Temporal Neighboring Multi-Modal Transformer with Missingness-Aware Prompt for Hepatocellular Carcinoma Prediction

Jingwen Xu[1], Ye Zhu[1], Fei Lyu[1], Grace Lai-Hung Wong[2], and Pong C. Yuen[1]

[1] Department of Computer Science, Hong Kong Baptist University
{csjwxu,csyzhu,feilyu,pcyuen}@comp.hkbu.edu.hk
[2] Department of Medicine and Therapeutics, The Chinese University of Hong Kong
wonglaihung@mect.cuhk.edu.hk

Abstract. Early prediction of hepatocellular carcinoma (HCC) is necessary to facilitate appropriate surveillance strategy and reduce cancer mortality. Incorporating CT scans and clinical time series can greatly increase the accuracy of predictive models. However, there are two challenges to effective multi-modal learning: (a) CT scans and clinical time series suffer from temporal misalignment. (b) CT scans can be missing compared with clinical time series. To tackle the above challenges, we propose a Temporal Neighboring Multi-modal Transformer with Missingness Aware Prompt (**TNformer-MP**) to integrate clinical time series and available CT scans for HCC prediction. To explore the inter-modality temporal correspondence, a Temporal Neighboring Multi-modal Tokenizer (**TN-MT**) is exploited to fuse CT embedding into neighboring clinical time series tokens across multiple scales. To mitigate the performance drop caused by missing CT modality, TNformer-MP exploits a Missingness-aware Prompt-driven Multi-modal Tokenizer (**MP-MT**) that adjusts the encoding of clinical time series tokens with learnable prompts. Experiments conducted on large-scale multi-modal datasets of 36,353 patients show that our method achieves superior performance compared to existing methods.

Keywords: Hepatocellular carcinoma · Multi-modal learning · Temporal neighboring · Prompt.

1 Introduction

Hepatocellular carcinoma (HCC) is the most common primary malignancy of the liver and a leading cause of cancer-related fatalities in the world [21, 8, 4]. Early prediction of HCC is vital to mitigate costs, complications, and mortality. Deep learning methods have shown promise in predicting HCC using clinical time series data extracted from electronic health records [23, 9, 1, 22]. However, these uni-modal approaches have inherent limitations. Contrast-enhanced computed tomography (CT) scans are a crucial component of screening and early diagnosis of HCC in clinical practice [14, 15, 7]. Joint learning from both clinical

time series and CT scans can enhance HCC prediction accuracy by leveraging complementary information [12, 10]. Nevertheless, developing an effective multi-modal model for these two modalities presents its challenges: (a) clinical time series and CT scans exhibit temporal misalignment. As depicted in Figure.1, there is a significant time gap between CT scans and clinical time series, posing difficulties in bridging modality-specific features. (b) CT scans, being the auxiliary modality, may be missing. Obtaining paired data is not always feasible. For instance, CT scans are collected less frequently compared to clinical time series in real-world practice.

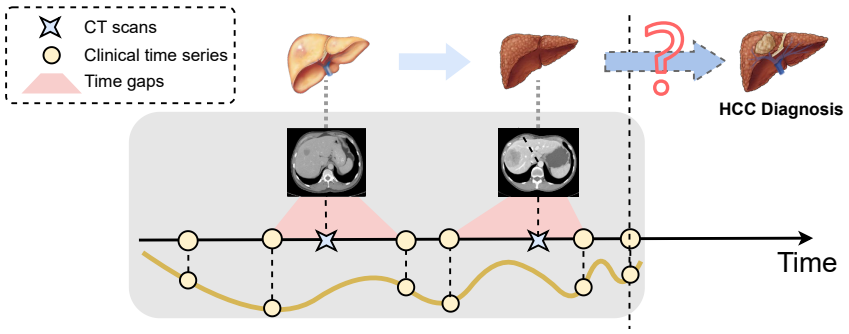


Fig. 1: Incorporating both pathology in CT scan and biomarker information in clinical time series is crucial for early HCC prediction. However, achieving this objective is challenging: (1) CT scans exhibit significant time gaps with clinical time series. (2) CT scans can be missing compared with clinical time series.

