Bridging Radiological Images and Factors with Vision-Language Model for Accurate Diagnosis of Proliferative Hepatocellular Carcinoma

Yanyan Huang[1⋆], Wanli Zhang[2⋆], Peixiang Huang[1], Yu Fu[3], Ruimeng Yang[2(✉)], and Lequan Yu[1(✉)]

[1] The University of Hong Kong, Hong Kong SAR, China
{yanyanh, paxson_huang}@connect.hku.hk, lqyu@hku.hk
[2] South China University of Technology, Guangzhou, China
zhangwanli2018@126.com, eyruimengyang@scut.edu.cn
[3] Lanzhou University, Lanzhou, China
fuyu@lzu.edu.cn

Abstract. The integration of multimodal data, particularly medical images and tabular data encompassing physician-assessed radiological factors, holds significant promise for enhancing clinical decision-making. However, effective fusion of these heterogeneous data modalities remains challenging due to their disparate feature spaces and the limitations of current independent encoding approaches. We introduce **FM-Bridge**, a novel methodology leveraging vision-language foundation model (VLM) to address this challenge. Our approach capitalizes on the intrinsic image-text embedding space alignment within VLMs to achieve robust multimodal fusion. We propose transforming clinical expertise-rich tabular data into semantically coherent textual descriptions, subsequently utilizing the VLM's text encoder to generate textual features explicitly aligned with image features. This method facilitates a more semantically congruent and effective fusion of medical image and tabular data, demonstrating potential for improved performance in downstream medical image analysis tasks compared to conventional methods. Code is available at https://github.com/HKU-MedAI/FM-Bridge.

Keywords: Multimodal Data Fusion · Medical Image Analysis · Tabular Data · Vision-Language Foundation Models.

1 Introduction

Proliferative hepatocellular carcinoma (HCC) poses a significant global health challenge, necessitating accurate and timely diagnosis for effective treatment planning and improved patient outcomes [28,13,1,25]. Medical imaging techniques, such as CT, are indispensable tools in HCC diagnosis, providing crucial visual information about lesions [22,15,24]. However, directly applying image

⋆ The first two authors contributed equally.

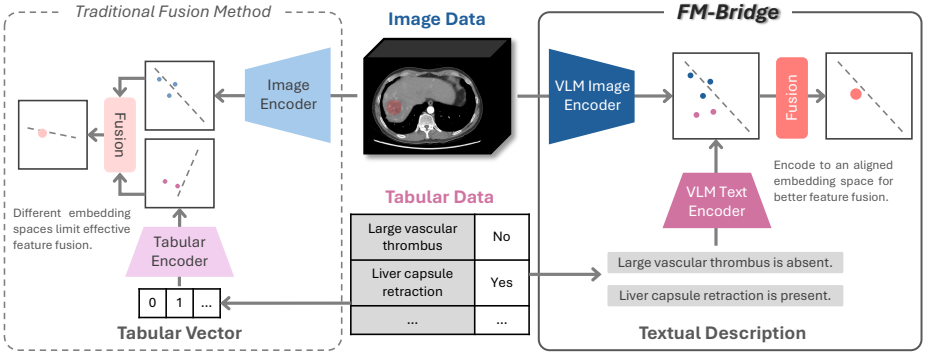

Fig. 1. The illustration of conventional multimodal fusion methods and our proposed FM-Bridge approach. Conventional independent encoding (left) results in disparate feature spaces and limited fusion. FM-Bridge (right) overcomes this by using a VLM to align image and text features derived from tabular data.

classification models to medical images for proliferative HCC diagnosis often leads to suboptimal outcomes. This is primarily due to the subtle visual cues and complex pathological features that characterize proliferative HCC, making it difficult to distinguish from other conditions based solely on images. To overcome these limitations, integrating multimodal data, particularly the synergistic combination of medical images and tabular data incorporating physician-assessed radiological factors, has emerged as a promising strategy [9,26,10,4]. Medical images can offer detailed visual representations of anatomical structures and indicate pathological alterations of lesions. Complementarily, tabular data, enriched with expert radiological factors, encapsulates valuable clinical expertise and experiential knowledge that is directly relevant to image interpretation. This combination promises a more comprehensive and robust diagnostic approach.

