Question-Driven Decision Tree via Knowledge Distillation for Interpretable Prediction

Xuancheng Yao[1], Zhong Zhang[2], Qiuhui Chen[3], and Yi Hong[1]*

[1] School of Computer Science, Shanghai Jiao Tong University, Shanghai, China
[2] Shanghai General Hospital, Shanghai Jiao Tong University, Shanghai, China
[3] East China University of Science and Technology, Shanghai, China

Abstract. Interpretable models are important for high-stakes medical image recognition, but often underperform unconstrained black-box architectures. We propose Question-Driven Decision Tree (QDT), a framework that formulates prediction as structured reasoning through sequential and interpretable sub-questions. To retain expressive power under such constraints, QDT adopts a teacher–student paradigm in which a high-capacity teacher distills knowledge to a logic-constrained student while preserving an auditable decision rule. Experiments on challenging dermatological benchmarks, including two public datasets (ISIC-2019 and MILK10k) and a private pressure injury dataset, show that QDT surpasses existing interpretable models and achieves performance comparable to, and in some cases exceeding, strong black-box methods, while maintaining auditable question-level decision paths.

Keywords: Interpretable Model · Question-Driven Decision · Knowledge Distillation

1 Introduction

While deep learning architectures, ranging from convolutional neural networks [1, 2] to Vision Transformers (ViTs) [18], achieve expert-level accuracy in medical imaging [3, 4], their opaque nature impedes clinical integration. Medical workflows require verifiable decision processes. Therefore, predictive precision should be coupled with verifiable decision logic to foster clinical trust [6]. For instance, early explainable AI (XAI) methods primarily relied on post-hoc techniques, such as saliency mapping [7] and perturbation-based approximations [8]. However, these retroactive explanations often fail to capture the actual reasoning of the model and remain highly sensitive to minor input variations, risking misleading interpretations rather than providing genuine insights [9, 10]. This fundamental unreliability necessitates a shift toward models that are interpretable by design.

Concept Bottleneck Models (CBMs) [12] offer concept-level interpretability by constraining predictions through predefined human-understandable attributes. Although this creates an explicit reasoning layer, the flat concept representation contradicts the sequential and conditional nature of clinical diagnostic

* Corresponding author: yi.hong@sjtu.edu.cn

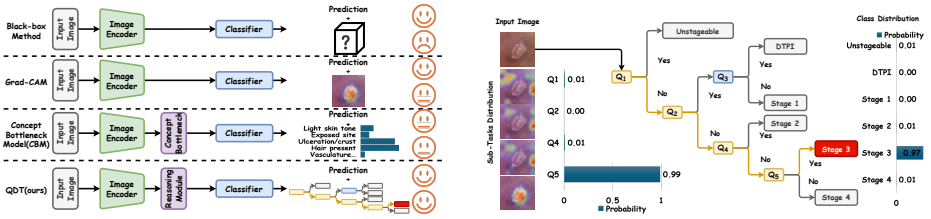

Fig. 1: Comparison of interpretability paradigms (left) and visualization of our QDT's structured reasoning process (right).

reasoning. Hence, CBMs struggle to capture subtle feature interactions, resulting in a trade-off where transparency comes at the cost of predictive performance.

Diverging from flat concept structures and opaque architectures (see the left subfigure of Fig. 1), we introduce the Question-Driven Decision Tree (QDT) framework. QDT formulates diagnosis as a hierarchical sequence of clinically meaningful sub-tasks, accurately mirroring established medical reasoning. However, enforcing such a deterministic clinical tree during training may constrain model capacity and limit flexible representation learning. To address this, we repurpose knowledge distillation [14] to transfer visual reasoning capabilities from a high-capacity teacher to a logic-constrained student via a teacher–student paradigm. By aligning intermediate representations [15, 19], the student inherits useful feature interactions modeled by the teacher, reducing the performance penalty associated with rule-constrained prediction.

Evaluated across diverse dermatological benchmarks (ISIC-2019 [16], MILK10-k [17], and a curated pressure injury dataset), our main contributions are:

- Formulating a question-driven decision tree paradigm that models predictions as a structured and clinically aligned reasoning process, overcoming the limitations of flat concept representations.
- Utilizing teacher–student distillation to transfer unconstrained visual knowledge to a logic-bound decision tree, bridging the gap between accuracy and interpretability.
- Demonstrating empirically that QDT generalizes as a framework across clinical datasets with clinically meaningful decompositions, outperforming existing interpretable baselines while maintaining auditable decision paths.