Recently, some multi-modal fusion models that combine medical images with non-imaging clinical data have been proposed to outperform single-modality models [19, 26, 5]. Multi-transSP [27] jointly learns the global feature correlations among different modalities for nasopharyngeal carcinoma prediction. TMSS [20] leverages the superiority of transformers in handling different modalities for cancer prediction. [11] propose a strategy to disentangle longitudinal signatures from clinical time series and integrate them with chest CT scans for SPN classification. Using head CT scans and tabular clinical data, [16] exploits a variational distributions combination model to integrate multi-modal information for intracerebral hemorrhage prediction. [13] leverages graph neural networks to capture the node-level and global-level relationships between MR images and tabular clinical data. it is worth noting that most existing models ignore the issue of inter-modal temporal misalignment: they either rely on paired-modality data in the time axis or only consider static medical information. Moreover, these models often assume modality completeness or require additional efforts to generate missing modalities for completeness, which limits their applicability in addressing the scenario of missing CT data in HCC prediction.

To tackle the aforementioned challenges, we proposed a Temporal Neighboring Multimodal Transformer with Missingness-Aware Prompt (**TNformer-MP**) that integrates clinical time series and CT scans for early HCC prediction. Specifically, To bridge the paired modalities from temporal correspondence, we introduce a Temporal Neighboring Multi-modal Tokenizer (**TN-MT**) to combine each CT embedding with a neighboring range of clinical time series tokens across various scales. Moreover, for CT-missing patients, we introduce a Missingness-aware Prompt-driven Multimodal Tokenizer (**MP-MT**) to adopt learnable prompts to adapt the clinical time series tokens to the missing modality scenario. Finally, TNformer-MP correlates the unified tokens using a transformer encoder and performs the final prediction score.

Our contributions are summarized as follows: (1) We introduce a multi-modal framework (TNformer-MP) to perform early HCC prediction by integrating clinical time series and CT scans. (2) TNformer-MP proposes a temporal neighboring multi-modal tokenizer and a missingness-aware prompt-driven multimodal tokenizer to bridge the paired modalities in the time domain while addressing the modality incompleteness. (3) The effectiveness of our method is validated in large-scale real-world patients.

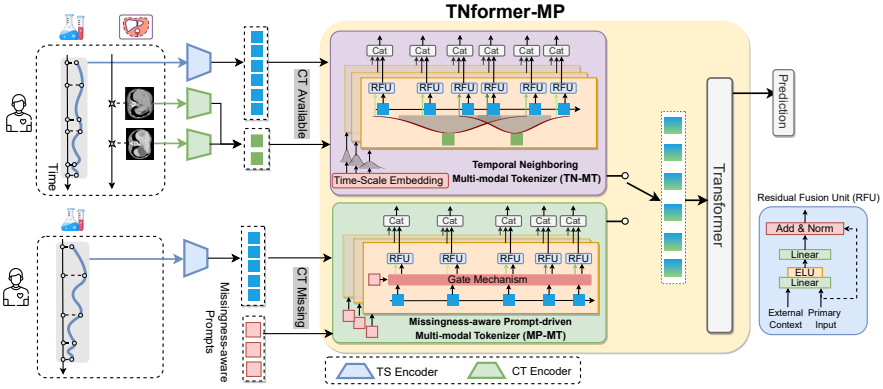


Fig. 2: Overview of the proposed Temporal Neighboring multi-modal transformer with Missingness-aware Prompt (**TNformer-MP**).

2 Proposed Method

Figure 2 illustrates the overview of our proposed method. Modality-specific features are first obtained by a CT encoder and a time series (TS) encoder. On top of that, we propose a Temporal Neighboring multi-modal transformer with Missingness-aware Prompt (**TNformer-MP**) to integrate both modality-specific features for early HCC prediction. Specifically, to encode the unified

tokens, a Temporal Neighboring Multi-modal Tokenizer (**TN-MT**) is exploited to incorporate each CT embedding into neighboring clinical time series tokens across multiple scales. To ensure the robustness of unified tokens in the CT-missing scenario, we introduce a Missingness-aware Prompt-driven Multi-modal Tokenizer (**MP-MT**) to adapt the clinical time series tokens with learnable prompts. Finally, a stack of transformer encoder layers is applied to correlate the unified tokens and perform prediction.

Given the multi-modal data of each patient, we extract modality-specific features $\{G, E\}$ by a CT encoder and TS encoder. $G = \left\{\boldsymbol{g}_{t_1}, ..., \boldsymbol{g}_{t_m}, ..., \boldsymbol{g}_{t_M}\right\}$ represents a collection of CT embeddings, where M indicates the number of CT scans. $E = \{\boldsymbol{e}_{t_1}, ..., \boldsymbol{e}_{t_n}, ..., \boldsymbol{e}_{t_N}\}$ represents a set of clinical time series tokens, where N denotes the sequence length. Then the following multi-modal tokenizers are exploited to integrate the modal-specific features and generate a set of unified tokens.