Despite the evident complementarity of these data sources, effectively fusing information from heterogeneous modalities remains a significant challenge. Existing methodologies frequently employ independent encoding pathways for image and tabular data, embedding them into distinct feature spaces before attempting fusion [19,18,6], as illustrated in the left part of Fig. 1. However, these conventional approaches often treat the two modalities as independent entities, failing to fully capitalize on the inherent inter-modality relationships and dependencies. Furthermore, the resulting disparate feature spaces can hinder truly effective and semantically rich information integration, limiting diagnostic accuracy.

To address these critical limitations and effectively leverage both image-based visual information and human expert knowledge for accurate proliferative HCC diagnosis, we propose **FM-Bridge**, a novel approach leveraging medical Vision-Language Foundation Models (VLMs) [27,11] to bridge the modality gap between radiological images and tabular data of radiological factors. Our core motivation is to leverage the intrinsic alignment between image and text embedding

spaces in VLMs, enabling the incorporation of pre-trained knowledge and facilitating robust multimodal fusion on the same semantic space. As shown in the right part of Fig. 1, FM-Bridge transforms clinical expertise-rich tabular data, encompassing physician-assessed radiological factors, into semantically coherent textual descriptions. These descriptions are then processed by the text encoder of a VLM, leveraging learnable prompts to generate textual features that are explicitly aligned with image features extracted by the VLM's vision encoder, which is also fine-tuned with vision prompt tuning [12]. This novel alignment process facilitates a more semantically congruent fusion of medical image and tabular data and enables the model to leverage the complementary information from both modalities. We demonstrate the effectiveness of FM-Bridge in a comprehensive evaluation on a private dataset of proliferative HCC diagnosis, showcasing its superior performance over conventional independent encoding strategies and underscoring its potential for advancing multimodal medical AI.

2 Method

2.1 Overview of FM-Bridge

The overview of the proposed FM-Bridge is depicted in Fig. 2. Our framework takes two primary inputs: a CT scan and pre-determined tabular radiological factors, which are assessments made by expert radiologists. Initially, these tabular factors are automatically transformed into natural language descriptions using predefined templates. Subsequently, these textual descriptions are input to the text encoder of a pre-trained VLM, while the corresponding CT scan is simultaneously processed by the VLM's image encoder. To preserve the VLM's pre-existing knowledge and the inherent alignment between its image and text feature spaces, we employ prompt learning [14,12,29,30] for both the image and text encoders instead of full fine-tuning. Finally, the resulting multimodal features are then integrated and fused to derive a comprehensive representation, which is used for proliferative HCC diagnosis prediction.

2.2 Textualization of Tabular Radiological Factors

To effectively integrate physician expertise encoded within tabular radiological factors, we textualized tabular features into natural language descriptions. Initially, experienced radiologists assessed predefined radiological factors for each CT scan and assigned binary scores (1 for presence, 0 for absence). This resulted in structured tabular data representing expert evaluations of visual characteristics. However, this tabular format is not directly interpretable by the text encoder component of VLMs, which are designed to process natural language. Therefore, we employed a rule-based textualization strategy to bridge this gap. This strategy converted each feature and its score into a sentence using predefined templates to indicate feature presence or absence. For instance, the feature "Liver capsule retraction" with a score of '1' was transformed into "Liver capsule retraction is present," while "Mosaic architecture" with a score of '0' became

