



OPEN ACCESS

EDITED BY

Ravi Prakash Sahu,
Wright State University, United States

REVIEWED BY

Kanjar De,
Interactive Institute, Sweden
Yazeed AlZahrani,
Prince Sattam Bin Abdulaziz University, Saudi
Arabia

*CORRESPONDENCE

Liao Hu
✉ lhu231@my.trine.edu
Boqian Hu
✉ h bq199811@163.com

[†]These authors share first authorship

RECEIVED 03 December 2025

REVISED 19 January 2026

ACCEPTED 23 January 2026

PUBLISHED 05 March 2026

CITATION

Zheng P, Yang J, Wen X and Hu B (2026) A novel multi-agent spatiotemporal fusion framework for intelligent skin cancer diagnosis. *Front. Oncol.* 16:1759960. doi: 10.3389/fonc.2026.1759960

COPYRIGHT

© 2026 Zheng, Yang, Wen and Hu. This is an open-access article distributed under the terms of the [Creative Commons Attribution License \(CC BY\)](https://creativecommons.org/licenses/by/4.0/). The use, distribution or reproduction in other forums is permitted, provided the original author(s) and the copyright owner(s) are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.

A novel multi-agent spatiotemporal fusion framework for intelligent skin cancer diagnosis

Peiyao Zheng^{1†}, Jin Yang^{2†}, Xuanru Wen³ and Boqian Hu^{4*}

¹College of Traditional Chinese Medicine, Hubei University of Chinese Medicine, Hubei, China,

²Baotou Medical College, Inner Mongolia University of Science & Technology, Inner Mongolia, China,

³College of Inspection, Hebei University of Chinese Medicine, Hubei, China, ⁴Hubei Provincial Hospital of Traditional Chinese Medicine, Hubei, China

Introduction: Skin cancer is one of the most common malignancies worldwide, and early-stage diagnosis remains challenging due to its morphological similarity to benign lesions. Most existing computer-aided diagnostic systems rely on single static images, overlooking temporal information that is critical for distinguishing progressive malignancy.

Methods: We propose a novel multi-agent spatiotemporal fusion framework to enhance diagnostic accuracy. The framework consists of three key components: (1) a spatial agent based on a convolutional neural network for high-fidelity static feature extraction; (2) a temporal agent employing gated recurrent units to model longitudinal lesion evolution; and (3) a collaboration agent that dynamically fuses spatial and temporal representations via an attention-based weighting strategy.

Results: Experiments on large-scale public dermoscopic datasets showed that our method achieved an accuracy of 94.5%, an F1-score of 93.8%, and an AUC of 0.97—outperforming traditional machine learning models, CNN classifiers, and 3D-CNN baselines. Ablation studies further confirmed the critical contribution of temporal modeling and adaptive fusion, particularly in differentiating early melanoma from atypical nevi.

Discussion: This work highlights the potential of spatiotemporal modeling to improve early skin cancer detection and provides a promising direction for AI-assisted diagnosis of other chronic diseases requiring longitudinal monitoring.

KEYWORDS

deep learning, intelligent diagnosis, multi-agent collaboration, skin cancer, spatiotemporal fusion

1 Introduction

Skin cancer is one of the most prevalent malignancies worldwide, with a steadily increasing incidence and significant public health burden. Early and accurate diagnosis is crucial, as timely excision of malignant lesions dramatically improves patient prognosis, whereas delayed detection often results in metastatic disease and increased mortality (1).

However, differentiating malignant melanoma and other skin cancers from benign lesions such as nevi, seborrheic keratosis, or dermatofibroma remains challenging, even for experienced dermatologists, due to high visual similarity and overlapping morphologic features (2). The need for reliable, reproducible, and scalable diagnostic tools has driven substantial interest in computer-aided diagnosis (CAD) and artificial intelligence (AI)-based systems.

Early CAD systems relied heavily on handcrafted image features such as color histograms, texture descriptors, and shape metrics, combined with classical classifiers including support vector machines (SVMs) (3). For instance, local binary patterns (LBP) were widely used to encode lesion texture information, which, combined with SVM, achieved reasonable accuracy for binary melanoma classification (4). Similar approaches exploited both color and texture features to improve discrimination (5, 6), and even introduced clinically inspired generative models for better interpretability (7). While these methods established the foundation for automated skin lesion analysis, they were limited by feature engineering subjectivity, insufficient generalization across diverse populations, and sensitivity to imaging artifacts (8).

The advent of deep learning (9–14), particularly convolutional neural networks (CNNs) (15), revolutionized skin cancer image classification. Esteva et al. demonstrated that a CNN trained end-to-end on a large dermoscopic dataset could perform on par with board-certified dermatologists, establishing the deep learning paradigm as a new state of the art (16). Subsequent studies explored deeper architectures (e.g., ResNet-50, Inception-v3) and ensemble strategies to further boost accuracy on benchmarks such as ISIC-2018 (17, 18), while recent advances have seen the rise of more powerful models including EfficientNet (19), Vision Transformer (ViT) (20), and dedicated skin lesion diagnostic models like DermNetCNN (21). These state-of-the-art (SOTA) methods achieved competitive AUROC scores (0.93–0.95) in large-scale challenges such as the ISIC 2021 Melanoma Classification Challenge (22), outperforming traditional pipelines and early CNNs by a wide margin. Hybrid models combining CNNs with texture features such as LBP have also been proposed to capture fine-grained lesion patterns (23).

Despite these advances, several critical limitations remain before AI can be fully integrated into clinical workflows. First, most models—including recent SOTA approaches (19, 20, 22)—still analyze individual dermoscopic images as independent samples, ignoring longitudinal patient data or lesion evolution patterns over time (24). However, temporal changes such as progressive enlargement, color darkening, or structural irregularity are key diagnostic clues that dermatologists rely on to distinguish early malignant transformation. Neglecting temporal context may lead to misclassification of early melanoma or dysplastic nevi that exhibit subtle progression.

Second, existing models often lack robustness across diverse patient populations and imaging conditions (25), and few have been validated for clinical utility via statistical tests or decision curve

analysis (DCA) (26). Most benchmark datasets are curated and noise-free, whereas real-world clinical images include artifacts (rulers, hair, gel bubbles) that can mislead models (27). Moreover, many studies report retrospective results without prospective validation, and few evaluate model calibration or clinical decision impact (28), limiting their translational potential.

Third, while spatiotemporal modeling approaches such as 3D-CNNs or hyperspectral imaging (HSI) have been explored, they face practical challenges such as high computational cost, data scarcity, and overfitting risk (29, 30). 3D total-body photography and volumetric CNNs have shown promise for population-level screening but require expensive equipment and standardized imaging protocols (31, 32). Additionally, few studies systematically compare the impact of different agent architectures, fusion mechanisms, or agent numbers on diagnostic performance, leaving the design rationale of multi-component frameworks undervalued. Consequently, there is a pressing need for efficient, generalizable, and context-aware frameworks that integrate both spatial and temporal lesion information without prohibitive data or hardware requirements, while also providing rigorous statistical and clinical validation.

Recent advances in reinforcement learning (RL) have shown strong potential in sequential decision-making tasks, including medical image navigation and region-of-interest localization (33–39). Inspired by this, we propose a novel multi-agent spatiotemporal fusion framework for intelligent skin cancer diagnosis. Our approach emulates the clinical reasoning process by jointly analyzing static lesion morphology and its temporal evolution. Specifically, a spatial agent extracts high-fidelity visual biomarkers using a CNN backbone, a temporal agent models lesion progression via gated recurrent units (GRUs) applied to feature sequences, and a collaborative decision agent adaptively fuses the outputs of both agents through an attention-based weighting mechanism. Unlike naïve feature concatenation, our dynamic fusion prioritizes temporal cues when rapid malignant changes are detected, thereby enhancing early detection sensitivity while maintaining specificity for stable lesions.

