Towards Multi-dimensional Explanation Alignment for Medical Classification

Lijie Hu*, Songning Lai*, Wenshuo Chen*, Hongru Xiao, Hongbin Lin, Lu Yu, Jingfeng Zhang, Di Wang

King Abdullah University of Science and Technology

Abstract

The lack of interpretability in the field of medical image analysis has significant ethical and legal implications. Existing interpretable methods in this domain encounter several challenges, including dependency on specific models, difficulties in understanding and visualization, as well as issues related to efficiency. To address these limitations, we propose a novel framework called **Med-MICN** (**Med**ical **M**ulti-dimensional **I**nterpretable **C**oncept **N**etwork). Med-MICN provides interpretability alignment for various angles, including neural symbolic reasoning, concept semantics, and saliency maps, which are superior to current interpretable methods. Its advantages include high prediction accuracy, interpretability across multiple dimensions, and automation through an end-to-end concept labeling process that reduces the need for extensive human training effort when working with new datasets. To demonstrate the effectiveness and interpretability of Med-MICN, we apply it to four benchmark datasets and compare it with baselines. The results clearly demonstrate the superior performance and interpretability of our Med-MICN.

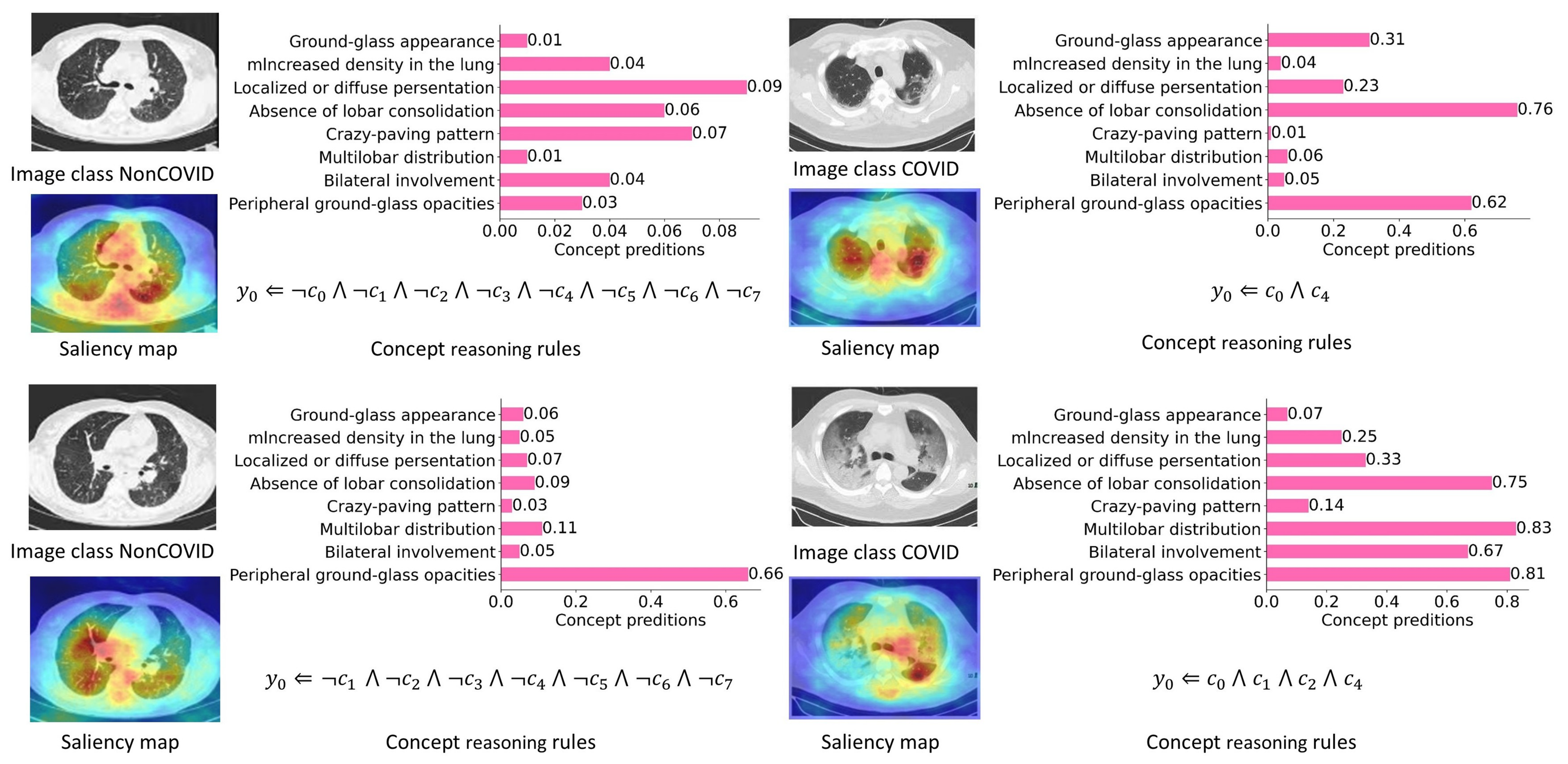

Figure 1: Med-MICN demonstrates multidimensional interpretability, encompassing concept score prediction, concept reasoning rules, and saliency maps, achieving alignment within the interpretative framework. The 'Peripheral ground-glass opacities' is c_0, and along the y-axis, it sequentially becomes $c_1, \ldots, c_7$.

Annotation Module Framework

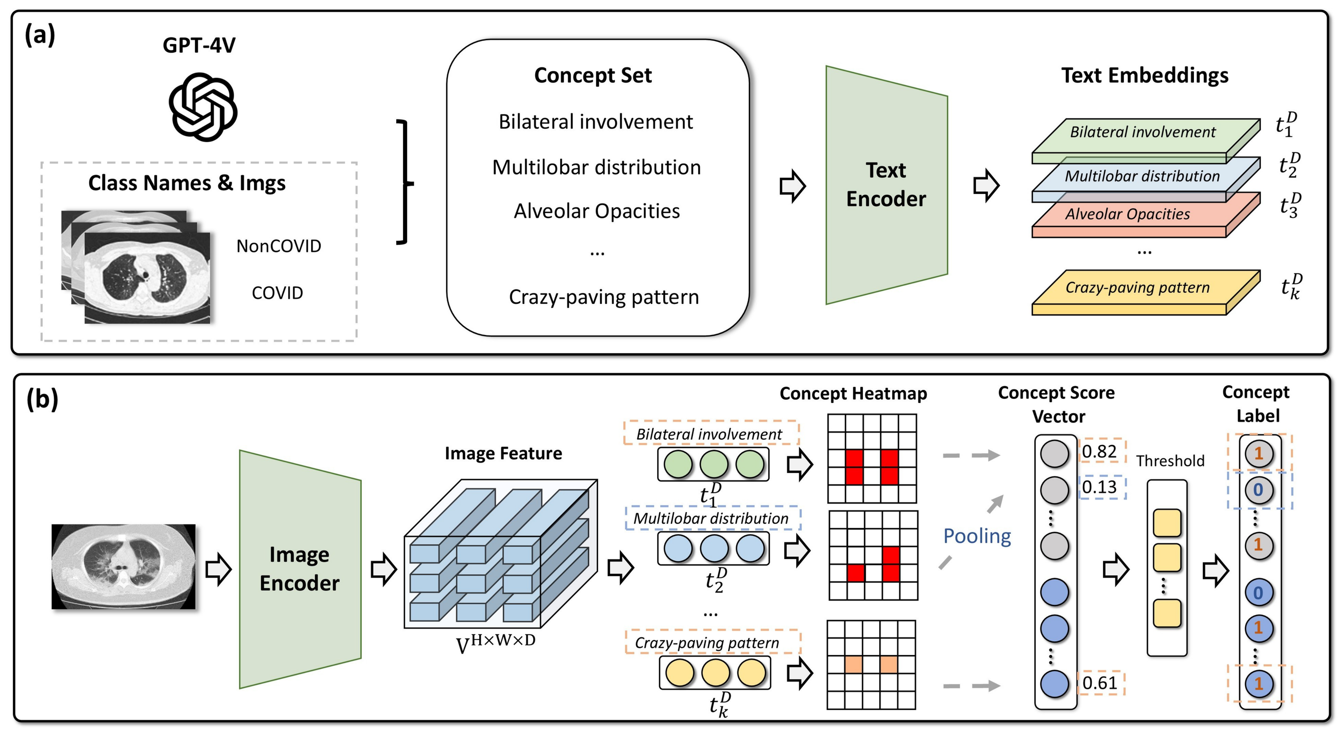

Figure 2: (a) module, output rich dimensional interpretable conceptual information for the specified disease through the multimodal model and convert the conceptual information into text vectors through the text embedding module; (b) module, access the image to the image embedder to get the image features, and then match with the conceptual textual information to get the relevant attention region. Then, we get the influence score of the relevant region information through pooling, and finally send it to the filter to sieve out the concept information with weak relevance to get the disease concept of image information.