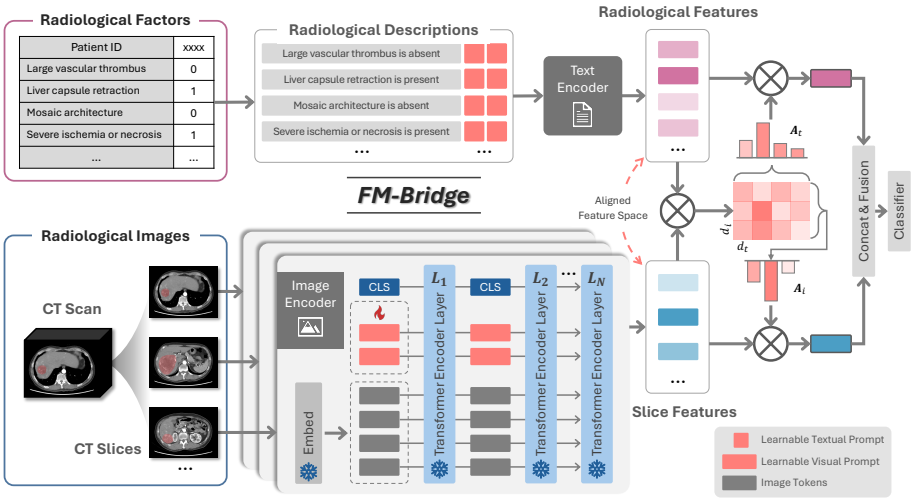

Fig. 2. The overview of the proposed FM-Bridge. The inputs are a CT scan and tabular radiological factors. It consists of textualization of tabular radiological factors, multimodal prompt learning, and multimodal feature fusion.

"Mosaic architecture is absent." Applying these templates consistently generated textual descriptions for each CT scan, encapsulating expert assessments in a language format. This textualization offers key advantages. First, it makes tabular data processable by VLM text encoders, leveraging their natural language understanding. Second, explicitly stating feature presence/absence in natural language conveys nuanced clinical information in a semantically rich format accessible to the model. This effectively injects expert knowledge from tabular data into the multimodal learning, contributing to more informed diagnostic predictions.

2.3 Multimodel Prompt Learning

To effectively adapt the pre-trained VLM for proliferative HCC diagnosis while leveraging its inherent knowledge, we employed multimodal prompt learning. Let f_{img} and f_{text} denote the image and text encoders of the VLM, respectively, with pre-trained parameters θ_{img} and θ_{text} (frozen during fine-tuning). The input CT scan $\boldsymbol{I} \in \mathbb{R}^{C\times H\times W\times L}$ is processed slice-by-slice. For the i-th slice, we divide it into M patches and extract patch embeddings $\boldsymbol{X}^i \in \mathbb{R}^{M\times D}$ using the frozen embedding layer. To guide visual feature extraction, we prepend a learnable class token $\boldsymbol{e}_{cls}$ and V learnable visual prompts $\boldsymbol{P}_v = \{\boldsymbol{p}_v^1, \boldsymbol{p}_v^2, \ldots, \boldsymbol{p}_v^V\}$ to the patch embeddings, forming the input to the image encoder:

$$\boldsymbol{E}_{\text{img}}^i = \{\boldsymbol{e}_{cls}, \boldsymbol{P}_v, \boldsymbol{X}^i\} \in \mathbb{R}^{(M+V+1)\times D}. \tag{1}$$

We utilize deep prompting, inserting these visual prompts at each layer of the image encoder, unlike shallow prompting which only prompts the first layer.

Deep prompting enhances the model's capacity to capture complex patterns in medical images. The image encoder then outputs the latent visual feature representation for each slice $\boldsymbol{Z}^i_{\text{img}} = f_{\text{img}}(\boldsymbol{E}^i_{\text{img}}; \theta_{\text{img}})$, and the CT scan's visual feature representation is the set of slice features $\hat{\boldsymbol{Z}}_{\text{img}} = \{\boldsymbol{Z}^1_{\text{img}}, \boldsymbol{Z}^2_{\text{img}}, \ldots, \boldsymbol{Z}^L_{\text{img}}\}$.