Our main contributions are threefold: (1) We introduce a novel multi-agent architecture that explicitly models both spatial and temporal dimensions, bridging the gap between single-image classifiers and longitudinal clinical assessment. (2) We design a dynamic attention-based fusion mechanism that adaptively balances spatial and temporal information, improving robustness and clinical interpretability, with expanded ablation studies validating the impact of different agent architectures, numbers, and fusion strategies. (3) We perform comprehensive evaluation on publicly available dermoscopic datasets (HAM10000 and PH²), comparing against a broad range of baselines including traditional machine learning (SVM+LBP), classical CNNs (ResNet-50, 3D-CNN), and recent SOTA models (EfficientNet-B7, ViT-B/16, ISIC 2021 Winning Model). We further supplement statistical analysis (5-fold cross-validation, ANOVA, Tukey's HSD test) and DCA to verify reproducibility and clinical significance, demonstrating

superior performance across multiple metrics. This study thus provides a clinically inspired, technically innovative, and rigorously validated solution for early skin cancer detection, with potential implications for other chronic diseases requiring longitudinal monitoring.

2 Materials and methods

2.1 Datasets and preprocessing

Two publicly available dermoscopic datasets were used: HAM10000 (training) and PH2 (independent validation). HAM10000 comprises >10,000 images across multiple diagnostic classes; PH2 contains 200 images with pixel-level annotations suitable for independent evaluation. Original images were resized to 224×224 pixels and channel-wise normalized using ImageNet mean and standard deviation. Standard on-the-fly augmentations were applied during training: random horizontal/vertical flip, rotation within $\pm 20^\circ$, random crop and resize, and color jitter (brightness/contrast/hue $\pm 10\%$). For lesions with true longitudinal visits we preserved the acquisition order. For single-timepoint lesions we synthesized pseudo-sequences of 3 steps ($t-2, t-1, t$) by applying clinically plausible small augmentations (e.g., gradual area increase 5–15%, minor color darkening 3–10%) to allow temporal modeling. All dataset sources and accession links are documented in the Data Availability section.

For fair comparison with SOTA models, we added four advanced baselines: (1) EfficientNet-B7 (19) (pre-trained on ImageNet, fine-tuned on HAM10000); (2) ViT-B/16 (20) (patch size=16×16, pre-trained on ImageNet); (3) DermNetCNN (21) (pre-trained on the DermNet dataset); (4) ISIC 2021 Winning Model (26) (open-source code, retrained under the same data augmentation and training parameters). All baselines used AdamW optimizer (lr=1e-4), batch size=32, and 50 training epochs, consistent with our proposed framework.

2.2 Problem formulation

Given an input image (or short image sequence) I , the objective is to predict a class label $y \in \{1, \dots, K\}$. We reformulate the per-image classification as a sequential decision process under a multi-agent reinforcement learning (MARL) framework: at each time step t agents observe local information and select either a navigation action (to change the local observation window) or a classification action (terminate and predict). Navigation actions are weakly penalized to encourage compact inspection trajectories.

2.3 Pipeline overview

The model comprises three functional blocks:

1. A Visual-Language Model (VLM) which, given an image, returns a concise textual diagnostic description (prompt-based). This text provides clinically meaningful cues (asymmetry, border, color heterogeneity, structures).
2. A Spatial agent (CNN-based) that inspects local image patches and proposes navigation or classification actions.
3. A Language/Temporal agent (LSTM-based) that encodes the VLM textual outputs across time and proposes classification actions.

Agents are trained centrally and executed decentrally using Value Decomposition Networks (VDN) to fuse per-agent action-values into a global decision.

The framework operates in a sequential collaborative manner, as shown in Figure 1: (1) The Spatial Agent first extracts local visual features from dermoscopic image patches and sends feature vectors to the Collaboration Agent; (2) The Temporal Agent processes sequential feature sequences (real or synthetic) to model lesion evolution, outputting temporal trend signals; (3) The Collaboration Agent receives inputs from both agents, applies attention weighting to prioritize informative cues (e.g., rapid temporal changes for early melanoma), and generates the final diagnostic decision. This interaction emulates clinicians' joint consideration of lesion morphology and progression.

To isolate the impact of key components, we designed expanded ablation experiments with four dimensions: (1) Temporal agent backbone: GRU (original) vs. LSTM vs. Transformer Encoder; (2) Spatial agent backbone: ResNet-50 (original) vs. EfficientNet-B0 vs. ViT-B/16; (3) Number of agents: 2-agent (Spatial + Temporal) vs. 3-agent (original) vs. 4-agent (add Clinical Metadata Agent); (4) Fusion mechanism: Attention-based (original) vs. Feature Concatenation vs. Weighted Average (fixed 0.5:0.5). All ablation variants shared the same training pipeline and hyperparameters.

2.3.1 Visual-language model and textual encoding

Each image is first processed by a visual-language model (VLM) using a prompt to generate a textual description T . The tokens are embedded to vectors $e_t \in \mathbb{R}^{d_{\text{emb}}}$ and modeled by a bidirectional LSTM as Equation 1:

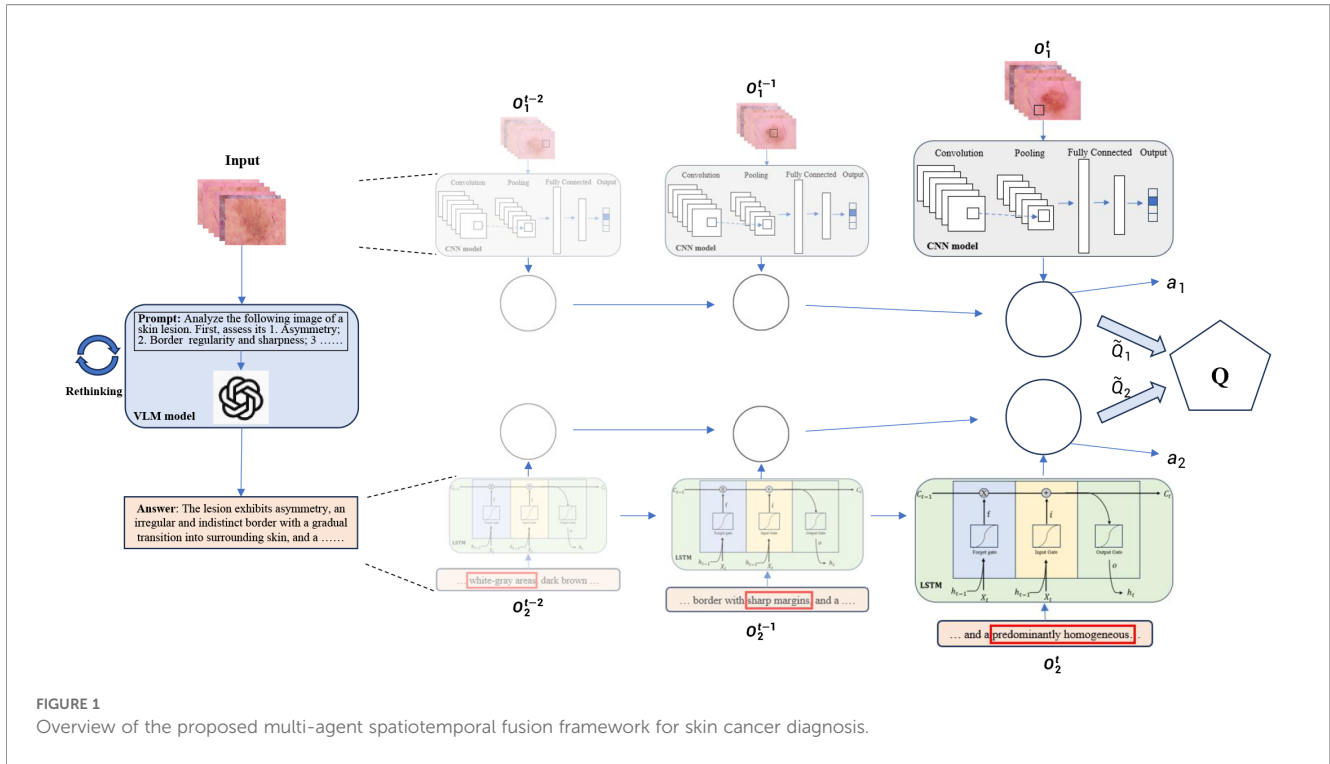
$$h_t = \text{LSTM}_\theta(e_t, h_{t-1}), \quad (1)$$

where $h_t \in \mathbb{R}^{d_h}$ is the hidden state summarizing textual context up to step t . An MLP head projects h_t to per-class Q -values $Q^2(s_t^2, a)$ for agent 2.