2.1 Temporal Neighboring Multi-Modal Tokenizer

To tackle the temporal misalignment, our motivation is that the pathology in CT scans generally changes slowly compared to the rapid fluctuations in clinical signals. Therefore, **TN-MT** learns the inter-modal temporal correspondence to bridge each CT scan with a nearby range of clinical time series. Meanwhile, we model the inter-modal temporal correspondence across multiple scales since the different abnormalities in CT scans can have short-term or long-term impacts on the clinical signal.

To model the multi-scale temporal correspondence, a time scale embedding is first developed to learn a set of temporal Gaussian kernels with learnable bandwidths:

$$\begin{aligned} \kappa_{\lambda_k}\left(t_m, t_n\right) &= \exp\left(-\frac{\left\|t_m - t_n\right\|^2}{\lambda_k}\right) \\ \lambda_k &= \exp\left(\boldsymbol{W}\boldsymbol{q}_k + b\right), \boldsymbol{q}_k \in \{\boldsymbol{q}_1, ..., \boldsymbol{q}_K\} \end{aligned} \tag{1}$$

where λ_k is bandwidth estimated by an affine network and a learnable query vector $\boldsymbol{q}_k$. Meanwhile, we exploite a set of query vectors $\{\boldsymbol{q}_1, ..., \boldsymbol{q}_K\}$ to parameterize multi-scale kernels. t_m and t_n denote the timestamps of CT and clinical time series respectively.

Then each $\kappa_{\lambda_k}(.)$ is utilized to distribute the CT embeddings to a set of clinical time series tokens by measuring their temporal similarities:

$$\bar{\boldsymbol{g}}_{t_n} = \sum_{m=1}^{M} \frac{\kappa_{\lambda_k}\left(t_m, t_n\right)}{Z\left(t_m, \mathbf{t}_N\right)} \boldsymbol{g}_{t_m}, \quad Z\left(t_m, \boldsymbol{t}_N\right) = \sum_{n=1}^{N} \kappa_{\lambda_k}\left(t_m, t_n\right) \tag{2}$$

where $\bar{\boldsymbol{g}}_{t_n}$ is the weighted sum of CT embeddings at each t_n. The unified token is obtained by incorporating $\bar{\boldsymbol{g}}_{t_n}$ into $\boldsymbol{e}_{t_n}$ in a Residual Fusion Unit (RFU):

$$\boldsymbol{u}_{t_n} = \text{RFU}\left(\boldsymbol{e}_{t_n}, \bar{\boldsymbol{g}}_{t_n}\right) = \text{LayerNorm}\left(\boldsymbol{e}_{t_n} + \boldsymbol{\eta}\right) \tag{3}$$

$$\boldsymbol{\eta} = \mathrm{ELU}\left(\boldsymbol{W}_1\boldsymbol{e}_{t_n} + \boldsymbol{W}_2\bar{\boldsymbol{g}}_{t_n} + \boldsymbol{b}\right) \tag{4}$$

where $\boldsymbol{e}_{t_n}$ is the primary input and $\bar{\boldsymbol{g}}_{t_n}$ acts as the optional context vector. RFU allows to control the extent to which the optional context vector contributes to the primary input.

Based on the multi-head mechanism, the aforementioned process is executed simultaneously with K temporal kernels. All output vectors at each time stamp are then combined to form the ultimate unified token.

2.2 Missingness-Aware Prompt-Driven Multi-Modal Tokenizer

The CT modality incompleteness causes the different distributions of input tokens, which can degrade the performance. Recent research on prompt learning research has shown promising adaptation in medical scenarios [2, 24, 25]. To tackle the above problem, inspired by the prompt learning, **MP-MT** adopts learnable prompts to tailor the model to the missing-modality input. In particular, instead of directly prepending prompts to input tokens, it mimics the paired-modality process to dynamically adapt the clinical time series tokens with prompts before attaching them to the transformer layers.