For the text encoder, the input is derived from the N textualized radiological factors $\boldsymbol{T} \in \mathbb{R}^N$ (obtained from tabular data as described in Section 2.2). Each textual description is tokenized and embedded to obtain text embeddings. For the j-th textual description, we denote its embedding sequence as $\boldsymbol{Y}^j = \{\boldsymbol{t}_{SOS}, \boldsymbol{t}^j_1, \boldsymbol{t}^j_2, \ldots, \boldsymbol{t}^j_N, \boldsymbol{t}_{EOS}\} \in \mathbb{R}^{(N+2)\times D}$, including start-of-sequence ($\boldsymbol{t}_{SOS}$) and end-of-sequence ($\boldsymbol{t}_{EOS}$) tokens. Furthermore, we insert T learnable textual prompts $\boldsymbol{P}_t = \{\boldsymbol{p}^1_t, \boldsymbol{p}^2_t, \ldots, \boldsymbol{p}^T_t\}$ into the text embeddings:

$$\boldsymbol{E}^j_{\text{text}} = \{\boldsymbol{t}_{SOS}, \boldsymbol{P}_t, \boldsymbol{t}^j_1, \boldsymbol{t}^j_2, \ldots, \boldsymbol{t}^j_N, \boldsymbol{t}_{EOS}\} \in \mathbb{R}^{(N+T+2)\times D}. \tag{2}$$

Similarly, the text encoder outputs the latent textual feature representation for each description $\boldsymbol{Z}^j_{\text{text}} = f_{\text{text}}(\boldsymbol{E}^j_{\text{text}}; \theta_{\text{text}})$, and the textual feature representation of the CT scan is the set of description features $\hat{\boldsymbol{Z}}_{\text{text}} = \{\boldsymbol{Z}^1_{\text{text}}, \boldsymbol{Z}^2_{\text{text}}, \ldots, \boldsymbol{Z}^N_{\text{text}}\}$.

2.4 Multimodal Feature Fusion and Objective Functions

To fuse visual and textual radiological factors for final diagnosis, we first compute a similarity matrix $\boldsymbol{S} \in \mathbb{R}^{L\times N}$ between them:

$$\boldsymbol{S} = \hat{\boldsymbol{Z}}_{\text{img}} \cdot \hat{\boldsymbol{Z}}^{\top}_{\text{text}} \in \mathbb{R}^{L\times N}. \tag{3}$$

This matrix captures the pairwise similarities between each CT slice's visual features and each textual radiological factor's embeddings. Next, we derive attention weights to emphasize relevant features. Textual attention weights $\boldsymbol{A}_{\text{text}} \in \mathbb{R}^N$ are computed by summing similarity scores for each textual feature across all slices and applying softmax normalization. Visual attention weights $\boldsymbol{A}_{\text{img}} \in \mathbb{R}^L$ are computed similarly, summing scores for each visual feature across all textual features and normalizing:

$$\boldsymbol{A}_{\text{text}} = \text{softmax}\left(\sum_{i=1}^{L} \boldsymbol{S}[i,:]\right), \quad \boldsymbol{A}_{\text{img}} = \text{softmax}\left(\sum_{j=1}^{N} \boldsymbol{S}[:,j]\right). \tag{4}$$

These attention weights reflect the importance of each feature in the multimodal context. We then aggregate textual and visual features using these weights to obtain attention-weighted representations:

$$\boldsymbol{Z}_{\text{text}} = \sum_{j=1}^{N} \boldsymbol{A}_{\text{text}}[j] \cdot \boldsymbol{Z}^j_{\text{text}}, \quad \boldsymbol{Z}_{\text{img}} = \sum_{i=1}^{L} \boldsymbol{A}_{\text{img}}[i] \cdot \boldsymbol{Z}^i_{\text{img}}. \tag{5}$$

Finally, we concatenate $\boldsymbol{Z}_{\text{text}}$ and $\boldsymbol{Z}_{\text{img}}$, and pass the concatenated vector through a linear layer and sigmoid activation to predict the proliferative HCC probability.