2 Method

Problem Formulation. Let $\mathcal{D} = \{(X_i, Y_i)\}_{i=1}^{N}$ denote a dataset of medical images, where $X_i \in \mathbb{R}^{H\times W\times 3}$ is an input image and Y_i consists of a primary diagnostic label and a set of supporting sub-task labels. Specifically, the primary diagnostic task is defined as $y_{\text{main}} \in \{1, \ldots, C\}$, where C denotes the number of diagnostic categories. In addition, each image is associated with K clinically meaningful sub-tasks, represented as $y_{\text{sub}} = (y_{\text{sub}_1}, \ldots, y_{\text{sub}_K})$, where each $y_{\text{sub}_k} \in \{0, 1\}$ corresponds to the binary outcome of a diagnostic question.

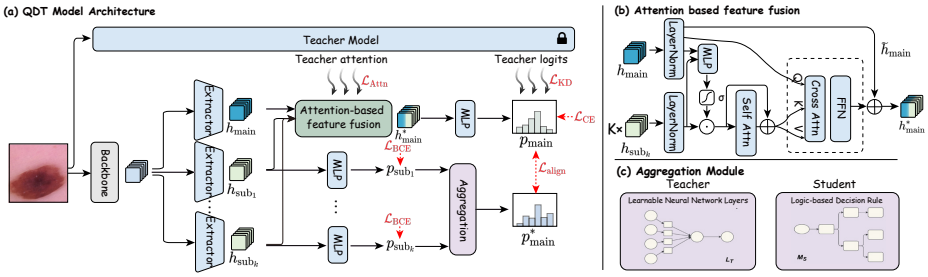

Fig. 2: Overview of QDT, showing multi-task feature extraction, attention-based fusion, and teacher–student aggregation for interpretable decisions.

Our goal is to learn an interpretable student model F_S that maps an input image X_i to a set of sub-task predictions p_{sub} and a final diagnostic prediction p_{main}. Crucially, the final diagnosis of F_S is constrained by a predefined and deterministic clinical logic tree M_S, such that $p_{\text{main}} = M_S(\hat{y}_{\text{sub}}), M_S : \{0,1\}^K \to \{1,\ldots,C\}$, $\hat{y}_{\text{sub}} = \mathbb{I}(p_{\text{sub}} > \delta)$. Here, $\mathbb{I}$ is the indicator function and we use a threshold $\delta = 0.5$ in all experiments. This formulation imposes a fundamental trade-off: enforcing the structured mapping M_S ensures transparency and verifiability, but often limits predictive performance. In contrast, an unconstrained teacher model F_T can achieve higher accuracy by directly optimizing p_{main} without structural constraints. The objective of our QDT is to learn F_S under the constraint induced by M_S, while minimizing the performance gap with F_T.

2.1 Model Architecture

As in Fig. 2, the teacher and student share a common backbone–head architecture, with distinct aggregation modules for combining sub-task predictions. The architecture consists of a shared feature encoder followed by multiple lightweight task-specific prediction heads. We use a CNN or Transformer as the backbone of the feature encoder B_{enc}. Given an input image $X_i \in \mathbb{R}^{H\times W\times 3}$, the backbone extracts a spatial feature representation $Z_i = B_{\text{enc}}(X_i)$.

Sub-task Heads. On top of the shared feature representation Z_i, the model includes K parallel sub-task heads $\{H_{\text{sub}_k}\}_{k=1}^{K}$, each corresponding to a clinically meaningful diagnostic question. Each sub-task head consists of a lightweight task-specific projection E_{sub_k} followed by a binary classifier C_{sub_k}, producing a probability $p_{\text{sub}_k} = C_{\text{sub}_k}(E_{\text{sub}_k}(Z_i)), p_{\text{sub}_k} \in [0,1]$.