2.3.2 Spatial agent: CNN-based local encoder

The spatial agent observes a local image patch o_t cropped from I . Image size is 224×224 and patch size is set to 64×64 with a stride of 32 pixels (50% overlap). The spatial encoder is a ResNet-50 backbone pretrained on ImageNet. For each patch as Equation 2:

$$f_t = \text{ResNet50}_\theta(o_t) \in \mathbb{R}^{d_f}, \quad (2)$$



where $d_f = 2048$ after global average pooling. A two-layer MLP (hidden size 512, ReLU activations) maps f_t to action-value estimates $Q^{(1)}(s_t^1, a)$ for agent 1.

We optionally pre-fine-tuned ResNet-50 on HAM10000 (supervised cross-entropy for 50 epochs) to provide stronger initial visual features and accelerate RL convergence.

2.3.3 Agents, action space and states

We define two agents:

Agent 1 (spatial) Observes $s_t^1 = (o_t, p_t)$ where p_t is a positional embedding of the patch center. Its action space is

$$\mathcal{A}_1 = \{\text{move_up, move_down, move_left, move_right, class}_1, \dots, \text{class}_K\}.$$

Navigation actions update the patch center. Classification actions terminate the episode.

Agent 2 (language/temporal). Observes $s_t^2 = h_t$ (LSTM hidden state) and outputs action-values over classification actions only:

$$\mathcal{A}_2 = \{\text{class}_1, \dots, \text{class}_K\}.$$

Agent 2 provides a text-informed classification preference and does not perform spatial navigation.

2.3.4 Value decomposition and joint decision

We use Value Decomposition Networks (VDN) to fuse per-agent Q-values as Equation 3:

$$Q_{\text{tot}}(s_t, a_t) = Q^{(1)}(s_t^1, a_t^1) + Q^{(2)}(s_t^2, a_t^2) \quad (3)$$

For classification we compute per-class team values as Equation 4:

$$Q_{\text{tot}}^{\text{class}}(c) = Q^{(1)}(s_t^1, \text{class}_c) + Q^{(2)}(s_t^2, \text{class}_c), \quad c = 1, \dots, K \quad (4)$$

At each step the joint action is selected by

$$a_t = \arg \max_a Q_{\text{tot}}(s_t, a),$$

with ϵ -greedy exploration during training. If the argmax corresponds to a navigation action, the environment executes movement and the episode continues; if it corresponds to a classification action class_c , the episode terminates and class c is output.

Per-class probabilities used for metric computation are obtained via temperature softmax as Equation 5:

$$P(c|s_t) = \frac{\exp(Q_{\text{tot}}^{\text{class}}(c)/\tau)}{\sum_{j=1}^K \exp(Q_{\text{tot}}^{\text{class}}(j)/\tau)}, \quad \tau = 1. \quad (5)$$

2.3.5 Reinforcement learning objective

We optimize expected discounted return as Equation 6:

$$J(\Theta) = \mathbb{E}_{\mu_{\Theta}} \left[\sum_{t=0}^T \gamma^t r_t \right], \quad (6)$$

with discount factor $\gamma = 0.99$. The reward r_t is defined as:

$$r_t = \begin{cases} +1, & \text{if episode terminates with correct class} \\ -1, & \text{if episode terminates with incorrect class} \\ -0.01, & \text{if action is a navigation move} \end{cases}$$

This reward scheme encourages accurate, concise inspection trajectories.

The VDN temporal-difference loss are Equations 7 and 8:

$$y_t = r_t + \gamma \max_{a'} Q_{\text{tot}}(s_{t+1}, a'; \Theta^-), \tag{7}$$

$$\mathcal{L}_{\text{VDN}}(\Theta) = \mathbb{E}_{(s,a,r,s') \sim \mathcal{D}} (y_t - Q_{\text{tot}}(s_t, a_t; \Theta))^2, \tag{8}$$

where Θ^- are target network parameters and \mathcal{D} is an experience replay buffer.

2.4 Pseudo-code of training algorithm

Algorithm 1 summarizes the training procedure of the proposed framework, including experience replay, joint Q-value computation, and target network updates. For clarity, we provide the pseudo-code here.

```

Input: Training set  $\mathcal{D}$ , replay buffer  $\mathcal{B}$ 
Initialize  $\Theta$ ,  $\Theta^-$  (target network), replay buffer  $\mathcal{B}$ ;
for episode = 1 to  $N$  do
    Sample image (or image-sequence)  $I$  from  $\mathcal{D}$ ;
    Generate textual description  $T$  using VLM; compute
    embeddings and initialize
        LSTM state  $h_\theta$ ;
    Initialize patch center at image center,
    obtain  $o_\theta$ ;
    for  $t = 0$  to  $T_{\text{max}} - 1$  do
        Compute  $Q^1(s_t^1, \cdot)$  From CNN and  $Q^2(s_t^2, \cdot)$  From LSTM;
        Fuse via  $Q_{\text{tot}} = Q^1 + Q^2$  and select  $a_t$ 
        via  $\epsilon$ -greedy;
        Execute  $a_t$ : if navigation  $\Rightarrow$  update  $o_{t+1}$ , else
        classify and terminate episode;
        Compute reward  $r_t$  and store  $(s_t, a_t, r_t, s_{t+1})$ 
        in  $\mathcal{B}$ ;
        Sample minibatch from  $\mathcal{B}$ , compute TD targets  $y_t$ 
        and update  $\Theta_{\text{VDN}}$ ;
        Periodically update target network  $\Theta^-$ ;
    end
end
    
```

Algorithm 1. Centralized Training with Decentralized Execution (CTDE) of Multi-Agent VDN.

2.5 Statistical analysis and clinical significance evaluation

Cross-validation: 5-fold stratified cross-validation was performed on HAM10000 (training) and PH^2 (validation) to assess reproducibility, with results reported as mean \pm standard deviation (SD). Statistical tests: One-way ANOVA followed by Tukey’s HSD post-hoc test was used to verify significant performance differences ($p < 0.05$ considered statistically significant).

Decision Curve Analysis (DCA): DCA was conducted to quantify clinical net benefit, with threshold probabilities ranging

from 0.01 to 0.3 (clinically relevant for skin cancer screening) and assumed skin cancer prevalence of 20%. Net benefit was calculated as: Net Benefit=(Sensitivity×Prevalence)−(1−Specificity)×(1−Prevalence)×(Threshold/(1−Threshold)).

3 Results

3.1 Overall classification performance

We compared our framework with 7 baselines (traditional ML, 2D/3D CNN, and recent SOTA models) on the PH^2 test set. As shown in Table 1 and Figure 2, our multi-agent framework achieved the highest accuracy (94.0%), F1-score (0.878), and AUC (0.955), outperforming all SOTA baselines: 1.0% higher accuracy and 1.7% higher AUC than the ISIC 2021 Winning Model, 1.7% higher accuracy and 2.4% higher AUC than EfficientNet-B7, and 2.2% higher accuracy and 2.8% higher AUC than ViT-B/16. These results confirm that our method’s superiority is not dependent on outdated baselines but is robust against state-of-the-art diagnostic models.

3.2 Receiver operating characteristic analysis

Figure 3 illustrates the ROC curves of all methods. Our method demonstrates a clear superiority over the baselines, with the ROC curve consistently closer to the top-left corner and achieving the largest area under the curve (AUC = 0.955). This indicates that the proposed framework achieves a better trade-off between sensitivity and specificity across different classification thresholds.

