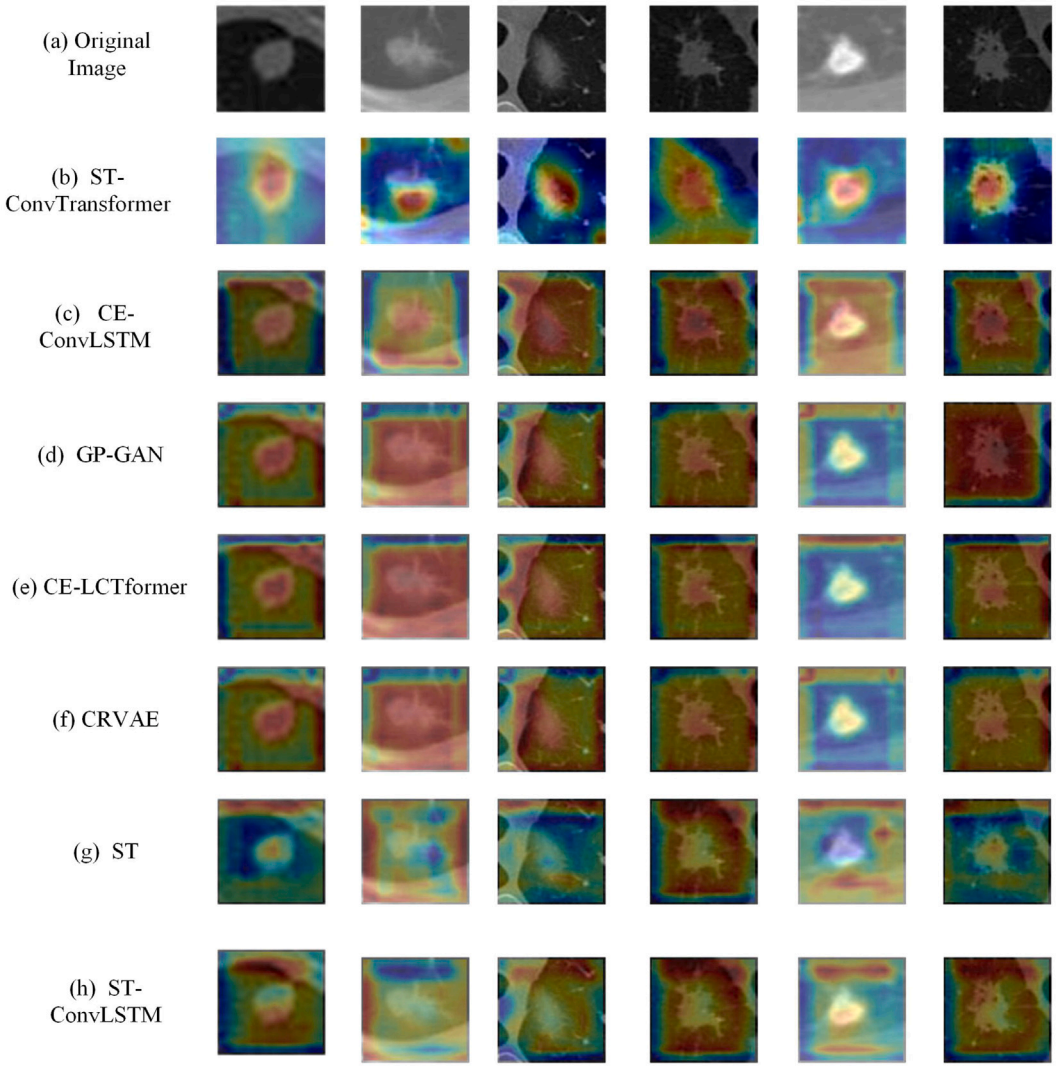

Fig. 5. THE visualization of pulmonary nodule predictions using different models based on GRAD-CAM.

descriptions of tumor evolution, they do not fully capture the patient-specific biological and physiological context. Clinical variables [31,32] can significantly influence tumor progression patterns and treatment responses.

