CTSL: Codebook-based Temporal-Spatial Learning for Accurate Non-Contrast Cardiac Risk Prediction Using Cine MRIs

Haoyang Su[1,2,4], Shaohao Rui[2,3,4], Jinyi Xiang[3], Lianming Wu[3], and Xiaosong Wang[4](✉)

[1] Fudan University, Shanghai, China
[2] Shanghai Innovation Institute, Shanghai, China
[3] Shanghai Jiao Tong University, Shanghai, China
[4] Shanghai Artificial Intelligence Laboratory, Shanghai, China
wangxiaosong@pjlab.org.cn

Abstract. Accurate and contrast-free Major Adverse Cardiac Events (MACE) prediction from cine MRI sequences remains a critical challenge. Existing methods typically necessitate supervised learning based on human-refined masks in the ventricular myocardium, which become impractical without contrast agents. We introduce a self-supervised framework, namely Codebook-based Temporal-Spatial Learning (CTSL), that learns dynamic, spatiotemporal representations from raw cine data without requiring segmentation masks. CTSL decouples temporal and spatial features through a multi-view distillation strategy, where the teacher model processes multiple cine views, and the student model learns from reduced-dimensional cine-SA sequences. By leveraging codebook-based feature representations and dynamic lesion self-detection through motion cues, CTSL captures intricate temporal dependencies and motion patterns. High-confidence MACE risk predictions are achieved through our model, providing a rapid, non-invasive solution for cardiac risk assessment that outperforms traditional contrast-dependent methods, thereby enabling timely and accessible heart disease diagnosis in clinical settings.

Keywords: Motion-aware Multi-view Distillation · Temporal-Spatial Feature Disentangling · Non-contrast Survival Prediction.

1 Introduction

The application of MACE in survival analysis within cardiology is of paramount importance, serving as a critical indicator of long-term cardiac health and treatment outcomes [3, 27, 26]. In this context, cine cardiac MRI imaging is widely accessible, while its prognostic efficacy is significantly hindered by the inherent intricacy of myocardial tissue and the entanglement of its temporal and spatial dynamics [24]. Classical methods [5, 15, 21], modeled through electronic health records (EHR) or radiomics, purely rely on manual interpretations of structural and functional abnormalities [1, 2], which are subject to inter-observer variability and often fail to capture subtle, yet crucial, prognostic features. Though the landscape of state-of-the-art survival models for 3D medical imaging is vast, limitations still persist. XSurv [19], which utilizes multi-modal data such as PET and CT scans, struggles with the scarcity of paired samples and the challenges of data co-registration. AdaMSS [20], which requires physician-driven lesion refinement, is both time-consuming and labor-intensive. Furthermore, models specialized in pathology [29, 25, 13, 11] are limited by their reliance on 2D imaging, resulting in poor generalization to high-dimensional images. As a result, while cine imaging is a commonly available modality, its integration of multi-dimensional data, including multi-chamber dynamics from short-axis and longitudinal views of cardiac morphology over time, still remains a challenge in survival analysis.

In this work, we first present a self-supervised pre-training scheme, denoted as CTSL, which operates independently of heart masks or contrast imaging data. The framework mainly comprises two stages: motion-aware multi-view model distillation and spatiotemporal disentangling. Initially, we extend the classical distillation learning paradigm, DINOv2 [23], from a patient-level perspective, innovatively incorporating multi-view cine sequences as input for the distillation, i.e., injecting the information from other views than short-axis (SA) images into the pre-trained model. In this stage, motion queries extracted through SA cine sequences are treated as myocardium-oriented key tokens by the student network, which aligns with long-axis cine tokens from the teacher network via Kullback-Leibler (KL) divergence [17]. Subsequently, drawing upon the latent space discretization techniques of VQVAE [22], we extract query tokens from the preceding KL-aligned student model and design trainable temporal and spatial codebook embeddings, disentangling the spatiotemporal representations from the compressed 4D cine data. Finally, a survival prediction framework is presented using the learned image tokens from CTSL and EHR features to perform MACE-based survival analysis.

Our contributions in the proposed framework are threefold: 1) We demonstrate the feasibility of adopting contrast-free imaging techniques together with EHR for the MACE survival analysis. 2) We introduce a self-supervised framework, CTSL, that learns codebook-based spatiotemporal representations from raw cine data via a motion-aware multi-view model distillation module and a spatiotemporal feature disentanglement module. 3) We evaluate the proposed survival analysis framework on three private datasets and demonstrate its superior performance compared to prior arts.