We employ binary cross-entropy loss for classification. To encourage diverse and informative textualized radiological factors, we add an orthogonal projection loss (OPL) $\mathcal{L}_{\text{ortho}}$ [23] as a regularization term. OPL encourages textual feature representations to be well-separated in the feature space, preventing redundancy and degradation of the textual features to a single point. The final objective function is:

$$\mathcal{L} = -\mathrm{E}\left(y_i \log(\hat{y}_i) + (1 - y_i) \log(1 - \hat{y}_i)\right) + \lambda \cdot \mathcal{L}_{\text{ortho}}, \tag{6}$$

where y_i and $\hat{y}_i$ are the ground-truth label and predicted probability for the i-th sample, and λ is a hyperparameter controlling the OPL's weight.

Table 1. Proliferative hepatocellular carcinoma diagnosis results with different methods. The best results are highlighted in bold.

Method	$\mathcal{I}$	$\mathcal{T}$	AUC	F_1
Tabular Data-only				
MLP		✓	0.668 ± 0.124	0.435 ± 0.127
FT-Transformer [5]		✓	0.572 ± 0.118	0.204 ± 0.142
TabPFN [8]		✓	0.627 ± 0.119	0.000 ± 0.000
Random Forest [21]		✓	0.677 ± 0.113	0.293 ± 0.181
XGBoost [2]		✓	0.667 ± 0.113	0.167 ± 0.161
CatBoost [20]		✓	0.683 ± 0.104	0.167 ± 0.167
CT Image-only				
ResNet-50 [7]	✓		0.655 ± 0.114	0.483 ± 0.128
ConvNeXt [17]	✓		0.692 ± 0.110	0.538 ± 0.133
ViT [3]	✓		0.693 ± 0.117	0.518 ± 0.126
Swin Transformer [16]	✓		0.687 ± 0.110	0.472 ± 0.146
Multi-model				
Addition	✓	✓	0.671 ± 0.110	0.426 ± 0.173
Concatenate	✓	✓	0.680 ± 0.108	0.451 ± 0.145
FiLM [18]	✓	✓	0.713 ± 0.112	0.533 ± 0.131
DAFT [19]	✓	✓	0.714 ± 0.111	0.423 ± 0.169
TabMixer [6]	✓	✓	0.715 ± 0.108	0.436 ± 0.172
FM-Bridge	✓	✓	**0.762 ± 0.102**	**0.567 ± 0.122**

3 Experiments and Results

3.1 Experimental Settings

Tasks and Evaluation. We evaluated FM-Bridge on a private dataset for proliferative HCC diagnosis collected from Guangzhou First People's Hospital. This dataset comprises CT images and corresponding tabular radiological factors (*e.g.*, tumor capsulate state, mosaic architecture presence) for 337 patients,

assessed by experienced radiologists. The task was to predict proliferative HCC diagnosis based on both CT images and radiological factors, with pathological diagnosis as the ground truth. The dataset includes 112 positive and 225 negative samples, with 104 randomly chosen patients reserved for testing and the remainder for training and validation. We provided two representative images of proliferative and non-proliferative HCC in Fig. 3. We used Area Under the ROC Curve (AUC) and F_1 score as primary evaluation metrics. To ensure robustness, we report 95% confidence intervals calculated via bootstrap resampling (1000 iterations).

Comparative Methods. For comprehensive evaluation, we compare the proposed FM-Bridge with different types of methods, including tabular data-only methods, image-only methods, and multi-model methods. *Tabular Data-only methods* included MLP, FT-Transformer [5], TabPFN [8], Random Forest [21], XGBoost [2], and CatBoost [20], which are widely used in tabular data analysis. *Image-only methods* comprised ResNet-50 [7], ConvNeXt [17], Vision Transformer (ViT) [3], and Swin Transformer [16]. For *Multimodal methods*, we included simple fusion strategies (Addition, Concatenation) and other state-of-the-art image-tabular fusion methods: FiLM [18], DAFT [19], and TabMixer [6]. For fair comparison in multimodal methods and FM-Bridge, we utilized the image encoder from GenMedClip [11] as the image backbone and its text encoder for encoding tabular features in FM-Bridge.