While previous studies have demonstrated the utility of radiomics, CNN/RNN-based models, and transformer architectures for predicting tumor progression, each has exhibited limitations in fully capturing the complexity of tumor dynamics. By jointly modeling longitudinal imaging data and clinical information through multimodal contrastive learning, the proposed ST-ConvTransformer addresses these gaps and consistently outperforms representative baselines across key evaluation metrics. These findings not only corroborate but also extend the current body of literature, underscoring the value of holistic spatiotemporal modeling for achieving more precise and clinically actionable predictions of tumor progression.

Although the proposed ST-ConvTransformer demonstrates strong performance in predicting tumor progression, a significant limitation is the limited discussion of model interpretability. In healthcare applications, explainability is a prerequisite for clinical adoption, as clinicians must be able to understand and trust the outputs of AI models when making treatment decisions. This study utilized Grad-CAM to elucidate the different model's key focus areas. Specifically, through the model's feature activations, it can intuitively display the image regions that different model primarily relies on when predicting tumor progression, as shown in Fig. 5.

As shown in Fig. 5, heatmaps reveal the model's focus on the tumor boundary and surrounding tissues. This visualization not only provides explanatory support for the model but also helps clinical doctors understand the basis for model decisions, thereby enhancing the acceptability and credibility of prediction results in clinical applications. Additionally, the model could revolutionize follow-up care by providing objective, data-driven insights into tumor behavior, allowing clinicians to personalize monitoring schedules. It can be observed that the ST and ST-ConvLSTM and CE-ConvLSTM generally exhibit dispersed attention distributions, making it difficult for them to consistently focus on the core lesion regions of the nodules. Although generative models, such as GP-GAN and CRVAE are able to capture certain structural components of the nodules, their saliency responses show evident blurring and boundary diffusion, indicating limited ability in preserving morphological details and distinguishing subtle changes. In contrast, LCTformer enhances the response to localized abnormal structures to some extent, yet its saliency maps still contain noticeable high-frequency noise and unstable focus patterns. The proposed ST-ConvTransformer demonstrates the most clear and consistent attention patterns across all cases, with saliency strongly concentrated along the nodule boundaries and interiors, accurately delineating the lesion morphology while also reflecting subtle evolutionary trends over longitudinal follow-up. By maintaining temporal coherence in segmentation across longitudinal imaging examinations, it

enables accurate quantification of volumetric and morphological tumor dynamics, which are critical for assessing therapeutic efficacy and disease progression. Moreover, its demonstrated robustness to patient-specific anatomical variations underscores its suitability for deployment in clinical follow-up settings, thereby facilitating reliable AI-driven monitoring in tumor imaging practice.

In addition, the model has exhibited some failure cases, particularly with nodules that exhibit invasive growth in localized areas, where predictions deviated significantly from actual outcomes. These nodules, which may indicate a higher malignancy risk, require closer clinical attention. While clinical data is considered, the model may not fully account for other key patient-specific factors, such as genetic markers, that could influence tumor behavior. Future work will focus on incorporating additional morphological features and genetic data to improve the model's accuracy and robustness.

5. Conclusion

In this paper, we introduce the Spatio-Temporal Convolutional Transformer (ST-ConvTransformer) to address the complex challenge of predicting tumor growth. The model is designed to simultaneously capture intra-slice structures, inter-slice spatial contexts, and temporal dynamics, all of which are essential for precise tumor prediction. By effectively capturing both spatial and temporal relationships, the ST-ConvTransformer provides a more holistic and precise prediction of tumor progression. The quantitative results presented in this study clearly demonstrate the superior performance of our proposed prediction algorithm in forecasting tumor growth. When compared to recent deep learning-based tumor growth prediction models, the proposed model demonstrated superior performance in terms of accuracy, AUC, and Dice similarity; it showed slightly lower Recall compared to ST-ConvLSTM. Despite this trade-off, the ST-ConvTransformer offers substantial clinical utility by providing more reliable predictions for patients with high tumor progression risks. Additionally, the ability to integrate clinical data, including patient history and risk factors, enables the model to generate more personalized and accurate tumor growth prediction.