2 Method

We propose the CTSL framework shown in Fig. 1, which operates through a two-stage self-supervised learning paradigm, followed by a final survival prediction stage. In stage I, multi-view cine sequences are processed independently by teacher and student networks, where KL loss $\mathcal{L}_{KL}$ aligns SA dynamics with long-axis anatomical patterns, while motion-aware distillation is enforced via motion contrastive loss $\mathcal{L}_{MCL}$, generating motion queries that represent myocardial dynamics. Spatiotemporal codebooks are composed based on student-derived fea-

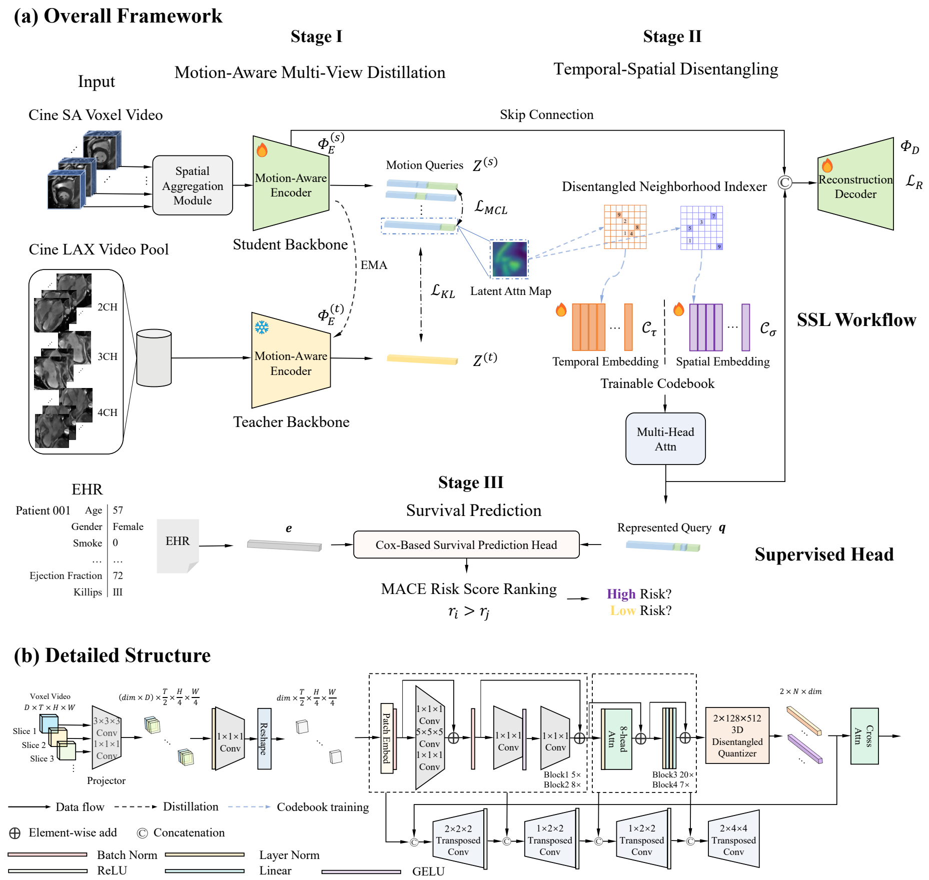


Fig. 1. Overall framework and detailed structure of CTSL. (a) Self-supervised cardiac risk prediction framework. (b) Model architecture with Uniformer [18] backbone.

tures using nearest-neighbor indexing, generating compact representations for survival prediction. At last, final represented queries, fused with EHR data, drive Cox-based risk stratification through parameterized temporal and spatial embeddings.