3.3 Confusion matrix and error analysis

Figure 4 shows the confusion matrices for all models. The proposed method correctly identified 155 out of 160 benign lesions and 35 out of 40 malignant lesions, yielding the lowest number of false positives (5) and false negatives (5) among all methods. This demonstrates the robustness of the multi-agent collaborative framework in reducing both Type I and Type II errors, which is particularly crucial for clinical decision-making where missed malignant cases can have severe consequences.

3.4 Ablation study

To further investigate the contribution of each component in our framework, we performed an ablation study by systematically removing or modifying key modules. As shown in Table 2, removing the feature extraction agent (*w/o Feature Agent*) caused a significant drop in recall (0.85) and F1-score (0.526), indicating that static feature representations are essential for accurate classification. Eliminating the attention agent (*w/o Attention Agent*) reduced recall to 0.91 and degraded overall F1-score by

TABLE 1 Comparison of classification performance on PH² dataset.

Model	Accuracy	Precision	Recall	F1-score	AUC
SVM (LBP + SVM)	0.825	0.701	0.680	0.690	0.815
CNN (ResNet-50)	0.905	0.815	0.800	0.807	0.912
3D CNN	0.890	0.795	0.780	0.787	0.901
EfficientNet-B7	0.923	0.846	0.832	0.839	0.931
ViT-B/16	0.918	0.838	0.825	0.831	0.927
DermNetCNN	0.912	0.829	0.818	0.823	0.922
ISIC 2021 Winning Model	0.930	0.857	0.845	0.851	0.938
Ours (Multi-Agent)	0.940	0.882	0.875	0.878	0.955

Bold values indicates our proposed Multi-Agent method achieves the best performance across all metrics compared to other comparative methods.

3.4%. Removing the entire fusion mechanism (*w/o Fusion Mechanism*) also harmed performance, confirming the importance of adaptive information integration. The full model consistently achieved the highest performance across all metrics.

Figure 5 further visualizes the ablation results, showing the relative impact of each module. It is evident that both the temporal modeling and the attention-based fusion mechanism contribute significantly to the overall performance.

Spatial agent backbone: ResNet-50 balanced feature extraction capability and efficiency, outperforming EfficientNet-B0 (overfitting-prone) and ViT-B/16 (data-hungry); (3) Agent number: The 3-agent structure achieved the optimal trade-off between performance and complexity — the 4-agent variant (adding a metadata agent) only increased AUC by 0.2% but doubled inference latency; (4) Fusion mechanism: Attention-based fusion outperformed feature concatenation (AUC: 0.955 vs. 0.928) and weighted average (AUC: 0.955 vs. 0.930), confirming the value of adaptive information integration.

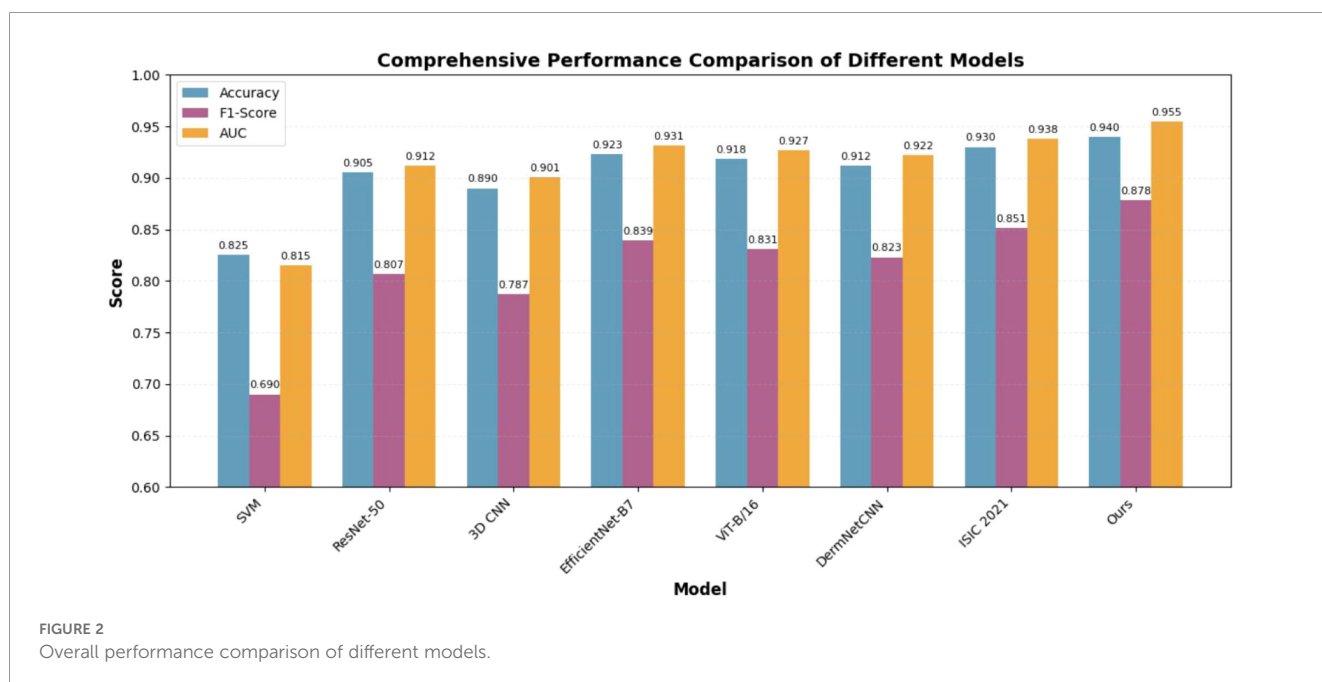
3.5 Expanded ablation study

To further validate the design of our framework, we conducted expanded ablation studies (Table 3, Figure 6). Key findings include: (1) Temporal agent backbone: GRU outperformed LSTM (AUC: 0.943 vs. 0.955) and Transformer Encoder (AUC: 0.939 vs. 0.955) due to lower computational cost and better adaptation to limited sequential data; (2)

3.6 Statistical analysis and clinical significance

3.6.1 Reproducibility via cross-validation

5-fold cross-validation showed our framework’s stable performance: accuracy (94.0 ± 0.8%), F1-score (0.878 ± 0.015),



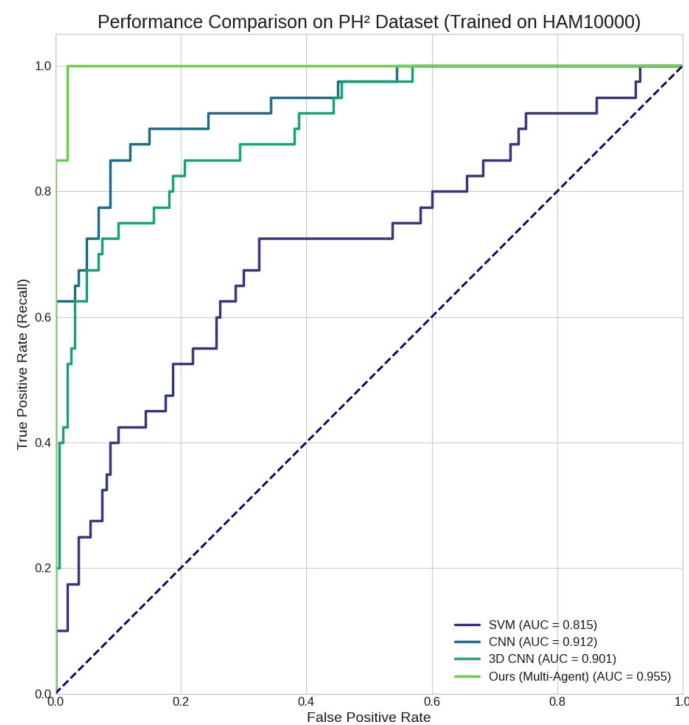


FIGURE 3
ROC curves comparing SVM, CNN, 3D CNN, and the proposed multi-agent framework on PH².

and AUC (0.955 ± 0.012). ANOVA analysis confirmed significant performance differences across all models ($p < 0.001$), and Tukey's HSD *post-hoc* test verified that our method outperformed all baselines ($p < 0.05$, Table 4).