3.2 Experimental Results

Table 1 presents the experimental results. FM-Bridge achieves a superior AUC of 0.762 and F_1 score of 0.567, significantly outperforming all comparative methods. Both image-only and tabular data-only methods show reasonable performance, highlighting the importance of both modalities. However, simple multimodal fusion (Addition, Concatenation) underperforms, indicating that effective multimodal fusion is crucial. While methods like FiLM, DAFT, and TabMixer improve upon simple fusion, their performance remains limited due to separate encoding and less effective interaction mechanisms within misaligned embedding spaces. In contrast, the proposed FM-Bridge can effectively encode the image and tabular data into unified embedding space and learn the interaction between them more effectively, which leads to better performance. Moreover, FM-Bridge uniquely treats tabular data as an equally important input alongside images, rather than a secondary modality, allowing for better utilization of tabular information and contributing to its improved results.

3.3 Ablation Study

To further dissect the contribution of each component within FM-Bridge, we performed ablation studies by systematically removing individual components. Table 2 reveals that ablating either tabular data or image data results in a substantial performance decrease. This underscores the importance of both modalities for accurate diagnosis. Notably, even when relying solely on tabular data

Table 2. Ablation study of the proposed FM-Bridge on proliferative HCC diagnosis.

Method	$\mathcal{I}$	$\mathcal{T}$	AUC	F_1
FM-Bridge	✓	✓	**0.762 ± 0.102**	**0.567 ± 0.122**
- w/o Tabular	✓		0.669 ± 0.119	0.438 ± 0.150
- w/o Image		✓	0.695 ± 0.095	0.424 ± 0.113
- w/o OPL	✓	✓	0.726 ± 0.106	0.523 ± 0.129

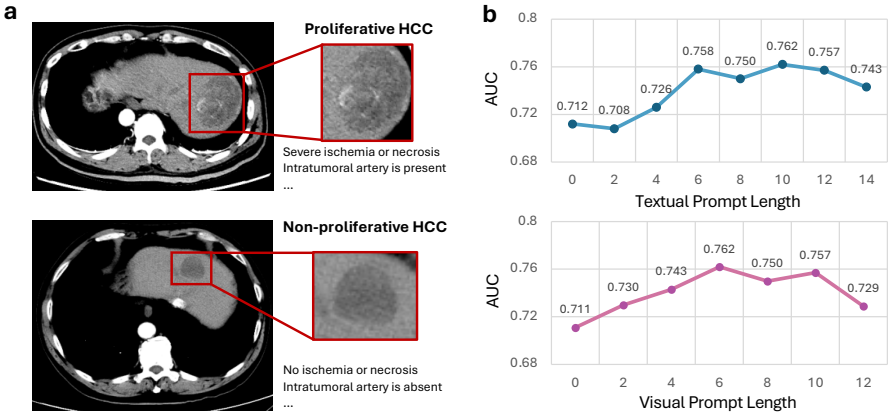

Fig. 3. (a) Images show typical slices of proliferative HCC and non-proliferative HCC, respectively. (b) Ablation study of the length of learnable visual and textual prompt in FM-Bridge.

(w/o Image), FM-Bridge still surpasses other tabular-only methods in AUC. This suggests that transforming tabular data into textual descriptions and encoding them with the text encoder not only preserves the original expressive power but may also enhance it by leveraging inherent knowledge within the foundation model. Furthermore, the Orthogonal Prompt Loss (OPL) also plays a critical role in performance, which could be attributed to its ability to prevent homogenization of the learnable textual prompt. The impact of varying learnable visual and textual prompt lengths is detailed in Fig. 3 b. Optimal performance was observed with a visual prompt length of 10 and a textual prompt length of 6. When the prompt lengths are too short, the model may not capture sufficient information from the input data, while excessively long prompts may introduce overfitting and hinder generalization.