For each prompt $\boldsymbol{p}$, it is dynamically distributed to each timestamp by measuring the compatibility with the corresponding clinical time series token via a gate mechanism:

$$\bar{\boldsymbol{p}}_{t_n} \;=\; \mathrm{Gate}\left(\boldsymbol{p}, \boldsymbol{e}_{t_n}\right) \;=\; \sigma\left(\boldsymbol{W}_1\boldsymbol{p} + \boldsymbol{W}_2\boldsymbol{e}_{t_n} + \boldsymbol{b}_1\right) \odot \left(\boldsymbol{W}_3\boldsymbol{p} + \boldsymbol{b}_2\right) \tag{5}$$

where the $\bar{\boldsymbol{p}}_{t_n}$ denotes the weighted prompt vector at each t_n. The weights represent the compatibility between each clinical time series token and prompt in Eq.(5), which is generated by a sigmoid function σ. In this way, the gate mechanism allows the network to selectively adjust input tokens.

To sufficiently adapt the corresponding clinical time series token with the weighted prompt vector $\bar{\boldsymbol{p}}_{t_n}$, **MP-MT** mimics the paired-modality process to integrate $\boldsymbol{e}_{t_n}$ and $\bar{\boldsymbol{p}}_{t_n}$ via an RFU:

$$\begin{aligned} \boldsymbol{u}_{t_n} = \mathrm{RFU}\left(\boldsymbol{e}_{t_n}, \bar{\boldsymbol{p}}_{t_n}\right) &= \mathrm{LayerNorm}\left(\boldsymbol{e}_{t_n} + \boldsymbol{W}_3\,\boldsymbol{\eta}\right) \\ \boldsymbol{\eta} \;&=\; \mathrm{ELU}\left(\boldsymbol{W}_1\boldsymbol{e}_{t_n} + \boldsymbol{W}_2\bar{\boldsymbol{p}}_{t_n} + \boldsymbol{b}\right) \end{aligned} \tag{6}$$

where RFU fully updates $\boldsymbol{e}_{t_n}$ by fusing with $\bar{\boldsymbol{p}}_{t_n}$ and controling the extent to which the updated embedding contributes to the original token. A set of P prompts are simultaneously utilized for adaptation based on the multi-head mechanism. Furthermore, all output vectors are combined.

Finally, the unified tokens are input into a stack of transformer encoder layers and the prediction probability is obtained by a prediction layer.

3 Experiments

3.1 Experiment Settings

Dataset. We conduct the HCC prediction experiments on a territory-wide cohort of patients with chronic viral hepatitis (CVH). The data is collected from

Table 1: HCC prediction performance in different test sets.

Modality	Prediction Window	Model	$(\mathbf{TS}+\mathbf{CT})_{\mathbf{ALL}}$		$(\mathbf{TS}+\mathbf{CT})_{\mathbf{PARTIAL}}$		$(\mathbf{TS}+\mathbf{CT})_{\mathbf{PAIR}}$	
			AUROC	**AUPRC**	**AUROC**	**AUPRC**	**AUROC**	**AUPRC**
Uni-Modal	1-year	CT Encoder	-	-	-	-	80.2±1.6	32.6±1.9
		TS Encoder	87.6±1.3	32.5+2.6	86.6±2.1	33.9±1.7	90.1±0.8	47.9±2.1
	2-year	CT Encoder	-	-	-	-	77.5±2.4	22.3±1.9
		TS Encoder	85.3±1.4	27.5±3.1	84.2±1.7	25.2±2.3	88.7±2.3	38.3±2.7
	3-year	CT Encoder	-	-	-	-	75.4±3.1	12.7±2.5
		TS Encoder	84.7±2.2	21.2±2.8	83.7±2.3	19.7±3.5	87.5±2.7	24.5±3.3
Multi-Modal	1-year	MedFuse[6]	88.9±2.1	35.7±1.8	87.5±1.9	34.7±3.1	89.3±2.5	48.7±1.9
		TMSS[20]	88.5±3.1	37.2±2.3	86.8±2.2	35.1±2.5	89.6±1.9	49.6±2.3
		TDSig[11]	89.6±1.8	38.3±1.6	88.2±1.5	35.9±1.8	90.6±2.1	49.4±2.4
		Ours	**91.5±2.3**	**39.8±2.2**	**89.6±1.7**	**37.2±2.1**	**92.8±2.4**	**51.7±2.4**
	2-year	MedFuse[6]	86.7±2.4	33.8±1.7	86.1±2.2	27.8±2.7	89.4±2.8	40.1±2.3
		TMSS[20]	87.3±1.9	34.2±2.5	86.9±2.4	28.3±3.3	90.1±2.1	42.3±2.7
		TDSig[11]	87.9±2.5	34.9±2.1	87.1±1.8	29.2±2.5	90.6±1.6	43.1±2.2
		Ours	**89.2±2.1**	**35.7±1.9**	**88.2±1.7**	**31.2±2.3**	**91.1±2.3**	**45.3±1.9**
	3-year	MedFuse[6]	85.1±2.9	23.9±2.3	84.3±2.5	21.2±3.6	88.9±2.3	28.7±3.1
		TMSS[20]	85.8±2.5	24.6±2.6	84.9±2.8	22.7±3.1	89.3±3.1	30.1±2.5
		TDSig[11]	85.3±2.1	26.1±2.7	84.5±2.1	23.8±1.7	89.7±2.7	31.4±3.3
		Ours	**86.7±2.6**	**28.1±2.1**	**86.1±2.3**	**25.1±2.7**	**90.1±2.5**	**33.6±2.7**