Main Task Head. In addition, both the teacher and student models include a main task head $H_{\text{main}} = \{E_{\text{main}}, C_{\text{main}}\}$. The main head first extracts a main-task feature from the shared representation: $h_{\text{main}} = E_{\text{main}}(Z_i)$. Each sub-task head also produces an intermediate representation $h_{\text{sub}_k} = E_{\text{sub}_k}(Z_i)$. To allow the final diagnosis to leverage clinically relevant sub-task evidence, we fuse the main-task feature with the set of sub-task features via a fusion module $\Phi(\cdot)$: $h^*_{\text{main}} = \Phi\big(h_{\text{main}}, \{h_{\text{sub}_k}\}_{k=1}^{K}\big)$. The diagnostic logits are then obtained by $p_{\text{main}} = C_{\text{main}}(h^*_{\text{main}}), p_{\text{main}} \in \mathbb{R}^C$. Below is the detailed design of $\Phi(\cdot)$.

Attention-Based Feature Fusion. As shown in Fig. 2(b), sub-task representations are adaptively integrated into the main diagnostic stream through a gated dual-attention module. After layer normalization, $\tilde{h}_{\text{main}} = \text{LN}(h_{\text{main}})$ and $\tilde{h}_{\text{sub}_k} = \text{LN}(h_{\text{sub}_k})$, each sub-task feature is modulated by a learnable gate $\hat{g}_k = g_k/(\frac{1}{K}\sum_{j=1}^{K} g_j + \epsilon)$, $g_k = \sigma\Big(\text{MLP}\big([\tilde{h}_{\text{main}}, \tilde{h}_{\text{sub}_k}]\big)/\tau\Big)$, which controls its relative contribution. The gated sub-task features are first refined via self-attention with a residual connection, $h'_{\text{sub}} = \text{SelfAttn}\Big(\tilde{h}_{\text{sub}} \odot \hat{g}\Big) + \tilde{h}_{\text{sub}} \odot \hat{g}$, where $\tilde{h}_{\text{sub}} = [\tilde{h}_{\text{sub}_1}, \ldots, \tilde{h}_{\text{sub}_K}]$ denotes the stacked normalized sub-task representations and $\hat{g} = [\hat{g}_1, \ldots, \hat{g}_K]$ is the normalized gate vector broadcast along the feature dimension. Then, the refined sub-task features are injected into the primary stream through cross-attention, i.e., $h_{\text{cross}} = \text{CrossAttn}\Big(\tilde{h}_{\text{main}}, h'_{\text{sub}}, h'_{\text{sub}}\Big)$. A residual feed-forward block then produces the fused representation $h^*_{\text{main}} = \text{FFN}\big([\tilde{h}_{\text{main}}, h_{\text{cross}}]\big) + \tilde{h}_{\text{main}}$, integrating clinically relevant sub-task evidence while preserving main-task semantics.

Teacher–Student Distillation. The teacher model F_T is introduced to compensate for the limited expressiveness imposed by the deterministic logic tree. Although both teacher and student share the same multi-task backbone, they differ in sub-task aggregation (Fig. 2(c)). Specifically, F_T adopts a learnable MLP aggregator L_T, which directly maps sub-task probabilities to the primary diagnosis without structural constraints: $p^*_{\text{main}} = L_T(p_{\text{sub}_1}, \ldots, p_{\text{sub}_K})$. This unconstrained mapping allows complex interactions among sub-tasks to be modeled implicitly. In contrast, the student model F_S replaces L_T with a predefined decision rule M_S, which deterministically maps binarized sub-task predictions to the final diagnosis. While this design enforces transparency and ensures an auditable reasoning path, it restricts the hypothesis space available to the model. The practical formulation of M_S is detailed below using the Pressure Injury (PI) staging task as an example.

2.2 Clinically Guided Decomposition

The student's logic tree is constructed via a semi-automatic and clinically validated knowledge acquisition pipeline, with the number of questions determined by the task-specific taxonomy rather than fixed globally. Authoritative medical guidelines, such as the NPIAP protocol for pressure injury (PI) staging [24], serve as the primary source of diagnostic knowledge. Based on these guidelines, we use a large language model (LLM) to organize the staging criteria into a hierarchy of binary clinical questions, where each question corresponds to a medically meaningful decision point. Importantly, the diagnostic criteria are not generated by the LLM but are directly derived from the source guidelines. The resulting question hierarchy is then rigorously reviewed and refined by collaborating clinicians to ensure medical correctness, logical completeness, and consistency with clinical practice before being integrated into the model. This process yields a clinician-validated reasoning structure for interpretable PI staging.