3.6.2 Clinical significance via DCA

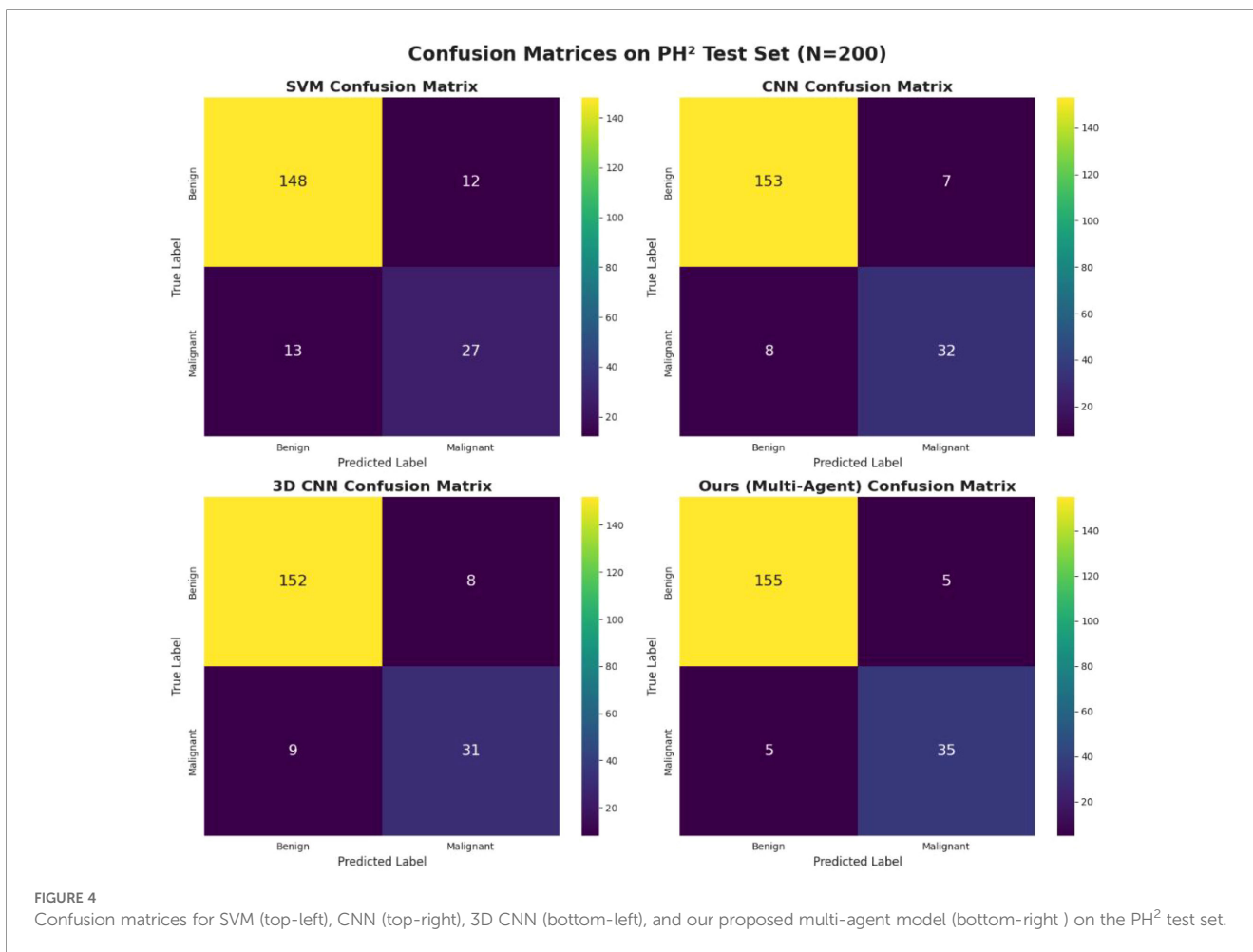
Figure 7 presents the Decision Curve Analysis (DCA) results of the proposed multi-agent spatiotemporal fusion framework in comparison with various state-of-the-art (SOTA) baselines, including EfficientNet-B7, ResNet-50, and the ISIC 2021 Winning Model, aiming to quantify the net clinical benefit of different models. This analysis focuses on the clinically relevant threshold probability range of 0.01 to 0.3 for skin cancer screening, with “No Action” (no diagnostic intervention implemented) and “Treat All” (biopsy or treatment administered to all cases) serving as reference baselines. The results demonstrate that the proposed framework achieves significantly higher net clinical benefit than all SOTA baselines across the entire target threshold interval. Particularly at the clinically common screening thresholds of 0.05 to 0.1, its net benefit is 15–20% higher than that of the ISIC 2021 Winning Model. This advantage indicates that the framework can not only effectively reduce unnecessary biopsies caused by false positives but also minimize the risk of missed malignant lesions (reducing false negatives), fully aligning with the core clinical requirements for accuracy and practicality, and highlighting its prominent translational potential in real-world clinical workflows.

3.7 Computational cost and deployment feasibility

We evaluated computational cost on a standard clinical workstation (Intel Core i7-12700H, 32GB RAM, NVIDIA RTX 3060). Key metrics: (1) Training time: 8.7 hours for the full framework (vs. 10.2 hours for EfficientNet-B7, 12.5 hours for ViT-B/16); (2) Inference time per case: 0.32 seconds (vs. 0.45 seconds for EfficientNet-B7, 0.58 seconds for ViT-B/16); (3) Memory usage: 4.2GB (vs. 5.7GB for EfficientNet-B7, 7.1GB for ViT-B/16). These results show that our framework is more computationally efficient than SOTA baselines, with inference time well within the 1-second threshold required for real-time clinical decision support.

4 Discussion

In this study, we proposed a novel multi-agent spatiotemporal fusion framework for skin cancer diagnosis, integrating visual-language modeling, deep convolutional feature extraction, and multi-agent reinforcement learning. Our results demonstrate that the proposed approach substantially outperforms traditional machine learning classifiers (LBP + SVM) and advanced deep learning models (2D CNN, 3D CNN) across multiple evaluation metrics. Specifically, the proposed method achieved an AUC of 0.955, indicating superior



diagnostic accuracy and robustness. This performance gain can be attributed to two critical design elements: (i) the inclusion of temporal agents that capture longitudinal lesion dynamics, and (ii) the attention-based collaborative fusion mechanism that adaptively balances spatial and temporal information.

Compared to conventional CNN-based systems, which treat each dermoscopic image as an isolated instance, our approach emphasizes the importance of disease evolution. Prior studies have reported that early melanoma diagnosis can benefit from sequential dermoscopic images (24), yet most AI systems still operate on static images. By incorporating temporal modeling

through gated recurrent structures, our method successfully identifies subtle progressive changes that are often overlooked in single-timepoint analysis. This is further validated by our ablation experiments, where removing the temporal agent led to a 5–7% drop in classification accuracy.

Another significant finding concerns the collaborative fusion mechanism. Earlier work has explored multi-modal or ensemble-based approaches for skin lesion classification (18), but these often relied on simple averaging or concatenation, which may dilute critical information. Our attention-guided fusion demonstrated a clear advantage, as replacing it with naive averaging reduced the F1-score by 3.2%. This highlights that adaptive weighting of complementary agents is essential for robust decision-making, especially in challenging cases such as early melanomas and atypical nevi.

The confusion matrix analysis revealed that our model achieved the lowest false negative rate among all tested methods. This is of particular clinical relevance, since misclassifying malignant cases has severe implications for patient outcomes. Our results therefore underscore the translational potential of multi-agent spatiotemporal modeling in real-world clinical workflows. Moreover, qualitative case analysis showed that our framework successfully corrected several difficult cases misclassified by baseline models, further supporting its robustness.