2.1 Preprocessing: Adaptive Myocardial Motion Localization

We apply a mask-free Region-of-Interest (ROI) preprocessing strategy, where optical flow is employed to extract the myocardial motion-focused region $\mathcal{V}$ from the full heart Voxel Video $\mathcal{V}^{(total)} \in \mathbb{R}^{H^{(total)} \times W^{(total)} \times T \times D}$. The Farneback dense optical flow algorithm [7] is applied to estimate the motion field $\mathbf{F}^{(t)}$ between adjacent Cine frame slices.

$$\mathbf{F}^{(t)} = \Psi_{\mathrm{FB}}\left(\mathcal{I}_t, \mathcal{I}_{t+1}\right) \in \mathbb{R}^{H \times W \times 2}, \ \forall t \in \{0, \ldots, T-1\}. \tag{1}$$

The global-level ROI center $\bar{\mathbf{c}}$ is determined by aggregating the centroid trajectories across time windows. A window width of $s = 96$ is utilized as the resolution of the ROI, obtaining the resulting cine myocardial voxel video to be fed into the SSL framework as $\mathcal{V} = \mathcal{V}^{(total)}[\bar{c_y} - s/2 : \bar{c_y} + s/2, \bar{c_x} - s/2 : \bar{c_x} + s/2, : , :] \in \mathbb{R}^{H \times W \times T \times D}$.

2.2 Stage I: Motion-aware Multi-view Model Distillation

Given the preprocessed 4D cine ROI sequence $\mathcal{V} \in \mathbb{R}^{H \times W \times T \times D}$, we designed paired motion-aware encoders $\Phi_E^{(s)}$ and $\Phi_E^{(t)}$ through a teacher-student distillation paradigm. A spatial aggregation module Γ_p first processes the input through depth-wise feature extraction and obtain

$$\Gamma_p(\mathcal{V}) = \text{concat}\Big[\Phi_p^{(d)}(\mathcal{V}_{:,:,:,d})\Big], \tag{2}$$

where each depth-specific operator $\Phi_p^{(d)} : \mathbb{R}^{1 \times T \times H \times W} \rightarrow \mathbb{R}^{64 \times \frac{T}{2} \times \frac{H}{4} \times \frac{W}{4}}$ implements temporal-dominant 3D convolutions. The concatenation operator concat$[\cdot]$ preserves motion patterns across depth dimensions, which further yields Z_0.

The classical video architecture Uniformer [18] is employed as the backbone, where we extract the pre-logits motion queries $Z^{(s)} = \Phi_E^{(s)}(Z_0^{SA})$ from the student network's cine SA input, while aggregating the teacher network's long-axis features as $Z^{(t)} = \Phi_E^{(t)}([Z_0^{CH_2}, Z_0^{CH_3}, Z_0^{CH_4}])$.

To reconcile motion disparity and enforce patient-level alignment, we formulate a hybrid loss to enable the teacher network to be updated through Exponential Moving Average (EMA),

$$\mathcal{L}_{\text{StageI}} = \tau^2 D_{\text{KL}}\left(p^{(s)} | \bar{p}^{(t)}\right) + \lambda \mathbb{E}\left[\log \frac{\exp(\langle \mathbf{z}_i^{(s)}, \mathbf{z}_i^{(s,+)} \rangle / \tau_c)}{\sum\limits_j \exp(\langle \mathbf{z}_i^{(s)}, \mathbf{z}_j^{(s,-)} \rangle / \tau_c)}\right], \tag{3}$$

where $p^{(s)} = \text{Softmax}(Z^{(s)}/\tau), \bar{p}^{(t)} = \text{Softmax}(Z^{(t)}/\tau)$, and $\mathbf{z}_i^{(s)} = \frac{Z_i^{(s)}}{||Z_i^{(s)}||_2}$. The distillation term minimizes KL divergence between student predictions and teacher ensembles, while the contrastive term aligns SA features $\mathbf{z}_i^{(s)}$ with temporally synchronized positives $\mathbf{z}_i^{(s,+)}$ and repulsing negatives $\mathbf{z}_j^{(s,-)}$ from other patients in the same batch. This dual mechanism leverages anatomical consistency, strictly aligning motion trajectory, especially at the End-Diastole and End-Systole phases.

2.3 Stage II: Spatiotemporal Codebook Learning with Disentangled Representation

The temporal and spatial motion queries $Z_\tau^{(s)}$ and $Z_\sigma^{(s)}$ are derived from the trained encoder $\Phi_E^{(s)}$ in Stage I, where SA data are processed into two distinct 3D forms, (T, H, W) and (D, H, W), for temporal and spatial branches, respectively. This design yields disentangled motion features, which are quantized via separate codebooks. Let $\mathcal{C}_\tau, \mathcal{C}_\sigma \in \mathbb{R}^{n_e \times d_c}$ denote the temporal and spatial codebooks, respectively, with $n_e = 128$ and $d_c = 512$.