4 Conclusion

This paper introduced FM-Bridge, a novel multimodal approach to proliferative HCC diagnosis explicitly designed to overcome the limitations of current methods in bridging the modality gap between medical images and expert-derived

radiological tabular data. By transforming tabular features into semantically coherent textual descriptions and employing prompt learning within a VLM framework, FM-Bridge achieves robust fusion of image and tabular representations. Our experimental results demonstrate that FM-Bridge significantly outperforms existing methods in proliferative HCC diagnosis, validating the effectiveness of our prompt-guided VLM approach for semantically congruent fusion. By moving beyond independent encoding strategies and fostering a more integrated approach, FM-Bridge represents a significant advancement in multimodal medical AI, offering a promising pathway for more accurate and clinically relevant diagnostic systems.

Acknowledgments. This work was supported in part by the Research Grants Council of Hong Kong (27206123, C5055-24G, and T45-401/22-N), the Hong Kong Innovation and Technology Fund (ITS/273/22 and GHP/318/22GD), the National Natural Science Foundation of China (No. 62201483), and Guangdong Natural Science Fund (No. 2024A1515011875).

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

References

1. Bao, Y., Li, J.X., Zhou, P., Tong, Y., Wang, L.Z., Chang, D.H., Cai, W.W., Wen, L., Liu, J., Xiao, Y.D.: Identifying proliferative hepatocellular carcinoma at pretreatment ct: implications for therapeutic outcomes after transarterial chemoembolization. Radiology **308**(2), e230457 (2023)
2. Chen, T., Guestrin, C.: Xgboost: A scalable tree boosting system. In: Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining. pp. 785–794 (2016)
3. Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., Uszkoreit, J., Houlsby, N.: An image is worth 16x16 words: Transformers for image recognition at scale. In: International Conference on Learning Representations (2021), `https://openreview.net/forum?id=YicbFdNTTy`
4. Fu, Y., Huang, Y., Zhang, Z., Dong, S., Xue, L., Niu, M., Li, Y., Shi, Z., Wang, Y., Zhang, H., et al.: Otfpf: Optimal transport based feature pyramid fusion network for brain age estimation. Information Fusion **100**, 101931 (2023)
5. Gorishniy, Y., Rubachev, I., Khrulkov, V., Babenko, A.: Revisiting deep learning models for tabular data. Advances in Neural Information Processing Systems **34**, 18932–18943 (2021)
6. Grzeszczyk, M.K., Korzeniowski, P., Alabed, S., Swift, A.J., Trzciński, T., Sitek, A.: Tabmixer: Noninvasive estimation of the mean pulmonary artery pressure via imaging and tabular data mixing. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 670–680. Springer (2024)
7. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 770–778 (2016)