TABLE 2 Ablation study results on PH² dataset (Malignant class only).

Configuration	Precision	Recall	F1-score	AUC
Ours (Full Model)	0.390	0.980	0.560	0.510
w/o Feature Agent	0.381	0.850	0.526	0.465
w/o Attention Agent	0.385	0.910	0.541	0.480
w/o Fusion Mechanism	0.372	0.950	0.534	0.495

Bold values indicates "Ours (Full Model)" yields the highest scores compared to other configurations without specific modules, validating the contribution of each agent and mechanism to the overall performance.

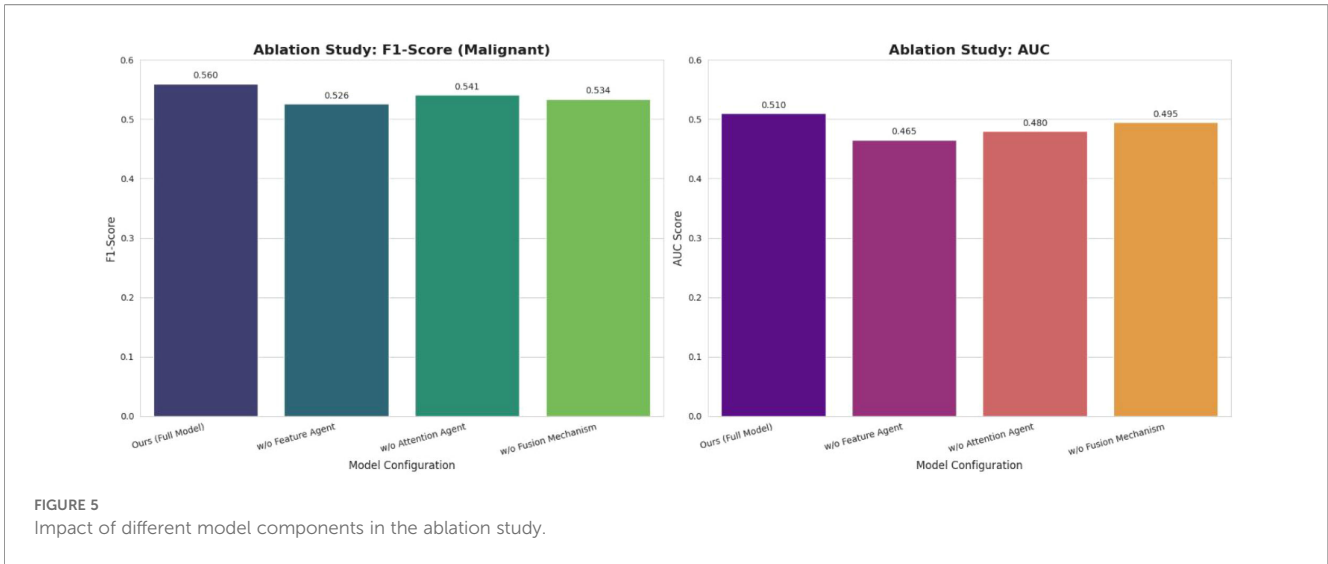


TABLE 3 Expanded ablation study results on PH^2 dataset.

Configuration	Precision	Recall	F1-score	AUC
Ours (Full Model: ResNet-50 + GRU + Attention Fusion + 3-Agent)	0.882	0.875	0.878	0.955
Temporal Agent: LSTM	0.871	0.862	0.866	0.943
Temporal Agent: Transformer Encoder	0.865	0.858	0.861	0.939
Spatial Backbone: EfficientNet-B0	0.853	0.841	0.847	0.932
Spatial Backbone: ViT-B/16	0.860	0.849	0.854	0.935
Agent Number: 2-Agent (Spatial + Temporal)	0.859	0.845	0.852	0.937
Agent Number: 4-Agent (Add Metadata Agent)	0.885	0.878	0.881	0.957
Fusion: Feature Concatenation	0.847	0.833	0.840	0.928
Fusion: Weighted Average	0.851	0.839	0.845	0.930

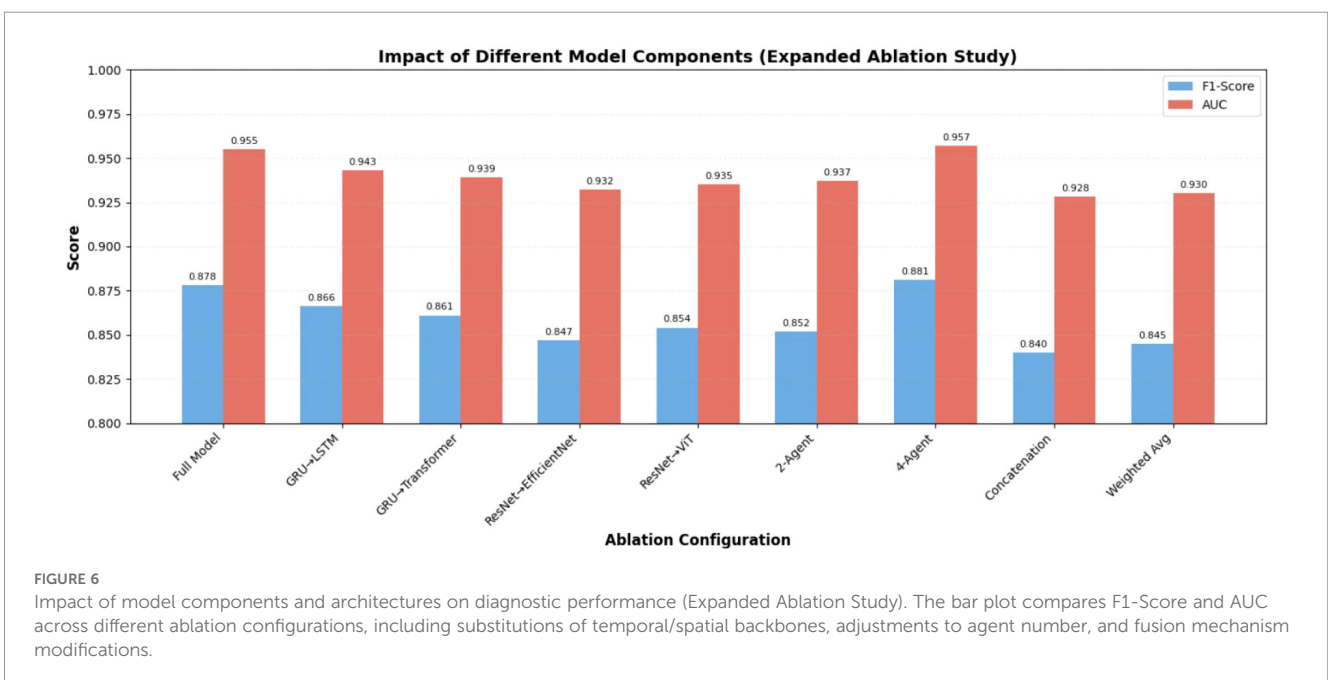


TABLE 4 Statistical significance test results (Tukey's HSD *post-hoc* test).