As a result, the diagnostic process of PI staging decomposes into five sequential binary (*yes/no*) sub-tasks (Q_1–Q_5):

- Q_1: Is the wound obscured by slough or eschar?
- Q_2: Is the skin surface intact?
- Q_3: Are DTPI signs (e.g., deep red/purple discoloration) present?
- Q_4: Is the defect confined to the dermis, lacking subcutaneous exposure?
- Q_5: Is subcutaneous fat exposed without deeper structural involvement?

These cascaded sub-tasks formulate the deterministic decision rule $M(\cdot)$ (see the right subfigure of Fig. 1). Defining $p_{\text{sub}_k} = P(Q_k = 1 \mid X)$ as the affirmative probability for sub-task Q_k, the diagnostic distribution is computed via path multiplication: $P(\text{Unstageable}) = p_{\text{sub}_1}$, $P(\text{DTPI}) = (1 - p_{\text{sub}_1})\, p_{\text{sub}_2}\, p_{\text{sub}_3}$, $P(\text{Stage 1}) = (1 - p_{\text{sub}_1})\, p_{\text{sub}_2}\, (1 - p_{\text{sub}_3})$, $P(\text{Stage 2}) = (1 - p_{\text{sub}_1})\, (1 - p_{\text{sub}_2})\, p_{\text{sub}_4}$, $P(\text{Stage 3}) = (1 - p_{\text{sub}_1})\, (1 - p_{\text{sub}_2})\, (1 - p_{\text{sub}_4})\, p_{\text{sub}_5}$, and $P(\text{Stage 4}) = (1 - p_{\text{sub}_1})\, (1 - p_{\text{sub}_2})\, (1 - p_{\text{sub}_4})\, (1 - p_{\text{sub}_5})$. This continuous formulation allows gradients to propagate through all sub-task predictors during end-to-end training, computing the final diagnosis via soft path probabilities. Conversely, during inference, p_{sub_k} values are strictly thresholded, enforcing a deterministic tree traversal that ensures a transparent and auditable reasoning trajectory.

This general methodology extends to the ISIC-2019 [16] and MILK10k [17] datasets by using dataset-specific trees rather than a universal topology. By deriving criteria from established dermatological heuristics, such as the *ABCDE* rule for melanoma (Asymmetry, Border irregularity, Color variegation, Diameter and Elevation) [25], and dataset-specific label structure, domain-specific logic trees are analogously formulated. These auxiliary labels are used for training the sub-task heads; at test time, QDT requires only the input image and does not require metadata or manual answers to the questions. Consequently, this question-driven design gives the student an auditable decision path, coupling predictive accuracy with clinically verifiable rationales.

2.3 Two-Stage Training via Multi-Level Distillation

Stage 1: Training the Teacher Model. The teacher model F_T is first trained to achieve high diagnostic performance. Its objective function consists of supervised losses on both the main task and the clinically guided sub-tasks, together with an auxiliary alignment term that encourages coherence between them.

Main Task Loss. The main task head predicts the diagnostic class distribution $p_{\text{main}} \in \mathbb{R}^C$. We apply the standard multi-class cross-entropy loss: $\mathcal{L}_{\text{CE}} = -\sum_{c=1}^{C} y^c_{\text{main}} \log p^c_{\text{main}}$, where C is the number of diagnostic categories.

Sub-Task Loss. Each of the K sub-task heads performs binary classification. The corresponding loss is computed using binary cross-entropy: $\mathcal{L}_{\text{BCE}} = -\frac{1}{K} \sum_{k=1}^{K} \Big[y_{\text{sub}_k} \log p_{\text{sub}_k} + (1 - y_{\text{sub}_k}) \log(1 - p_{\text{sub}_k}) \Big]$.

Aggregation Alignment Loss. To encourage agreement between the main task prediction and the aggregation of sub-task outputs, the teacher employs a learnable aggregation module L_T to produce an auxiliary prediction p^*_{main}. We minimize the KL divergence between them: $\mathcal{L}_{\text{align}} = D_{\text{KL}}\big(p_{\text{main}} \,\|\, p^*_{\text{main}}\big)$.

The **overall teacher loss** is $\mathcal{L}_T = \mathcal{L}_{\text{CE}} + \lambda_{\text{sub}}\mathcal{L}_{\text{BCE}} + \lambda_{\text{align}}\mathcal{L}_{\text{align}}$, where λ_{sub} and λ_{align} are weighting coefficients.