8. Hollmann, N., Müller, S., Purucker, L., Krishnakumar, A., Körfer, M., Hoo, S.B., Schirrmeister, R.T., Hutter, F.: Accurate predictions on small data with a tabular foundation model. Nature **637**(8045), 319–326 (2025)
9. Huang, S.C., Pareek, A., Seyyedi, S., Banerjee, I., Lungren, M.P.: Fusion of medical imaging and electronic health records using deep learning: a systematic review and implementation guidelines. NPJ digital medicine **3**(1), 136 (2020)
10. Huang, W.: Multimodal contrastive learning and tabular attention for automated alzheimer's disease prediction. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 2473–2482 (2023)
11. Ikezogwo, W.O., Zhang, K., Seyfioglu, M.S., Ghezloo, F., Shapiro, L., Krishna, R.: Medicalnarratives: Connecting medical vision and language with localized narratives. arXiv preprint arXiv:2501.04184 (2025)
12. Jia, M., Tang, L., Chen, B.C., Cardie, C., Belongie, S., Hariharan, B., Lim, S.N.: Visual prompt tuning. In: European Conference on Computer Vision. pp. 709–727. Springer (2022)
13. Kang, H.J., Kim, H., Lee, D.H., Hur, B.Y., Hwang, Y.J., Suh, K.S., Han, J.K.: Gadoxetate-enhanced mri features of proliferative hepatocellular carcinoma are prognostic after surgery. Radiology **300**(3), 572–582 (2021)
14. Khattak, M.U., Rasheed, H., Maaz, M., Khan, S., Khan, F.S.: Maple: Multi-modal prompt learning. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 19113–19122 (2023)
15. Li, M., Fan, Y., You, H., Li, C., Luo, M., Zhou, J., Li, A., Zhang, L., Yu, X., Deng, W., et al.: Dual-energy ct deep learning radiomics to predict macrotrabecular-massive hepatocellular carcinoma. Radiology **308**(2), e230255 (2023)
16. Liu, Z., Lin, Y., Cao, Y., Hu, H., Wei, Y., Zhang, Z., Lin, S., Guo, B.: Swin transformer: Hierarchical vision transformer using shifted windows. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 10012–10022 (2021)
17. Liu, Z., Mao, H., Wu, C.Y., Feichtenhofer, C., Darrell, T., Xie, S.: A convnet for the 2020s. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 11976–11986 (2022)
18. Perez, E., Strub, F., De Vries, H., Dumoulin, V., Courville, A.: Film: Visual reasoning with a general conditioning layer. In: Proceedings of the AAAI conference on artificial intelligence. vol. 32 (2018)
19. Pölsterl, S., Wolf, T.N., Wachinger, C.: Combining 3d image and tabular data via the dynamic affine feature map transform. In: Medical Image Computing and Computer Assisted Intervention–MICCAI 2021: 24th International Conference, Strasbourg, France, September 27–October 1, 2021, Proceedings, Part V 24. pp. 688–698. Springer (2021)
20. Prokhorenkova, L., Gusev, G., Vorobev, A., Dorogush, A.V., Gulin, A.: Catboost: unbiased boosting with categorical features. Advances in neural information processing systems **31** (2018)
21. Qi, Y.: Random forest for bioinformatics. Ensemble machine learning: Methods and applications pp. 307–323 (2012)
22. Qu, H., Zhang, S., Guo, M., Miao, Y., Han, Y., Ju, R., Cui, X., Li, Y.: Deep learning model for predicting proliferative hepatocellular carcinoma using dynamic contrast-enhanced mri: Implications for early recurrence prediction following radical resection. Academic Radiology (2024)
23. Ranasinghe, K., Naseer, M., Hayat, M., Khan, S., Khan, F.S.: Orthogonal projection loss. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 12333–12343 (2021)

24. Ronot, M., Chernyak, V., Burgoyne, A., Chang, J., Jiang, H., Bashir, M., Fowler, K.J.: Imaging to predict prognosis in hepatocellular carcinoma: current and future perspectives. Radiology **307**(3), e221429 (2023)
25. Wang, G., Ding, F., Chen, K., Liang, Z., Han, P., Wang, L., Cui, F., Zhu, Q., Cheng, Z., Chen, X., et al.: Ct-based radiomics nomogram to predict proliferative hepatocellular carcinoma and explore the tumor microenvironment. Journal of Translational Medicine **22**(1), 683 (2024)
26. Wang, Z., Yu, L., Ding, X., Liao, X., Wang, L.: Shared-specific feature learning with bottleneck fusion transformer for multi-modal whole slide image analysis. IEEE Transactions on Medical Imaging **42**(11), 3374–3383 (2023)
27. Zhang, S., Xu, Y., Usuyama, N., Xu, H., Bagga, J., Tinn, R., Preston, S., Rao, R., Wei, M., Valluri, N., et al.: A multimodal biomedical foundation model trained from fifteen million image–text pairs. NEJM AI **2**(1), AIoa2400640 (2025)
28. Zhang, W., Li, N., Li, J., Zhao, Y., Long, Y., He, C., Zhang, C., Li, B., Zhao, Y., Lai, S., et al.: Noninvasive identification of proliferative hepatocellular carcinoma on multiphase dynamic ct: quantitative and li-rads lexicon-based evaluation. European Radiology pp. 1–16 (2024)
29. Zhou, K., Yang, J., Loy, C.C., Liu, Z.: Conditional prompt learning for vision-language models. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 16816–16825 (2022)
30. Zhou, K., Yang, J., Loy, C.C., Liu, Z.: Learning to prompt for vision-language models. International Journal of Computer Vision **130**(9), 2337–2348 (2022)