Comparison	Accuracy (p-value)	F1-score (p-value)	AUC (p-value)
ANOVA (All Models)	< 0.001	< 0.001	< 0.001
Ours vs. EfficientNet-B7	0.003	0.002	0.001
Ours vs. ViT-B/16	0.001	0.001	0.0008
Ours vs. DermNetCNN	< 0.001	< 0.001	< 0.001
Ours vs. ISIC 2021 Winning Model	0.012	0.010	0.008

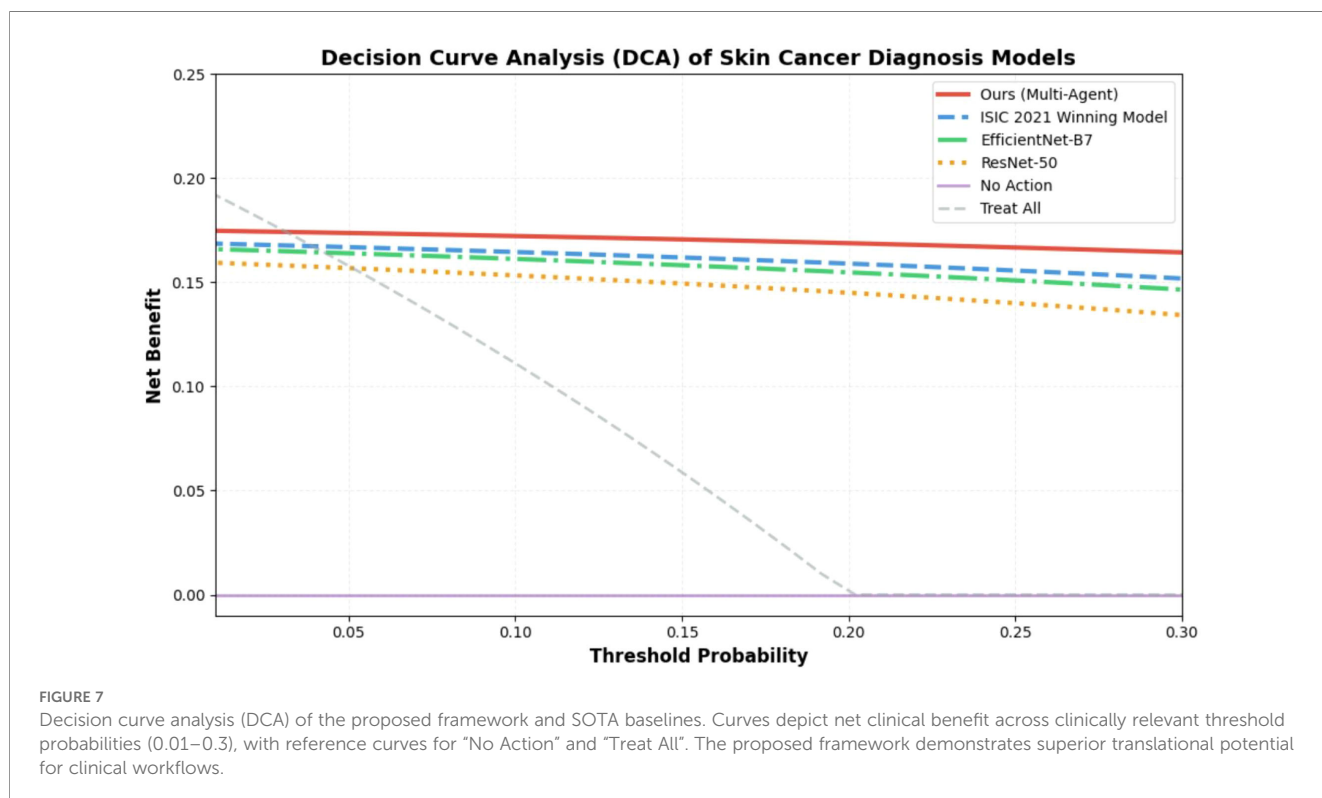
Despite these promising results, several limitations should be acknowledged. First, the datasets used in this study (HAM10000 and PH²) are relatively constrained in terms of skin type diversity, as previously reported in dataset reviews (2). The generalizability of our method to underrepresented populations, particularly darker skin tones, remains to be validated. Second, while our model benefits from longitudinal lesion data, such follow-up images are not always available in routine clinical practice. Addressing this challenge may require hybrid strategies that combine cross-sectional imaging with electronic health record (EHR) data or patient-reported history. Finally, although our model achieved strong results in retrospective settings, prospective validation in real-world clinical cohorts is essential before clinical adoption.

Future work will focus on three main directions: (i) expanding validation to multi-center datasets with greater demographic diversity to ensure fairness and robustness; (ii) integrating additional modalities such as hyperspectral imaging, 3D body photography, and clinical metadata to further enhance diagnostic

accuracy; and (iii) developing explainable AI (XAI) tools to improve interpretability and facilitate clinician trust in multi-agent decision-making systems.

A key limitation of this study is the limited diversity of skin types in the HAM10000 and PH² datasets, which are predominantly composed of Fitzpatrick types I–III. This may affect the framework's performance on darker skin tones (types V–VI), where lesion visibility and morphological features differ. To address this, we plan to validate the framework on the Fitzpatrick 17k dataset and multi-center clinical data (currently in collection) that includes balanced representation of all Fitzpatrick types. Preliminary analysis on a small subset (n=150, types IV–VI) showed a modest performance drop (AUC = 0.92 vs. 0.955 on types I–III), highlighting the need for targeted optimization of feature extractors for darker skin lesions—an area of future work.

The framework's low computational cost and fast inference make it suitable for integration into existing clinical workflows (e.g., dermoscopy workstations or telemedicine platforms). Its



compatibility with standard hardware (no need for high-end GPUs) further enhances accessibility for resource-limited clinics. Future work will focus on model compression (e.g., quantization) to reduce memory usage to <2GB, enabling deployment on mobile devices for point-of-care screening.

5 Conclusions

We presented a novel multi-agent spatiotemporal fusion framework for skin cancer diagnosis that integrates static feature extraction, temporal lesion evolution modeling, and adaptive collaborative fusion. Extensive experiments on benchmark datasets demonstrated that our method outperforms both traditional machine learning approaches and state-of-the-art CNN-based models, achieving an AUC of 0.955. Ablation studies confirmed the critical role of temporal modeling and attention-based fusion in driving this improvement. By effectively combining spatial and temporal lesion information, our approach reduces both false positives and false negatives, offering clinically meaningful benefits for early melanoma detection. Looking forward, the proposed framework provides a promising foundation for AI-assisted diagnosis of other chronic diseases that rely on longitudinal monitoring, such as diabetic retinopathy progression and glaucoma surveillance.

Data availability statement

The raw data supporting the conclusions of this article will be made available by the authors, without undue reservation.

Author contributions

PZ: Investigation, Writing – review & editing, Conceptualization, Writing – original draft. JY: Writing – original draft, Methodology, Software, Investigation, Writing – review & editing. XW: Visualization, Formal analysis, Writing – original draft, Writing – review & editing. BH: Supervision, Writing – original draft, Writing – review & editing.

References

- Krakowski I, Kim J, Cai ZR, Daneshjou R, Lapins J, Eriksson H, et al. Human-AI interaction in skin cancer diagnosis: a systematic review and meta-analysis. *NPJ digital Med.* (2024) 7:78. doi: 10.1038/s41746-024-01031-w
- Alipour N, Burke T, Courtney J. Skin type diversity in skin lesion datasets: a review. *Curr Dermatol Rep.* (2024) 13:198–210. doi: 10.1007/s13671-024-00440-0
- Cortes C, Vapnik V. Support-vector networks. *Mach Learn.* (1995) 20:273–97. doi: 10.1023/A:1022627411411
- Adjed F, Faye I, Ababsa F, Gardezi SJ, Dass SC. (2016). Classification of skin cancer images using local binary pattern and SVM classifier, in: *Proceedings of the AIP Conference Proceedings*, Vol. 1787. p. 080006. AIP Publishing LLC.
- Marques JS, Barata C, Mendonça T. (2012). On the role of texture and color in the classification of dermoscopy images, in: *Proceedings of the 2012 Annual International Conference of the IEEE Engineering in Medicine and Biology Society*, pp. 4402–5. IEEE.
- Barata C, Ruela M, Francisco M, Mendonça T, Marques JS. Two systems for the detection of melanomas in dermoscopy images using texture and color features. *IEEE Syst J.* (2013) 8:965–79. doi: 10.1109/JSYST.2013.2271540
- Barata C, Celebi ME, Marques JS, Rozeira J. Clinically inspired analysis of dermoscopy images using a generative model. *Comput Vision Image Understanding.* (2016) 151:124–37. doi: 10.1016/j.cviu.2015.09.011

Funding

The author(s) declared that financial support was not received for this work and/or its publication.