For each codebook entry $\mathrm{e}_k \in \mathcal{C} = \{\mathcal{C}_\tau, \mathcal{C}_\sigma\}$, quantized embeddings $Q = \{Q_\tau, Q_\sigma\}$ are obtained through

$$Q = \text{VecQuant}(Z^{(s)}, \mathcal{C}) = \sum_{k=1}^{n_e} \mathbb{I}\Big(k = \text{argmin}_{\text{j}} ||Z^{(s)} - \mathbf{e}_j||_2^2\Big)\mathbf{e}_k. \tag{4}$$

Represented query Q_{img} is obtained by cross attention through spatiotemporal quantization interaction,

$$Q_{img} = \text{Softmax}\Big(\frac{Q_\tau Q_\sigma}{\sqrt{d_c}}\Big)Q_\sigma \in \mathbb{R}^{N_\tau \times d_c}. \tag{5}$$

The joint optimization objective integrates codebook learning with spatiotemporal reconstruction

$$\mathcal{L}_{\text{StageII}} = ||\varPhi_D(Q_{\text{img}}, Z_l^{(s)}) - \mathcal{V}||_2^2 + \alpha\Big[||Z_\tau^{(s)} - \text{sg}(Q_\tau)||_2^2 + ||Z_\sigma^{(s)} - \text{sg}(Q_\sigma)||_2^2\Big], \tag{6}$$

where $\text{sg}(\cdot)$ denotes stop-gradient, $\varPhi_D$ denotes the reconstruction decoder, and α balances loss components. The dual codebook design $(\mathcal{C}_\tau, \mathcal{C}_\sigma)$ encourages resolving ambiguities where temporal blurring obscures spatial boundaries, such as the confusion between trabeculations and papillary muscles [4, 12], which often occurs in entangled encoding paradigms.

2.4 Survival Prediction Head

Clinical biomarkers $\mathbf{e} \in \mathbb{R}^{d_m}$ contained in tabular EHR data are exploited to formulate the multimodal fusion head with refined image features $\mathbf{q} = \mathbb{E}(Q_{img}) \in \mathbb{R}^{d_c}$. The fusion features $x_{\text{fused}} = \text{concat}[\mathbf{e}, \mathbf{q}] \in \mathbb{R}^{d_m + d_c}$ are then obtained.

A classical Cox head is defined, whose coefficients β_k automatically weight cross-modal interactions. The hazard function is subsequently defined as

$$h(t|\mathbf{x}_{\text{fused}}) = h_0(t)\text{exp}\Big(\sum_k \beta_k x_{\text{fused}_k}\Big). \tag{7}$$

The loss function minimizes negative log partial likelihood:

$$\mathcal{L}_{\text{Cox}} = -\sum_{i:\delta_i = 1}\Big[\theta^{\text{T}}\mathbf{x}_{\text{fused}}^{(\mathbf{i})} - \log\sum_{j \in R(t_i)} \exp(\theta^{\text{T}}\mathbf{x}_{\text{fused}}^{(j)})\Big] + \lambda||\theta||_2^2, \tag{8}$$

where $\theta = [\beta_1, ..., \beta_m]^{\text{T}}$, $\delta_i \in \{0, 1\}$ indicates event occurrence, and $R(t_i)$ is the risk set at time t_i.

3 Experiments

Datasets. Three in-house cardiac cine MRI datasets, i.e., RJCCM, AZCCM, and TJCCM, were utilized in the experiments. Each set includes four standardized views: short-axis, 2-chamber, 3-chamber, and 4-chamber orientations, comprising 407, 673, and 313 studies from patients, along with matched EHR data containing 135, 173, and 74 cardiovascular risk factors, respectively.

All sequences from the three datasets apply a magnetic field strength of 3.0 T and a 16-bit allocated intensity resolution for each image. Protocols across acquisition sites are variable. RJCCM employed a system with repetition time (TR) = 2.95-3.02 ms and echo time (TE) = 1.45-1.5 ms, capturing 30-phase cardiac cycles; AZCCM utilized a system with TR = 12.4-13.5 ms and TE = 1.55-1.61 ms with 25-phase cardiac cycles; TJCCM adopted a system with TR = 31.67-36.32 ms and TE = 1.39-1.41 ms with 25-phase cardiac cycles.