This article primarily focuses on the longitudinal growth prediction of tumors during the complete disease course and does not explicitly address the issue of missing stage data. In the process of collecting medical imaging data, it is often desirable to obtain comprehensive longitudinal data to track the progression of diseases over time. However, in real-world clinical practice, the availability of complete disease course data is often limited due to various patient-related subjective factors and objective constraints. As a result, a substantial number of cases lack complete stage data, leading to a significant number of missing samples. Although the current study does not explore this aspect, the effective use of incomplete data remains a significant challenge.

CRediT authorship contribution statement

Ning Xiao: Writing – original draft, Conceptualization. **Aoyu Li:** Validation, Software. **Yan Qiang:** Supervision. **Juanjuan Zhao:** Supervision. **Yan Geng:** Data curation. **Leqin Li:** Writing – review & editing.

Declaration of competing interest

The authors declare the following financial interests/personal relationships which may be considered as potential competing interests: Ning Xiao reports financial support was provided by Shanxi Province Fundamental Research. If there are other authors, they declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgment

This work is supported by Shanxi Province Fundamental Research (Grant number 202303021212169).

Data availability

Data will be made available on request.

References

[1] Rebecca L. Siegel, et al., Cancer statistics, 2025, Ca 75 (1) (2025) 10.
[2] K. Kay, K. Dolcy, R. Bies, et al., Estimation of solid tumor doubling times from progression-free survival plots using a novel statistical approach, AAPS J. 26 (5) (2024) 92.
[3] G. Cazoulat, D. Owen, M.M. Matuszak, et al., Biomechanical deformable image registration of longitudinal lung CT images using vessel information, Phys. Med. Biol. 61 (2016) 13.
[4] A. Gandolfi, S. De Franciscis, A. dnofrio, et al., Angiogenesis and vessel co-option in a mathematical model of diffusive tumor growth: The role of chemotaxis, J. Theoret. Biol. 512 (2021) 110526.
[5] F. Iranmanesh, M.A. Nazari, Finite element modeling of avascular tumor growth using a stress-driven model, J. Biomech. Eng. 139 (2024) 081009.
[6] R. Interian, R. Rodrguez-Ramos, F. Valdes-Ravelo, et al., Tumor growth modelling by cellular automata, Math. Mech. Complex Syst. 5 (3) (2017) 239–259.
[7] E. Lujan, M.S. Rosito, A. Soba, et al., Libregrowth: a tumor growth code based on reaction -diffusion equations using shared memory, Comput. Phys. Comm. 243 (2019) 97–105.
[8] L. Zhang, L. Lu, R.M. Summers, et al., Convolutional invasion and expansion networks for tumor growth prediction, IEEE Trans. Med. Imaging 37 (2) (2017) 638–648.
[9] L. Pei, S. Bakas, A. Vossough, et al., Longitudinal brain tumor segmentation prediction in MRI using feature and label fusion, TBiomedical Signal Process. Control. 55 (2020) 101648.
[10] M. Ghita, V. Chandrashekar, D. Copot, et al., Lung tumor growth modeling in patients with NSCLC undergoing radiotherapy, IFAC-PapersOnLine 54 (15) (2021) 233–238.
[11] H.J. Yoon, H. Park, H.Y. Lee, et al., Prediction of tumor doubling time of lung adenocarcinoma using radiomic margin characteristics, AAPS J. 26 (5) (2024) 92.
[12] P.T. Teo, Autonomous lung tumor and critical structure tracking using optical flow computation and neural network prediction, Thorac. Cancer 11 (9) (2020) 2600–2609.
[13] A. Kamli, R. Saouli, H. Batatia, et al., Synthetic medical image generator for data augmentation and anonymisation based on generative adversarial network for glioblastoma tumors growth prediction, IET Image Process. 14 (16) (2020) 4248–4257.
[14] Zhang Ling, Le, et al., Spatio-temporal convolutional LSTMs for tumor growth prediction by learning 4D longitudinal patient data, IEEE Trans. Med. Imaging 39 (4) (2019) 1114–1126.
[15] S. Trajanovski, C. Shan, P.J.C. Weijtmans, et al., Tongue tumor detection in hyperspectral images using deep learning semantic segmentation, IEEE Trans. Biomed. Eng. 68 (4) (2021) 1330–1340.
[16] H. Wang, N. Xiao, J. Zhang, et al., Static–dynamic coordinated transformer for tumor longitudinal growth prediction, Comput. Biol. Med. 148 (2022) 105922.
[17] Li Zhang, Bin Li, Lianfang Tian, et al., Medical image registration based on log-euclidean covariance matrices descriptor, Chinese J. Comput. 42 (9) (2019) 2087–2099.
[18] A. Dosovitskiy, P. Fischer, E. Ilg, et al., Flownet: Learning optical flow with convolutional networks, in: Proceedings of the IEEE International Conference on Computer Vision, 2015, pp. 2758–2766.
[19] E. Ilg, N. Mayer, T. Saikia, et al., Flownet 2.0: Evolution of optical flow estimation with deep networks, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2017, pp. 2462–2470.
[20] Q. Gang, X. Li, H. Wenxing, et al., Ensemble manifold regularized multimodal graph convolutional network for cognitive ability prediction, IEEE Trans. Biomed. Eng. 68 (10) (2021) 3564–3573.
[21] Qirun Hao, Jianwu Li, Yao Lu, et al., Variation-based ring artifact correction in CT images, Acta Automat. Sinica 45 (9) (2021) 1713–1726.
[22] D.R. Aberle, S. DeMello, C.D. Berg, et al., Results of the two incidence screenings in the national lung screening trial, N. Engl. J. Med. 369 (10) (2013) 920–931.
[23] National Lung Screening Trial Research Team, Reduced lung-cancer mortality with low-dose computed tomographic screening, N. Engl. J. Med. 365 (5) (2011) 395–409.
[24] M. Heusel, H. Ramsauer, T. Unterthiner, et al., Gans trained by a two time-scale update rule converge to a local nash equilibrium, in: Advances in Neural Information Processing Systems, 2017, p. 30.