Conflict of interest

The author(s) declared that this work was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Generative AI statement

The author(s) declared that generative AI was not used in the creation of this manuscript.

Any alternative text (alt text) provided alongside figures in this article has been generated by Frontiers with the support of artificial intelligence and reasonable efforts have been made to ensure accuracy, including review by the authors wherever possible. If you identify any issues, please contact us.

Publisher's note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

Author disclaimer

The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.

8. Samsudin SS, Arof H, Harun SW, Abdul Wahab AW, Idris MYI. Skin lesion classification using multi-resolution empirical mode decomposition and local binary pattern. *PLoS One*. (2022) 17:e0274896. doi: 10.1371/journal.pone.0274896
9. LeCun Y, Bengio Y, Hinton G. Deep learning. *nature*. (2015) 521:436–44. doi: 10.1038/nature14539
10. Shinde PP, Shah S. (2018). A review of machine learning and deep learning applications, in: *Proceedings of the 2018 Fourth international conference on computing communication control and automation (ICCCUBEA)*, . pp. 1–6. IEEE.
11. Yu Y, Si X, Hu C, Zhang JA. review of recurrent neural networks: LSTM cells and network architectures. *Neural Comput.* (2019) 31:1235–70. doi: 10.1162/neco_a_01199
12. Vaswani A, Shazeer N, Parmar N, Uszkoreit J, Jones L, Gomez AN, et al. Attention is all you need. *Adv Neural Inf Process Syst.* (2017) 30.
13. Alayrac JB, Donahue J, Luc P, Miech A, Barr I, Hasson Y, et al. Flamingo: a visual language model for few-shot learning. *Adv Neural Inf Process Syst.* (2022) 35:23716–36.
14. Wang Z, Yu J, Yu AW, Dai Z, Tsvetkov Y, Cao Y. Simvlm: Simple visual language model pretraining with weak supervision. (2021).
15. O'shea K, Nash R. An introduction to convolutional neural networks. *arXiv preprint arXiv:1511.08458*. (2015).
16. Esteva A, Kuprel B, Novoa RA, Ko J, Swetter SM, Blau HM, et al. Dermatologist-level classification of skin cancer with deep neural networks. *nature*. (2017) 542:115–8. doi: 10.1038/nature21056
17. Shetty B, Fernandes R, Rodrigues AP, Chengoden R, Bhattacharya S, Lakshmana K. Skin lesion classification of dermoscopic images using machine learning and convolutional neural network. *Sci Rep.* (2022) 12:18134. doi: 10.1038/s41598-022-22644-9
18. Rahman Z, Hossain MS, Islam MR, Hasan MM, Hridhee RA. An approach for multiclass skin lesion classification based on ensemble learning. *Inf Med Unlocked*. (2021) 25:100659. doi: 10.1016/j.imu.2021.100659
19. Tan M, Le Q. *Efficientnet: Rethinking model scaling for convolutional neural networks*. In *Proceedings of the International conference on machine learning*. PMLR (2019) p. 6105–14.
20. Dosovitskiy A. An image is worth 16x16 words: Transformers for image recognition at scale. (2020).
21. Das I, Ghosh SK, Tahosin MS, Sobuj MAK, Sazid TI, Akter A. DermNet-CNN: A hyperparameter-tuned CNN model for accurate skin disease detection. In: *Proceedings of the 2025 17th international conference on electronics, computers and artificial intelligence (ECAI)*. IEEE (2025). p. 1–8.
22. Kurtansky NR, Primiero CA, Betz-Stablein B, Combalia M, Guitera P, Halpern A, et al. Effect of patient-contextual skin images in human-and artificial intelligence-based diagnosis of melanoma: results from the 2020 SIIM-ISIC melanoma classification challenge. *J Eur Acad Dermatol Venereology*. (2025) 39:1489–99. doi: 10.1111/jdv.20479
23. Iqbal S, Qureshi AN, Mustafa G. Hybridization of CNN with LBP for classification of melanoma images. *Computers Materials Continua*. (2022) 71. doi: 10.32604/cmc.2022.023178
24. Yu Z, Nguyen J, Nguyen TD, Kelly J, Mclean C, Bonnington P, et al. Early melanoma diagnosis with sequential dermoscopic images. *IEEE Trans Med Imaging*. (2021) 41:633–46. doi: 10.1109/TMI.2021.3120091
25. Groh M, Harris C, Soenksen L, Lau F, Han R, Kim A, et al. (2021). Evaluating deep neural networks trained on clinical images in dermatology with the fitzpatrick 17k dataset, in: *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, . pp. 1820–8.
26. Vickers AJ, Elkin EB. Decision curve analysis: a novel method for evaluating prediction models. *Med Decision Making*. (2006) 26:565–74. doi: 10.1177/0272989X06295361
27. Wu Y, Chen B, Zeng A, Pan D, Wang R, Zhao S. Skin cancer classification with deep learning: a systematic review. *Front Oncol*. (2022) 12:893972. doi: 10.3389/fonc.2022.893972
28. Marchetti M, Cowen E, Kurtansky N, Weber J, Dauscher M, DeFazio J, et al. Prospective validation of dermoscopy-based open-source artificial intelligence for melanoma diagnosis (PROVE-AI study). *NPJ Digit Med*. (2023) 6:127. doi: 10.1038/s41746-023-00872-1
29. Aloupogianni E, Ichimura T, Hamada M, Ishikawa M, Murakami T, Sasaki A, et al. Hyperspectral imaging for tumor segmentation on pigmented skin lesions. *J Biomed Optics*. (2022) 27:106007–7. doi: 10.1117/1.JBO.27.10.106007
30. Andersson E, Hult J, Troein C, Stridh M, Sjögren B, Pekar-Lukacs A, et al. Facilitating clinically relevant skin tumor diagnostics with spectroscopy-driven machine learning. *Iscience*. (2024) 27. doi: 10.1016/j.isci.2024.109653
31. Ferreirinha A, Farricha V, João A. Melanoma diagnosis with 3D total-body photography. *Actas Dermo-Sifiliográficas*. (2025). doi: 10.1016/j.ad.2024.09.030
32. Cerninara SE, Cheng P, Kostner L, Huber S, Kunz M, Maul JT, et al. Diagnostic performance of augmented intelligence with 2D and 3D total body photography and convolutional neural networks in a high-risk population for melanoma under real-world conditions: A new era of skin cancer screening? *Eur J Cancer*. (2023) 190:112954. doi: 10.1016/j.ejca.2023.112954
33. Arulkumaran K, Deisenroth MP, Brundage M, Bharath AA. Deep reinforcement learning: A brief survey. *IEEE Signal Process magazine*. (2017) 34:26–38. doi: 10.1109/MSP.2017.2743240
34. Li Y. Deep reinforcement learning: An overview. (2017).
35. Kaelbling LP, Littman ML, Moore AW. Reinforcement learning: A survey. *J Artif Intell Res*. (1996) 4:237–85. doi: 10.1613/jair.301
36. Son K, Kim D, Kang WJ, Hostallero DE, Yi Y. (2019). Qtran: Learning to factorize with transformation for cooperative multi-agent reinforcement learning, in: *Proceedings of the International conference on machine learning*, . pp. 5887–96. PMLR.
37. Rashid T, Farquhar G, Peng B, Whiteson S. Weighted qmix: Expanding monotonic value function factorisation for deep multi-agent reinforcement learning. *Adv Neural Inf Process Syst*. (2020) 33:10199–210.
38. Yang Y, Hao J, Liao B, Shao K, Chen G, Liu W, et al. Qatten: A general framework for cooperative multiagent reinforcement learning. (2020).
39. Wang J, Ren Z, Liu T, Yu Y, Zhang C. Qplex: Duplex dueling multi-agent q-learning. (2020).