Stage 2: Distilling Knowledge to the Student. After training, the teacher model F_T is frozen. The student model F_S is then optimized using a combination of hard-label supervision and soft distillation objectives.

Hard-Target Loss. While identical in form to the teacher's objective $\mathcal{L}_T$, this ground-truth supervision $\mathcal{L}_{\text{hard}}$ differs solely in the derivation of the auxiliary prediction p^*_{main}: the student employs the parameter-free, deterministic logic tree M_S rather than a learnable aggregator.

Soft-Target Distillation Loss. Knowledge is transferred from the teacher to the student through two complementary distillation objectives: $\mathcal{L}_{\text{soft}} = \lambda_{\text{kd}}\mathcal{L}_{\text{KD}} + \lambda_{\text{attn}}\mathcal{L}_{\text{Attn}}$, where λ_{kd} and λ_{attn} are weighting coefficients. *(1) Output-Level Distillation.* We align the student and teacher predictions at both the main-task and sub-task levels by minimizing the KL divergence between their softened output distributions: $\mathcal{L}_{\text{KD}} = D_{\text{KL}}\big(p^T_{\text{main}} \,\|\, p^S_{\text{main}}\big) + D_{\text{KL}}\big(p^T_{\text{sub}} \,\|\, p^S_{\text{sub}}\big)$. *(2) Attention Distillation.* To transfer visual reasoning cues, we distill the spatial attention maps produced by the attention modules of the teacher and student: $\mathcal{L}_{\text{Attn}} = \|A^S - A^T\|_2^2$, where A^T and A^S denote the attention maps of the teacher and student, respectively. The final student objective is: $\mathcal{L}_S = \mathcal{L}_{\text{hard}} + \mathcal{L}_{\text{soft}}$.

3 Experiments

Datasets and Experimental Setup. We evaluate QDT across three dermatological datasets: a private Pressure Injury (PI) dataset (2,292 images, 6 classes), the public ISIC-2019 benchmark [16] (25,331 images, 8 classes), and MILK10k [17] (5,240 images, 10 classes). All images are resized to 224×224 and augmented via random flipping, rotation, and color jittering. To ensure rigorous evaluation, we employ 5-fold cross-validation for the smaller PI dataset, while enforcing a strict 7:1:2 train-validation-test split for ISIC-2019 and MILK10k. Given the inherent class imbalances across these domains, performance is quantified using overall Accuracy, alongside weighted and macro-averaged F1-scores.

Executed on NVIDIA L40S GPUs, all models are optimized using AdamW (batch size 256, initial learning rate 1×10^{-4} with cosine annealing). Teacher training applies supervisory loss weights of $\lambda_{\text{sub}} = 1.0$ and $\lambda_{\text{align}} = 0.5$. Subsequent student distillation operates at a temperature $\tau = 1.0$, governed by distillation weights $\lambda_{\text{kd}} = 0.5$ and $\lambda_{\text{attn}} = 0.5$. Training concludes via early stopping (50-epoch patience) monitored on the validation F1-score.

Performance Comparison with Baselines. Table 1 details the quantitative performance of the QDT framework, where deployable student models (_S) systematically outperform existing interpretable-by-design baselines. Utilizing a ResNet50 backbone on the ISIC-2019 benchmark, QDT achieves an accuracy of 84.18%, decisively surpassing both CBM (80.97%) and Deep Concept Reasoner (DCR [23], 72.21%). This margin indicates that structuring clinical evidence into a hierarchical, conditional reasoning sequence effectively circumvents the representational bottlenecks inherent to flat concept mappings.

Table 1: Comparison of our QDT with baseline methods on three datasets using accuracy (**Acc.**), weighted F1 (**F1-W**), and macro F1 (**F1-M**). Model_T and Model_S denote the teacher (upper-bound reference only) and student (deployable logic-constrained) models. The best and second-best results are highlighted in **bold** and underlined. All values are percentages (%).