Evaluation Metrics. The concordance index [10] (C-index) was used in our experiments as a metric that accounts for both continuous and interval-based survival prediction models. It quantifies the prediction effect based on the number of correct pairs. We have

$$\text{C-index} = \frac{\sum_{i,j} \mathbb{I}(t_i < t_j)\mathbb{I}(r_i > r_j)\delta_i}{\sum_{i,j} \mathbb{I}(t_i < t_j)\delta_i}, \tag{9}$$

where δ_i indicates event occurrence, $r_i = \beta^{\mathrm{T}} \mathrm{x}^i_{\text{fused}}$ is the risk score.

Implementation Details. The proposed model was developed utilizing the PyTorch framework and trained on a single NVIDIA-H100 GPU with CUDA 12.2. The optimization process utilized the Adam optimizer [16], with a learning rate of 5×10^{-5} and weight decay set to 1×10^{-5}. A batch size of 16 was employed, and training was conducted over 50 epochs with the StepLR scheduler. To prevent overfitting, a penalizer was applied with values of $10^{-4}, 10^{-2}$, and 10^{-2}, with feature correlation thresholds of 0.7, 0.9, and 0.7 for the RJCCM, AZCCM, and TJCCM datasets, respectively. The 4D image inputs were resized to a resolution of $24 \times 24 \times 96 \times 96$, where the dimensions correspond to depth, frame, height, and width.

3.1 Experimental Results

Table 1 presents comparative results between classical and SOTA models. Our proposed CTSL demonstrates robust risk prediction capabilities across three cohorts, with C-index values of 0.788, 0.826, and 0.863, respectively, outperforming both the clinical-dependent model cluster, including CoxPH, DeepSurv, DSM, as well as the SOTA models SurvRNC and Sparse BagNet.

Detailed Kaplan-Meier survival analysis is provided in Fig. 2, with p-values incorporated from the log-rank test to highlight statistical significance. Comparisons include the clinical gold-standard CoxPH and the multimodal SurvRNC (top-performing baseline). CTSL achieves the lowest p-values on average, with

Table 1. Performance comparison across three datasets (Metric: C-index↑ (p-value↓)). The radiomics features extracted using PyRadiomics v3.1.0 [9] serve as substitutes of images for non-imaging models like DeepSurv and DSM.

Model	EHR	Img	Radiomics	RJCCM	AZCCM	TJCCM
CoxPH [5]	✓	-	-	0.638 (.259)	0.745 (.002)	0.562 (.212)
DeepSurv [15]	✓	-	✓	0.608 (.120)	0.618 (.109)	0.623 (.088)
DSM [21]	✓	-	✓	0.690 (.010)	0.632 (.191)	0.746 (.201)
SurvRNC [28]	✓	✓	-	0.731 (.014)	0.739 (.099)	0.648 (.064)
Sparse BagNet [8]	-	✓	-	0.568 (.109)	0.715 (.008)	0.545 (.244)
CTSL (Ours)	✓	✓	-	**0.788** (.074)	**0.826** (.036)	**0.863** (.029)

complete separation between high- and low-risk groups while no intersection of the curve is observed. Besides, increasingly pronounced prognostic differentiation is detected over time.

Interpretable Comparison. Clinically, high-density lipoprotein (HDL) levels, diabetes status, and stroke volume (SV) emerged as key clinical determinants of MACEs, exhibiting cross-center stability in contribution magnitude as Fig. 3 shows. The CTSL-derived imaging biomarkers revealed myocardial motion signatures with superior predictive value. Notably, cine motion-driven features, including wall motion scoring, end-systolic volume (ESV), and dual-chamber right atrial end-diastolic volume index (Dual RAEDVi), in synergy with imaging data, collectively demonstrated significant risk stratification power.

Ablation Study. To evaluate the robustness of our framework, we design ablation experiments at three levels: (1) Model CTSL, which is obtained through the complete workflow; (2) Model Uniformer(Distilled), whose representation aggregated from motion queries directly without employing the discrete spatiotemporal codebook for refinement; (3) Model Uniformer(ImageNet), which does not undergo distillation or codebook discretization, and instead relies solely on pretrained ImageNet [6] weights as a feature extractor.