Pipeline

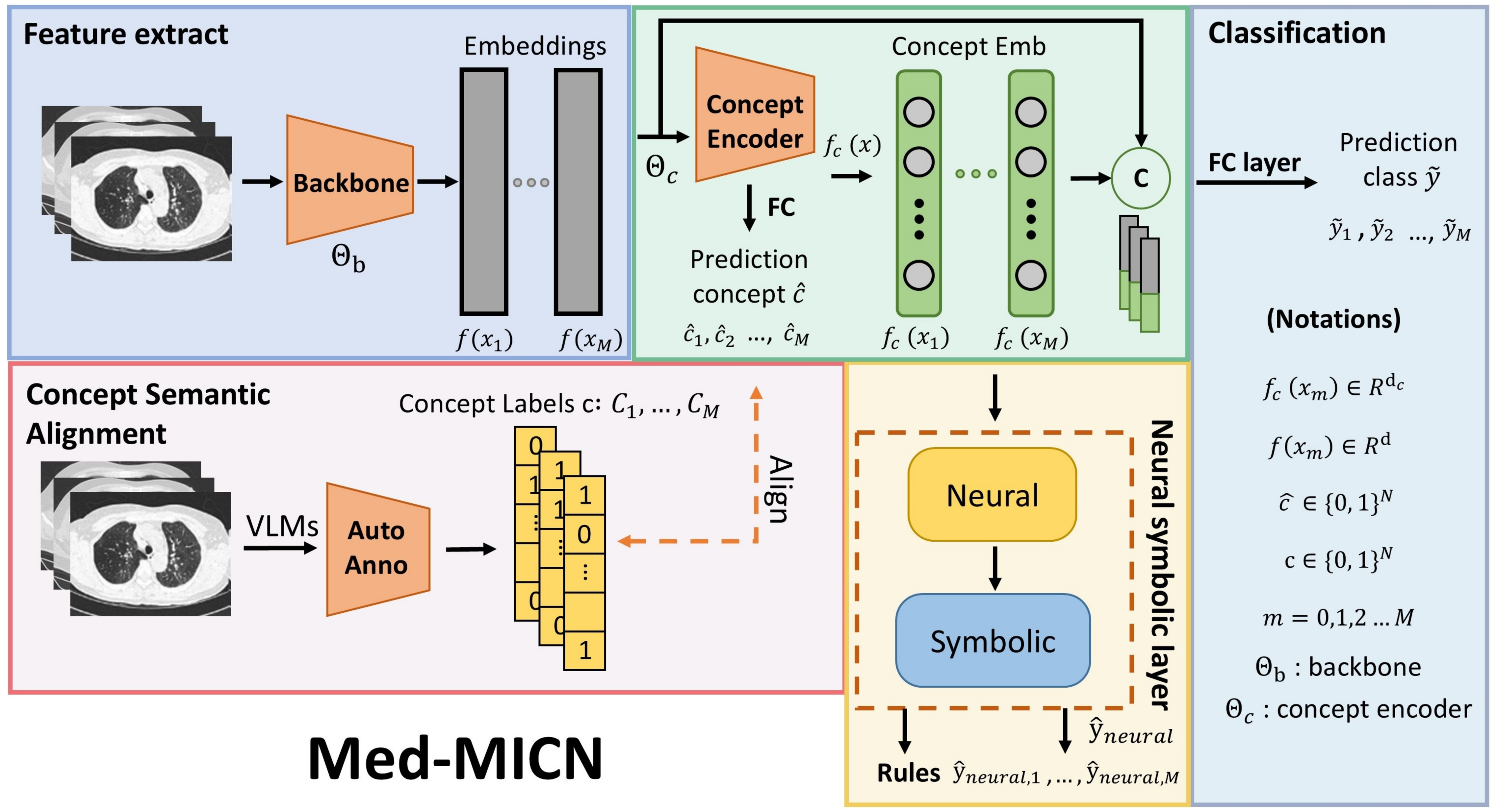

Figure 3: Overview of the Med-MICN Framework. The Med-MICN framework consists of four primary modules: (1) **Feature Extraction Module**: In the initial step, image features are extracted using a backbone network to obtain pixel-level features. (2) **Concept Embedding Module**: The extracted features are fed into the concept embedding module. This module outputs concept embeddings while passing through a category classification linkage layer to obtain predicted category information. (3) **Concept Semantic Alignment**: Concurrently, a Vision-Language Model (VLM) is used to annotate the image features, generating concept category annotations aligned with the predicted categories. (4) **Neural Symbolic Layer**: After obtaining the concept embeddings, they are input into the Neural Symbolic layer to derive conceptual rules. Finally, the concept embeddings obtained from module (2) are concatenated with the original image embeddings and fed into the final category prediction layer to produce the ultimate prediction results.