Baselines	**PI**			**ISIC-2019 [16]**			**MILK10k [17]**		
	Acc.↑	**F1-W↑**	**F1-M↑**	**Acc.↑**	**F1-W↑**	**F1-M↑**	**Acc.↑**	**F1-W↑**	**F1-M↑**
ResNet50 [2]	73.51	72.83	63.22	81.49	81.34	72.98	71.66	70.18	43.75
VGG16 [21]	72.09	71.29	61.69	78.55	78.34	68.02	68.90	67.05	38.80
DenseNet121 [22]	73.98	73.26	63.40	81.84	81.52	72.48	69.37	68.84	43.62
ViT-B/16 [18]	74.64	73.91	64.70	83.23	82.96	75.14	71.56	69.70	41.07
CBM [12]	72.24	71.11	61.06	80.97	80.57	71.93	70.42	68.70	41.60
DCR [23]	67.26	67.72	58.22	72.21	68.43	42.09	67.34	55.24	33.48
QDT (ours)	**Acc.↑**	**F1-W↑**	**F1-M↑**	**Acc.↑**	**F1-W↑**	**F1-M↑**	**Acc.↑**	**F1-W↑**	**F1-M↑**
ResNet50_T	74.84	74.23	72.34	**84.90**	**84.71**	75.51	72.80	70.38	42.32
ResNet50_S	74.32	73.30	71.64	84.18	83.91	78.17	72.32	69.99	42.71
VGG16_T	74.73	74.27	71.11	81.75	81.37	75.79	70.70	68.21	42.04
VGG16_S	74.21	73.54	71.15	81.40	81.34	74.38	69.37	68.63	42.80
DenseNet121_T	74.13	73.61	72.53	83.99	83.79	77.52	72.04	70.08	**45.95**
DenseNet121_S	74.08	73.32	71.35	83.48	83.15	75.58	72.32	**70.51**	43.18
ViT-B/16_T	**75.31**	**74.68**	73.39	83.71	83.46	74.48	**72.99**	69.84	42.61
ViT-B/16_S	75.04	74.45	**73.95**	83.15	82.89	**78.19**	72.61	69.60	39.18

Beyond specialized interpretable models, QDT shows robust architectural scalability by consistently improving upon the vanilla implementations of both classical CNNs (ResNet, VGG, DenseNet) and modern Vision Transformers (ViT-B/16). Notably, QDT variants exhibit pronounced enhancements in F1-Macro scores, evidenced by a 5.21% absolute increase over the standard ResNet50 on ISIC-2019. This consistent lift across diverse architectures suggests that granular sub-task supervision acts as a beneficial inductive bias, naturally regularizing the network against the severe class imbalances prevalent in medical datasets.

An internal comparison further elucidates how predictive power is preserved despite strict structural constraints. Unconstrained teacher models (_T), utilizing flexible MLP aggregation, establish the framework's empirical upper bounds. Through multi-level distillation, the logic-bound students (_S) closely track these peak metrics with negligible performance degradation (typically $< 1\%$). Consequently, the student successfully inherits complex visual feature interactions from the teacher, trading only a marginal fraction of predictive flexibility for an auditable and deterministic diagnostic path.

Ablation Studies. Table 2 details the component-wise contributions within the QDT framework, utilizing ResNet-50 as the baseline architecture. The architectural evolution reveals a clear trajectory of performance optimization bal-

Table 2: Ablation results (Accuracy %). STH: Sub-Task Head; AFM: Attention Fusion; AAL: Aggregation Alignment Loss; KD: Knowledge Distillation.

Modules	**PI**	**ISIC-2019**	**MILK10k**
Baseline(ResNet50)	73.51	81.49	71.66
+ STH	74.13	83.81	71.23
+ STH + AFM	73.75	84.09	71.85
+ STH + AFM + AAL	73.94	83.41	72.16
+ STH + AFM + AAL + KD (Full)	**74.32**	**84.18**	**72.32**

anced against interpretability constraints. Introducing explicit Sub-Task Heads (STH) immediately yields substantial gains on the ISIC-2019 benchmark (81.49% $\rightarrow$ 83.81%), highlighting the inherent value of granular clinical supervision. While an isolated STH introduces minor performance fluctuations on the PI and MILK10k datasets, potentially due to varying correlations among local features, integrating the Attention-based Fusion Module (AFM) stabilizes this behavior. By dynamically contextualizing sub-task evidence, the AFM collectively elevates the predictive baseline across diverse domains. The integration of training objectives exposes the fundamental trade-off inherent to structured reasoning. Enforcing deterministic logic via the Aggregation Alignment Loss (AAL) acts as a rigorous inductive bias. While occasionally serving as a beneficial regularizer (e.g., improving MILK10k to 72.16%), this strict constraint restricts expressive flexibility, evidenced by the accuracy reduction on ISIC-2019 (84.09% $\rightarrow$ 83.41%). However, this architectural cost for ensuring auditable trajectories is effectively neutralized by Knowledge Distillation (KD). By transferring soft probability distributions and spatial attention cues from the high-capacity teacher, KD compensates for the structural rigidity imposed by AAL. This final distillation phase systematically recovers prior performance dips, establishing peak accuracy across all three benchmarks and validating our core hypothesis.