Table 2. Results of ablation studies (Metric: C-index).

Model	Distillation	Quantization	RJCCM	AZCCM	TJCCM
Uniformer (ImageNet)	-	-	0.608	0.661	0.621
Uniformer (Distilled)	✓	-	**0.842**	0.754	0.648
CTSL	✓	✓	0.788	**0.826**	**0.863**

As evidenced in Table 2, our Stage I distillation framework demonstrates superior performance over natural image pre-trained counterparts through the synergistic integration of multi-view cardiac dynamics (mean ΔC-index: +0.118 vs. baselines). While achieving marginally lower performance in cohort RJCCM, potentially due to feature inconsistencies induced by motion artifacts, Stage II's disentangled spatiotemporal representations achieve statistically an overall performance improvement (mean ΔC-index: +0.078 vs. Stage I). This cross-

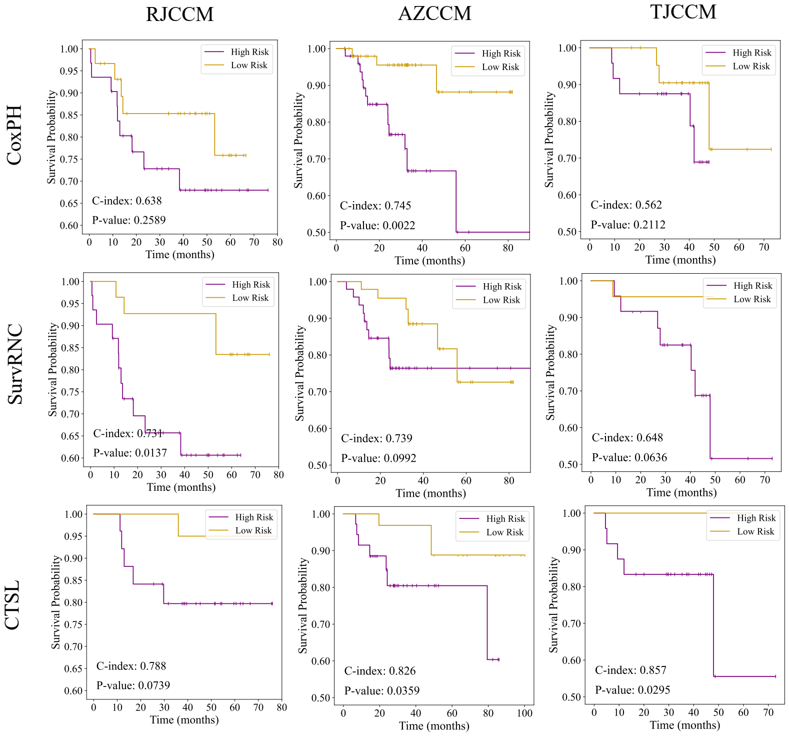


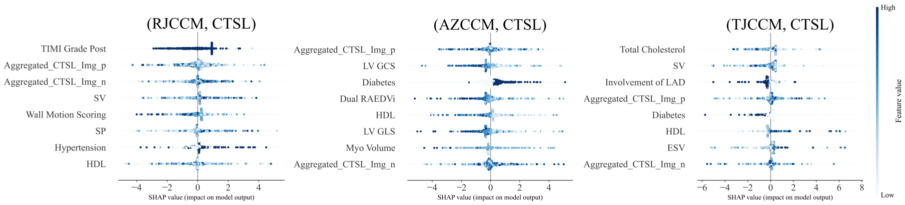
Fig. 2. Kaplan-Meier [14] analysis comparing risk stratification performance. Curves contrast our model against the clinical gold-standard CoxPH and SOTA SurvRNC baseline. Patients were stratified into high/low-risk groups by median predicted risk scores.

cohort consistency quantitatively validates the robustness of our latent space learning paradigm against anatomical variability.

4 Conclusion

This study introduces a self-supervised framework for non-contrast cardiac risk prediction, integrating motion-aware model distillation with codebook-based spatiotemporal disentanglement. By eliminating manual annotations, our approach effectively captures intrinsic myocardial dynamics. Experimental results demonstrate that CTSL not only enhances prognostic accuracy but also improves model interpretability by transforming raw 4D cine sequences into relevant image biomarkers. These findings highlight the potential of routine imaging for risk stratification, laying the groundwork for future advancements in personalized therapeutic planning through dynamic motion trajectory modeling.