the Hospital Authority Data Collaboration Lab (HADCL), Hong Kong. We take the 15-year follow-up for patients, starting from the first CVH diagnosis date. During this follow-up, we extracted the records of 46 clinical parameters following [21], which form the clinical time series data for each patient. We extracted all available CT scans for each patient during the follow-up period. Following the above procedures, we obtained a cohort of 36,353 patients, wherein 7.4% of the patients were diagnosed with HCC during the follow-up. The dataset consists of 36,353 clinical time series and 7,622 CT scans extracted from 36,353 patients, with 5,216 patients having both clinical time series and CT scan.

Prediction Window. We conduct the N-year (i.e., N=1,2,3) prediction task (predict the likelihood of HCC diagnosis for a patient N years after the last visit). In each N-year prediction configuration, we ascertain the inclusion of positive patients by imposing the criterion that the HCC diagnosis date must exceed N years beyond the date of the last visit.

Data Split. The dataset is randomly split into a training set (70%), a validation set (10%) and a test set (20%). The results of 5-fold cross-validation results are reported. The test set can be categorized into three subsets according to the modality availability: (1) $(\mathbf{TS}+\mathbf{CT})_{\mathbf{ALL}}$ includes all patients. (2) $(\mathbf{TS}+\mathbf{CT})_{\mathbf{PAIR}}$ includes patients with paired modalities. (3) $(\mathbf{TS}+\mathbf{CT})_{\mathbf{PARTIAL}}$ includes patients with only clinical time series.

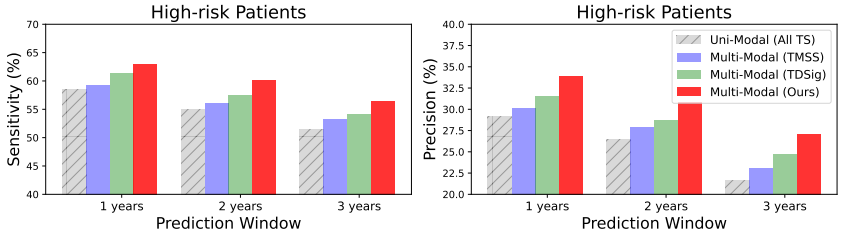
Metrics. The experimental results are evaluated in terms of the area under the receiver operator characteristic curves (AUROC) and the area under the precision-recall curve (AUPRC). In addition, we also evaluate the risk stratification performance based on the dual-cutoff strategy [18, 3, 17], where two cutoff values with sensitivity $> 90\%$ and specificity $> 90\%$ are selected. Accordingly,

patients are categorized into low-risk, intermediate-risk, and high-risk levels. The most predictive model is expected to maximize the sensitivity & precision for high-risk patients while prioritizing specificity & negative predictive value (NPV) for low-risk patients.

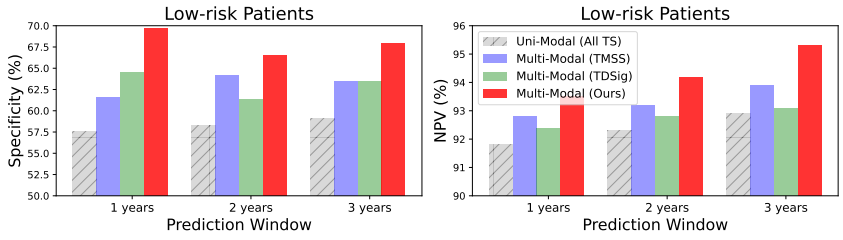
More implementation details are included in supplementary material.