Interpretability. The right subfigure of Fig. 1 illustrates the structured inference process of QDT on a Stage 3 pressure injury case. Instead of producing a single opaque prediction, the model traverses a deterministic sequence of clinically defined sub-tasks, each associated with a task-specific attention map. Early surface-level queries are confidently rejected, guiding the decision flow toward deeper assessments. At Q_5, a high affirmative probability (0.99) is assigned, with spatial attention concentrated on the lesion core, consistent with morphological criteria for Stage 3. The final prediction is obtained through rule-based aggregation of these binary decisions (probability 0.97). Across the PI dataset, sub-task classifiers achieve an average accuracy of 83.2%, indicating stable intermediate supervision. These results indicate that QDT enables verification of the decision process through exposed decision logic, confidence scores, and localized evidence, allowing alignment with established clinical guidelines, while the underlying CNN/Transformer feature extractor remains non-interpretable.

4 Conclusion

We presented **QDT**, a question-driven decision tree framework that unifies structured clinical reasoning with high-capacity representation learning. By coupling a deterministic logic tree with multi-level knowledge distillation, the proposed student model achieves competitive diagnostic performance while preserving an auditable question-level inference trajectory. Extensive evaluations across three dermatological benchmarks suggest that teacher-guided structured reasoning can partially alleviate the accuracy–interpretability trade-off.

Limitations and Future Work. Although this study focuses primarily on downstream task performance and qualitative interpretability analysis, comprehensive quantitative interpretability and fidelity metrics, as well as broader comparisons with recent neuro-symbolic methods, remain open challenges. Future work will investigate principled metrics for evaluating sub-task consistency and reasoning path stability, explore alternative tree topologies and automated tree construction strategies, and extend QDT to broader medical imaging domains.

Acknowledgments. This work was supported by the National Natural Science Foundation of China (NSFC) under Grant No. 62572308, and the Science and Technology Fund of Shanghai Jiao Tong University School of Medicine under Grant No. Jyhz2624.

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

References

1. Krizhevsky, A., Sutskever, I., Hinton, G.E.: Imagenet classification with deep convolutional neural networks. Advances in Neural Information Processing Systems **25** (2012)
2. 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)
3. Moor, M., Banerjee, O., Abad, Z.S.H., Krumholz, H.M., Leskovec, J., Topol, E.J., Rajpurkar, P.: Foundation models for generalist medical artificial intelligence. Nature **616**(7956), 259–265 (2023)
4. Esteva, A., Kuprel, B., Novoa, R.A., Ko, J., Swetter, S.M., Blau, H.M., Thrun, S.: Dermatologist-level classification of skin cancer with deep neural networks. Nature **542**(7639), 115–118 (2017)
5. Haenssle, H.A., Fink, C., Schneiderbauer, R., Toberer, F., Buhl, T., Blum, A., Kalloo, A., Hassen, A.B.H., Thomas, L., Enk, A., et al.: Man against machine: diagnostic performance of a deep learning convolutional neural network for dermoscopic melanoma recognition in comparison to 58 dermatologists. Annals of Oncology **29**(8), 1836–1842 (2018)
6. Ghassemi, M., Naumann, T., Schulam, P., Beam, A.L., Chen, I.Y., Ranganath, R.: A review of challenges and opportunities in machine learning for health. AMIA Summits on Translational Science Proceedings **2020**, 191 (2020)