Acknowledgments. This work was supported by the Shanghai Artificial Intelligence Laboratory, the National Natural Science Foundation of China (Grant Nos. 82171884



(a) (b) (c)

Fig. 3. SHAP-based interpretability analysis (a: RJCCM, b: AZCCM, c: TJCCM). Top-8 features are displayed for each dataset. The top-5 most prognostically influential imaging biomarkers with positive (Aggregated_CTSL_Img_p) and negative (Aggregated_CTSL_Img_n) contributions are aggregated, respectively.

and 82471931), and the Shanghai Municipal Commission of Science and Technology Medical Innovation Research Special Project (Grant No. 23Y11906900).

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

References

1. Baniecki, H., Sobieski, B., Szatkowski, P., Bombinski, P., Biecek, P.: Interpretable machine learning for time-to-event prediction in medicine and healthcare. Artificial Intelligence in Medicine **159**, 103026 (2025)
2. Bello, G.A., Dawes, T.J.W., Duan, J., Biffi, C., de Marvao, A., Howard, L.S., Gibbs, J.S.R., Wilkins, M.R., Cook, S.A., Rueckert, D., O'Regan, D.P.: Deep learning cardiac motion analysis for human survival prediction. Nature machine intelligence **1**, 95 – 104 (2018)
3. Bosco, E., Hsueh, L., McConeghy, K.W., Gravenstein, S., Saade, E.: Major adverse cardiovascular event definitions used in observational analysis of administrative databases: a systematic review. BMC Medical Research Methodology **21** (2021)
4. Chuang, M.L., Gona, P., Hautvast, G.L., Salton, C.J., Blease, S.J., Yeon, S.B., Breeuwer, M., O'Donnell, C.J., Manning, W.J.: Correlation of trabeculae and papillary muscles with clinical and cardiac characteristics and impact on cmr measures of lv anatomy and function. JACC: Cardiovascular Imaging **5**(11), 1115–1123 (2012)
5. Cox, D.R.: Regression models and life-tables. Journal of the Royal Statistical Society: Series B (Methodological) **34**(2), 187–202 (12 2018)
6. Deng, J., Dong, W., Socher, R., Li, L.J., Li, K., Fei-Fei, L.: Imagenet: A large-scale hierarchical image database. In: 2009 IEEE Conference on Computer Vision and Pattern Recognition. pp. 248–255 (2009)
7. Farnebäck, G.: Two-frame motion estimation based on polynomial expansion. In: Scandinavian Conference on Image Analysis (2003)
8. Gervelmeyer, J., Müller, S., Djoumessi, K., Merle, D., Clark, S.J., Koch, L., Berens, P.: Interpretable-by-design Deep Survival Analysis for Disease Progression Modeling . In: proceedings of Medical Image Computing and Computer Assisted Intervention – MICCAI 2024. vol. LNCS 15010. Springer Nature Switzerland (October 2024)