[25] B. Aea, D. Cwc, A. Sjsg, et al., GP-GAN: Brain tumor growth prediction using stacked 3D generative adversarial networks from longitudinal MR images, T Neural Netw. 132 (2024) 321–332.
[26] Y. Zhang, X. Lv, J. Qiu, et al., Deep learning with 3D convolutional neural network for noninvasive prediction of microvascular invasion in hepatocellular carcinoma, J. Magn. Reson. Imaging 54 (1) (2021) 134–143.
[27] M. Ma, X. Zhang, Y. Li, et al., ConvLSTM coordinated longitudinal transformer under spatio-temporal features for tumor growth prediction, Comput. Biol. Med. 163 (2023) 107313.
[28] N. Xiao, Y. Qiang, Z. Zhao, et al., Tumour growth prediction of follow-up lung cancer via conditional recurrent variational autoencoder, IET Image Process. 14 (15) (2020) 3975–3981.
[29] J. Yao, Y. Shi, K. Cao, et al., DeepPrognosis: Preoperative prediction of pancreatic cancer survival and surgical margin via comprehensive understanding of dynamic contrast-enhanced CT imaging and tumor-vascular contact parsing, Med. Image Anal. 73 (2021) 102150.
[30] M. Tan, W. Ma, Y. Sun, et al., Prediction of the growth rate of early-stage lung adenocarcinoma by radiomics, Front. Oncol. 11 (2021) 658138.
[31] M. Kwon, G. Rubio, H. Wang, et al., Smoking-associated downregulation of FILIP1l enhances lung adenocarcinoma progression through mucin production, inflammation, and fibrosis, Cancer Res. Commun. 2 (10) (2022) 1197–1213.
[32] A. Brito-Marcelino, R.J. Duarte-Tavares, K.B. Marcelino, et al., Breast cancer and occupational exposures: an integrative review of the literature, Rev. Bras. de Med. Do Trab. 18 (4) (2021) 488.