3.2 Evaluations

Compared with State-of-the-Art Methods. To evaluate the effectiveness of our method, we compare with three recent multi-modal models, namely MedFuse[6], TMSS[20] and TDSig[11]. Table 1 presents the accuracy performance in three prediction windows. We first observe that incorporating CT scans as an auxiliary modality during both training and inference enhances the performance of the uni-modal prediction. Our method outperforms all other methods consistently in terms of AUROC and AUPRC for both paired and unpaired test sets. Notably, as the prediction window increases, we notice that the impact of multi-modal fusion becomes more necessary compared to uni-modal learning. In this regard, our method demonstrates a larger improvement gain than the other approaches, particularly for larger prediction windows.



(a) High-risk Patients



(b) Low-risk Patients

Fig. 3: Performance of risk stratification across different methods

Risk Stratification In Different Prediction Windows. We conduct experiments to validate the risk stratification performance based on the dual-cutoff

strategy (Supplementary Information: Section C). The patients are categorized into low, intermediate, and high-risk levels. As depicted in Figure 3 (a), our method exhibits superior sensitivity and precision for high-risk patients compared to the other baselines. And the improvement is more pronounced as the prediction window expands. This indicates that our method can assist clinicians in providing timely interventions. In addition, as depicted in Figure 3 (b), our method achieves better specificity and negative predictive value for low-risk patients. This enables clinicians to allocate healthcare resources efficiently.

Table 2: Ablation study.

Prediction Window	Model	TN-MT	MP-MT	(TS + CT)ALL		(TS + CT)PARTIAL		(TS + CT)PAIR	
				AUROC	AUPRC	AUROC	AUPRC	AUROC	AUPRC
1-year	baseline			89.1±2.1	37.2±1.6	86.5±2.3	40.7±1.9	89.4±2.1	48.9±2.6
	w/o MP-MT	✓		90.1±1.8	38.1±2.7	88.2±2.7	42.6±2.5	91.7±3.2	51.1±2.1
	w/o TN-MT		✓	90.9±2.5	38.9±2.1	89.1±1.5	43.9±2.3	90.6±1.8	50.3±1.8
	Ours	✓	✓	**91.5±2.3**	**39.8±2.2**	**89.6±1.7**	**37.2±2.1**	**92.8±2.4**	**51.7±2.4**
2-year	baseline			86.6±1.7	32.7±2.7	84.1±2.5	26.5±1.5	89.1±1.5	43.5±2.3
	w/o MP-MT	✓		87.6±2.3	33.9±1.3	85.9±2.1	28.6±2.3	90.2±2.1	44.1±1.8
	w/o TN-MT		✓	88.7±2.7	34.9±2.5	87.4±2.3	30.7±3.2	89.5±1.7	42.9±1.6
	Ours	✓	✓	**89.2±2.1**	**35.7±1.9**	**88.2±1.7**	**31.2±2.3**	**91.1±2.3**	**45.3±1.9**
3-year	baseline			84.1±2.4	24.3±1.6	82.9±1.8	21.8±2.3	87.2±3.1	30.8±3.1
	w/o MP-MT	✓		85.3±3.1	26.2±2.5	84.3±2.6	23.6±3.1	89.6±2.3	32.9±2.1
	w/o TN-MT		✓	86.1±2.4	27.6±2.6	85.6±2.1	24.7±2.3	88.4±2.5	31.8±1.7
	Ours	✓	✓	**86.7±2.6**	**28.1±2.1**	**86.1±2.3**	**25.1±2.7**	**90.1±2.5**	**33.6±2.7**

Ablation Study. We perform ablation studies over each component of our method. We compare three variants of our method: (1) *baseline*: it directly concatenates CT embeddings to clinical time series tokens. (2) *w/o MP-MT*: it removes the MP-MT. (3) *w/o TN-MT*: it replaces the TN-MT with the concatenation of paired-modality tokens. Table 2 illustrates the comparison results. *Ours* significantly outperforms the *w/o MP-MT* and *w/o TN-MT* in the $(\mathbf{TS}+\mathbf{CT})_{\mathbf{ALL}}$, demonstrating the effectiveness of two novel modules. The significant improvement with *w/o MP-MT* over *baseline* in $(\mathbf{TS}+\mathbf{CT})_{\mathbf{PAIR}}$ demonstrates the advantage of temporal correspondence modeling in the multi-modal learning based on CT scans and clinical time series. There are large performance gaps between *w/o TN-MT* and *baseline* in the $(\mathbf{TS}+\mathbf{CT})_{\mathbf{ALL}}$ and $(\mathbf{TS}+\mathbf{CT})_{\mathbf{PARTIAL}}$, demonstrating that the sufficient interaction between prompts and clinical time series can substantially improve the performance of missing-modality samples with multi-modal learning.