7. Selvaraju, R.R., Cogswell, M., Das, A., Vedantam, R., Parikh, D., Batra, D.: Grad-cam: Visual explanations from deep networks via gradient-based localization. In: Proceedings of the IEEE International Conference on Computer Vision, pp. 618–626 (2017)
8. Ribeiro, M.T., Singh, S., Guestrin, C.: Why should I trust you? Explaining the predictions of any classifier. In: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp. 1135–1144 (2016)
9. Adebayo, J., Gilmer, J., Muelly, M., Goodfellow, I., Hardt, M., Kim, B.: Sanity checks for saliency maps. Advances in Neural Information Processing Systems **31** (2018)
10. Ghorbani, A., Abid, A., Zou, J.: Interpretation of neural networks is fragile. In: Proceedings of the AAAI Conference on Artificial Intelligence **33**(01), 3681–3688 (2019)
11. Rudin, C.: Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. Nature Machine Intelligence **1**(5), 206–215 (2019)
12. Koh, P.W., Nguyen, T., Tang, Y.S., Mussmann, S., Pierson, E., Kim, B., Liang, P.: Concept bottleneck models. In: International Conference on Machine Learning, pp. 5338–5348. PMLR (2020)
13. Wiggins, W.F., Tejani, A.S.: On the opportunities and risks of foundation models for natural language processing in radiology. Radiology: Artificial Intelligence **4**(4), e220119 (2022)
14. Hinton, G., Vinyals, O., Dean, J.: Distilling the knowledge in a neural network. arXiv preprint arXiv:1503.02531 (2015)
15. Zagoruyko, S., Komodakis, N.: Paying more attention to attention: Improving the performance of convolutional neural networks via attention transfer. arXiv preprint arXiv:1612.03928 (2016)
16. Combalia, M., Codella, N.C.F., Rotemberg, V., Helba, B., Vilaplana, V., Reiter, O., Carrera, C., Barreiro, A., Halpern, A.C., Puig, S., et al.: Bcn20000: Dermoscopic lesions in the wild. arXiv preprint arXiv:1908.02288 (2019)
17. Philipp, T., Akay, N.B., Rosendahl, C., Rotemberg, V., Todorovska, V., Weber, J., Wolber, A.K., Müller, C., Kurtansky, N., Halpern, A., et al.: MILK10k: A hierarchical multimodal imaging learning toolkit for diagnosing pigmented and non-pigmented skin cancer and its simulators. Journal of Investigative Dermatology (2025)
18. Dosovitskiy, A.: An image is worth 16x16 words: Transformers for image recognition at scale. arXiv preprint arXiv:2010.11929 (2020)
19. Romero, A., Ballas, N., Kahou, S.E., Chassang, A., Gatta, C., Bengio, Y.: Fitnets: Hints for thin deep nets. arXiv preprint arXiv:1412.6550 (2014)
20. Oquab, M., Darcet, T., Moutakanni, T., Vo, H., Szafraniec, M., Khalidov, V., Fernandez, P., Haziza, D., Massa, F., El-Nouby, A., et al.: Dinov2: Learning robust visual features without supervision. arXiv preprint arXiv:2304.07193 (2023)
21. Simonyan, K., Zisserman, A.: Very deep convolutional networks for large-scale image recognition. arXiv preprint arXiv:1409.1556 (2014)
22. Huang, G., Liu, Z., Van Der Maaten, L., Weinberger, K.Q.: Densely connected convolutional networks. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 4700–4708 (2017)
23. Barbiero, P., Ciravegna, G., Giannini, F., Zarlenga, M.E., Magister, L.C., Tonda, A., Lió, P., Precioso, F., Jamnik, M., Marra, G.: Interpretable neural-symbolic concept reasoning. In: International Conference on Machine Learning, pp. 1801–1825. PMLR (2023)

24. National Pressure Injury Advisory Panel, European Pressure Ulcer Advisory Panel, Pan Pacific Pressure Injury Alliance: Prevention and Treatment of Pressure Ulcers/Injuries: Clinical Practice Guideline. The International Guideline. 4th edn. EPUAP/NPIAP/PPPIA, Washington DC, Brussels, Perth (2025). https://internationalguideline.com
25. American Academy of Dermatology Ad Hoc Task Force for the ABCDEs of Melanoma, Tsao, H., Olazagasti, J.M., Cordoro, K.M., Brewer, J.D., Taylor, S.C., Bordeaux, J.S., Chren, M.M., Sober, A.J., Tegeler, C., Bhushan, R., Begolka, W.S.: Early detection of melanoma: reviewing the ABCDEs. J. Am. Acad. Dermatol. **72**(4), 717–723 (2015). https://doi.org/10.1016/j.jaad.2015.01.025