Stability Evaluation

Method	Backbone	Acc.(%)	Precision(%)	Recall(%)	F1(%)	AUC.(%)	Interpretability
Baseline	ResNet50	81.36	82.28	81.44	81.67	81.85	×
	VGG19	79.60	81.82	78.93	79.88	80.26	×
	DenseNet169	85.59	85.60	85.60	85.59	85.60	×
	SSSD-COVID	81.76	81.82	78.26	80.00	88.21	×
	Label Free CBM	69.49	68.62	69.82	69.21	64.84	✓
	DCR	55.93	58.38	55.43	51.41	55.43	✓
Ours	ResNet50	84.75	84.77	84.88	84.75	84.77	✓
	VGG19	83.05	86.74	82.93	84.37	84.26	✓
	DenseNet169	86.44	87.27	86.41	87.15	87.92	✓

Method	Backbone	Acc.(%)	Precision(%)	Recall(%)	F1(%)	AUC.(%)	Interpretability
Baseline	ResNet50	77.27	72.37	73.19	72.77	72.51	×
	VGG19	76.52	72.92	68.54	70.12	68.80	×
	DenseNet169	78.03	74.37	67.41	69.51	68.76	×
	Label Free CBM	70.34	68.62	69.82	69.21	69.49	✓
	DCR	76.52	71.79	63.88	65.32	63.88	✓
Ours	ResNet50	81.82	76.56	76.17	76.33	76.12	✓
	VGG19	82.58	81.59	76.05	78.07	75.63	✓
	DenseNet169	79.55	77.68	67.64	69.79	67.64	✓

Method	Backbone	Acc.(%)	Precision(%)	Recall(%)	F1(%)	AUC.(%)	Interpretability
Baseline	ResNet50	75.64	75.01	70.77	71.72	70.88	×
	VGG19	81.41	88.56	75.51	77.56	75.94	×
	DenseNet169	69.55	70.37	62.05	61.66	62.12	×
	Label Free CBM	71.21	71.89	71.45	70.84	74.12	✓
	DCR	62.02	66.25	51.50	41.33	50.56	✓
Ours	ResNet50	78.37	80.38	73.12	74.42	73.12	✓
	VGG19	88.30	92.59	85.43	88.16	87.09	✓
	DenseNet169	73.88	81.24	65.85	65.70	66.28	✓

Method	Backbone	Acc.(%)	Precision(%)	Recall(%)	F1(%)	AUC.(%)	Interpretability
Baseline	ResNet50	80.79	80.81	80.81	80.79	80.81	×
	VGG19	75.37	75.40	75.34	75.37	75.39	×
	DenseNet169	76.85	77.05	76.91	76.83	76.91	×
	Label Free CBM	75.24	75.15	74.92	75.41	75.02	✓
	DCR	68.05	67.55	65.33	66.12	67.01	✓
Ours	ResNet50	82.76	82.84	83.23	83.03	82.99	✓
	VGG19	77.34	77.72	77.33	77.53	77.58	✓
	DenseNet169	80.79	82.12	80.89	81.11	81.38	✓

Ablation Study

Dataset	Ablation Setting		Metrics					Interpretability
	$\mathcal{L}_c$	$\mathcal{L}_{neural}$	ACC.(%)	Precision(%)	Recall(%)	F1(%)	AUC.(%)	
COVID-CT			82.20	82.92	82.21	82.55	82.64	
	✓		83.05	83.62	83.16	83.01	83.16	
		✓	81.36	82.11	81.38	81.70	81.81	
	✓	✓	**84.75**	**84.77**	**84.88**	**84.75**	**84.77**	✓
DDI			78.03	74.97	66.88	69.24	67.41	
	✓		79.55	75.36	71.47	72.73	71.20	
		✓	78.79	76.38	66.29	68.69	67.64	
	✓	✓	**81.82**	**76.56**	**76.17**	**76.33**	**76.12**	✓
Chest X-Ray			68.59	69.63	61.11	61.02	62.05	
	✓		72.28	77.63	64.15	63.72	64.15	
		✓	70.03	73.83	61.84	61.25	62.39	
	✓	✓	**78.37**	**80.38**	**73.12**	**74.42**	**73.12**	✓
Fitzpatrick17k			78.33	79.50	78.32	78.91	79.06	
	✓		79.80	80.60	79.81	80.20	80.31	
		✓	80.79	81.28	80.82	81.28	81.07	
	✓	✓	**82.76**	**82.84**	**83.23**	**83.03**	**82.99**	✓

Table 1: Experimental results from ablation studies on each loss function demonstrate that each loss function is indispensable for both accuracy and interpretability.

Contact

Corresponding to lijie.hu@kaust.edu.sa