4 Conclusion

In this paper, we propose a temporal neighboring multi-modal Transformer with missingness-aware prompt for early HCC prediction based on clinical time series

and CT scans. A temporal neighboring multi-modal tokenizer and missingness-aware prompt-driven multi-modal tokenizer are introduced to bridge the modality-specific features with the temporal correspondence while addressing the modality incompleteness. Experiments in large-scale real-world patients show that our method yields superior prediction performance with promising risk stratification.

Acknowledgments. This work was supported by Hong Kong Research Grants Council General Research Fund under Grant RGC/HKBU12200122.

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

References

1. An, C., Choi, J.W., Lee, H.S., Lim, H., Ryu, S.J., Chang, J.H., Oh, H.C.: Prediction of the risk of developing hepatocellular carcinoma in health screening examinees: a korean cohort study. BMC cancer **21**, 1–12 (2021)
2. Chen, Z., Diao, S., Wang, B., Li, G., Wan, X.: Towards unifying medical vision-and-language pre-training via soft prompts. In: ICCV. pp. 23403–23413 (October 2023)
3. Chin, J., Powell, L.W., Ramm, L.E., Hartel, G.F., Olynyk, J.K., Ramm, G.A.: Utility of serum biomarker indices for staging of hepatic fibrosis before and after venesection in patients with hemochromatosis caused by variants in hfe. Clinical Gastroenterology and Hepatology **19**(7), 1459–1468 (2021)
4. Fan, R., Chen, L., Zhao, S., Yang, H., Li, Z., Qian, Y., Ma, H., Liu, X., Wang, C., Liang, X., et al.: Novel, high accuracy models for hepatocellular carcinoma prediction based on longitudinal data and cell-free dna signatures. Journal of Hepatology (2023)
5. Feng, Y., Chen, W., Gu, X., Xu, X., Zhang, M.: Multi-modal semi-supervised evidential recycle framework for alzheimer's disease classification. In: MICCAI. pp. 130–140. Springer (2023)
6. Hayat, N., Geras, K.J., Shamout, F.E.: Medfuse: Multi-modal fusion with clinical time-series data and chest x-ray images. In: Machine Learning for Healthcare Conference. pp. 479–503. PMLR (2022)
7. Hu, Q., Chen, Y., Xiao, J., Sun, S., Chen, J., Yuille, A.L., Zhou, Z.: Label-free liver tumor segmentation. In: CVPR. pp. 7422–7432 (2023)
8. Huang, D.Q., El-Serag, H.B., Loomba, R.: Global epidemiology of nafld-related hcc: trends, predictions, risk factors and prevention. Nature Reviews Gastroenterology & Hepatology **18**(4), 223–238 (2021)
9. Ioannou, G.N., Tang, W., Beste, L.A., Tincopa, M.A., Su, G.L., Van, T., Tapper, E.B., Singal, A.G., Zhu, J., Waljee, A.K.: Assessment of a deep learning model to predict hepatocellular carcinoma in patients with hepatitis c cirrhosis. JAMA network open **3**(9), e2015626–e2015626 (2020)
10. Li, S., Fang, Y., Wang, G., Zhang, L., Zhou, W.: Inter-modal conditional-guided fusion network with transformer for grading hepatocellular carcinoma. In: ISBI. pp. 1–5. IEEE (2023)
11. Li, T.Z., Still, J.M., Xu, K., Lee, H.H., Cai, L.Y., Krishnan, A.R., Gao, R., Khan, M.S., Antic, S., Kammer, M., et al.: Longitudinal multimodal transformer integrating imaging and latent clinical signatures from routine ehrs for pulmonary nodule classification. In: MICCAI. pp. 649–659. Springer (2023)