9. van Griethuysen, J.J.M., Fedorov, A., Parmar, C., Hosny, A., Aucoin, N., Narayan, V., Beets-Tan, R.G.H., Fillion-Robin, J.C., Pieper, S.D., Aerts, H.J.: Computational radiomics system to decode the radiographic phenotype. Cancer research **77 21**, e104–e107 (2017)
10. Harrell, J.F.E.: Regression modeling strategies: With applications to linear models, logistic regression, and survival analysis (2001)
11. Hou, W., He, Y., Yao, B., Yu, L., Yu, R., Gao, F., Wang, L.: Multi-scope analysis driven hierarchical graph transformer for whole slide image based cancer survival prediction. In: Medical Image Computing and Computer Assisted Intervention – MICCAI 2023. pp. 745–754. Springer Nature Switzerland, Cham (2023)
12. Inage, A., Mizuno, N.: Impacts of the systemic right ventricular trabeculae and papillary muscles on volumes and function assessed by novel magnetic resonance algorithm. Canadian Journal of Cardiology **31**(10, Supplement), S30 (2015), canadian Cardiovascular Congress 2015
13. Jaume, G., Vaidya, A., Chen, R.J., Williamson, D.F., Liang, P.P., Mahmood, F.: Modeling dense multimodal interactions between biological pathways and histology for survival prediction. In: 2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 11579–11590 (2024)
14. Kaplan, E.L., Meier, P.: Nonparametric estimation from incomplete observations. Journal of the American Statistical Association **53**, 457–481 (1958)
15. Katzman, J., Shaham, U., Cloninger, A., Bates, J., Jiang, T., Kluger, Y.: Deepsurv: personalized treatment recommender system using a cox proportional hazards deep neural network. BMC Medical Research Methodology **18** (2016)
16. Kingma, D.P., Ba, J.: Adam: A method for stochastic optimization. CoRR **abs/1412.6980** (2014)
17. Kullback, S., Leibler, R.A.: On information and sufficiency. Annals of Mathematical Statistics **22**, 79–86 (1951)
18. Li, K., Wang, Y., Peng, G., Song, G., Liu, Y., Li, H., Qiao, Y.: Uniformer: Unified transformer for efficient spatial-temporal representation learning. In: International Conference on Learning Representations (2022)
19. Meng, M., Bi, L., Fulham, M.J., wei Feng, D., Kim, J.: Merging-diverging hybrid transformer networks for survival prediction in head and neck cancer. ArXiv **abs/2307.03427** (2023)
20. Meng, M., Gu, B., Fulham, M., Song, S., Feng, D.D.F., Bi, L., Kim, J.: Adaptive segmentation-to-survival learning for survival prediction from multi-modality medical images. npj Precision Oncology **8** (10 2024)
21. Nagpal, C., Li, X., Dubrawski, A.: Deep survival machines: Fully parametric survival regression and representation learning for censored data with competing risks. IEEE Journal of Biomedical and Health Informatics **25**(8), 3163–3175 (2021)
22. van den Oord, A., Vinyals, O., Kavukcuoglu, K.: Neural discrete representation learning (2018)
23. Oquab, M., Darcet, T., Moutakanni, T., Vo, H., Szafraniec, M., Khalidov, V., Fernandez, P., Haziza, D., Massa, F., El-Nouby, A., Assran, M., Ballas, N., Galuba, W., Howes, R., Huang, P.Y., Li, S.W., Misra, I., Rabbat, M., Sharma, V., Synnaeve, G., Xu, H., Jegou, H., Mairal, J., Labatut, P., Joulin, A., Bojanowski, P.: Dinov2: Learning robust visual features without supervision (2024)
24. Rajiah, P., Francois, C., Leiner, T.: Cardiac mri: State of the art. Radiology **307**, 223008 (04 2023)
25. Ramanathan, V., Pati, P., McNeil, M., Martel, A.L.: Ensemble of prior-guided expert graph models for survival prediction in digital pathology. In: Medical Image

Computing and Computer Assisted Intervention – MICCAI 2024. vol. LNCS 15005, pp. 262 – 272. Springer Nature Switzerland (October 2024)

26. Razipour, A., Grodecki, K., Manral, N., Geers, J., Gransar, H., Shanbhag, A., Miller, R.J., Rozanski, A., Berman, D.S., Slomka, P.J., Dey, D.: Ai-derived automated quantification of cardiac chambers and myocardium from non-contrast ct: Prediction of major adverse cardiovascular events in asymptomatic subjects. Atherosclerosis **401**, 119098 (2025)
27. Rossello, X., González-Del-Hoyo, M.: Survival analyses in cardiovascular research, part i: the essentials. Revista Española de Cardiología (English Edition) **75**(1), 67–76 (2022)
28. Saeed, N., Ridzuan, M., Maani, F.A., Alasmawi, H., Nandakumar, K., Yaqub, M.: Survrnc: Learning ordered representations for survival prediction using rank-n-contrast. In: proceedings of Medical Image Computing and Computer Assisted Intervention – MICCAI 2024. vol. LNCS 15005. Springer Nature Switzerland (October 2024)
29. Zhao, L., Hou, R., Zhao, W., Qiu, L., Teng, H., Han, Y., Fu, X., Zhao, J.: Self-supervised learning guided transformer for survival prediction of lung cancer using pathological images. In: 2023 IEEE 20th International Symposium on Biomedical Imaging (ISBI). pp. 1–5 (2023)