12. Li, Z., Wang, Y., Zhu, Y., Xu, J., Wei, J., Xie, J., Zhang, J.: Modality-based attention and dual-stream multiple instance convolutional neural network for predicting microvascular invasion of hepatocellular carcinoma. Frontiers in Oncology **13** (2023)
13. Liu, S., Zhang, B., Fang, R., Rueckert, D., Zimmer, V.A.: Dynamic graph neural representation based multi-modal fusion model for cognitive outcome prediction in stroke cases. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 338–347. Springer (2023)
14. Lyu, F., Ma, A.J., Yip, T.C.F., Wong, G.L.H., Yuen, P.C.: Weakly supervised liver tumor segmentation using couinaud segment annotation. IEEE Transactions on Medical Imaging **41**(5), 1138–1149 (2021)
15. Lyu, F., Ye, M., Ma, A.J., Yip, T.C.F., Wong, G.L.H., Yuen, P.C.: Learning from synthetic ct images via test-time training for liver tumor segmentation. IEEE transactions on medical imaging **41**(9), 2510–2520 (2022)
16. Ma, W., Chen, C., Abrigo, J., Mak, C.H.K., Gong, Y., Chan, N.Y., Han, C., Liu, Z., Dou, Q.: Treatment outcome prediction for intracerebral hemorrhage via generative prognostic model with imaging and tabular data. In: MICCAI. pp. 715–725. Springer (2023)
17. Noureddin, M., Mena, E., Vuppalanchi, R., Samala, N., Wong, M., Pacheco, F., Polanco, P., Sakkal, C., Antaramian, A., Chang, D., et al.: Increased accuracy in identifying nafld with advanced fibrosis and cirrhosis: independent validation of the agile 3+ and 4 scores. Hepatology Communications **7**(5) (2023)
18. Park, S., Kwon, J.H., Kim, S.Y., Kang, J.H., Chung, J.I., Jang, J.K., Jang, H.Y., Shim, J.H., Lee, S.S., Kim, K.W., et al.: Cutoff values for diagnosing hepatic steatosis using contemporary mri-proton density fat fraction measuring methods. Korean Journal of Radiology **23**(12), 1260 (2022)
19. Ren, C.X., Xu, G.X., Dai, D.Q., Lin, L., Sun, Y., Liu, Q.S.: Cross-site prognosis prediction for nasopharyngeal carcinoma from incomplete multi-modal data. Medical Image Analysis p. 103103 (2024)
20. Saeed, N., Sobirov, I., Al Majzoub, R., Yaqub, M.: Tmss: An end-to-end transformer-based multimodal network for segmentation and survival prediction. In: MICCAI. pp. 319–329. Springer (2022)
21. Wong, G.L.H., Hui, V.W.K., Tan, Q., Xu, J., Lee, H.W., Yip, T.C.F., Yang, B., Tse, Y.K., Yin, C., Lyu, F., et al.: Novel machine learning models outperform risk scores in predicting hepatocellular carcinoma in patients with chronic viral hepatitis. JHEP Reports **4**(3), 100441 (2022)
22. Wong, G.L.H., Yuen, P.C., Ma, A.J., Chan, A.W.H., Leung, H.H.W., Wong, V.W.S.: Artificial intelligence in prediction of non-alcoholic fatty liver disease and fibrosis. Journal of gastroenterology and hepatology **36**(3), 543–550 (2021)
23. Xu, J., Lyu, F., Yuen, P.C.: Density-aware temporal attentive step-wise diffusion model for medical time series imputation. In: Proceedings of the 32nd ACM International Conference on Information and Knowledge Management. pp. 2836–2845 (2023)
24. Ye, Y., Xie, Y., Zhang, J., Chen, Z., Xia, Y.: Uniseg: A prompt-driven universal segmentation model as well as a strong representation learner. In: MICCAI. p. 508–518. Springer (2023)
25. Yin, C., Liu, S., Zhou, K., Wong, V.W.S., Yuen, P.C.: Prompting vision foundation models for pathology image analysis. In: CVPR. pp. 11292–11301 (2024)
26. Zhang, L., Li, Z., Chandra, S.S., Nasrallah, F.: Multi-modal traumatic brain injury prognosis via structure-aware field-wise learning. IEEE Transactions on Knowledge and Data Engineering (2024)

27. Zheng, H., Lin, Z., Zhou, Q., Peng, X., Xiao, J., Zu, C., Jiao, Z., Wang, Y.: Multi-transsp: Multimodal transformer for survival prediction of nasopharyngeal carcinoma patients. In: MICCAI. pp. 234–243. Springer (2022)