Akash saini, 2026, 14:3
ISSN (Online): 2348-4098
ISSN (Print): 2395-4752

International Journal of Science, Engineering and Technology
An Open Access Journal

A Review of Deep Learning Architectures for Multi-Class Skin Lesion Classification: From CNN-Recurrent Hybrids to Class-Balanced Peephole LSTM Frameworks

Akash saini[1], Jitender Kumar[2]
[1]Student, Department of Computer Science and EngineeringGITAM , kablana
[2]Assistant Professor, Department of Computer Science and Engineering GITAM, Kablana,

Abstract- Automated dermoscopic image analysis has emerged as a critical tool for addressing the global burden of skin cancer, yet two persistent obstacles continue to limit clinical translation: high inter-class morphological similarity among diagnostic categories and severe class imbalance within benchmark datasets such as HAM10000. This paper reviews the methodological evolution of deep learning approaches for multi-class skin lesion classification, tracing the progression from handcrafted feature-based classical machine learning pipelines through deep convolutional architectures, hybrid CNN-recurrent frameworks, and Vision Transformer-based models. Particular attention is given to the gate desynchronisation limitation inherent in standard Long Short-Term Memory units, which excludes the internal cell state from gate computations and can cause premature loss of diagnostically relevant sequential information. The review further examines loss-level and data-level strategies for class-imbalance mitigation, including Weighted Cross-Entropy, Focal Loss, SMOTE, and generative adversarial augmentation, and surveys the explainability frameworks required for regulatory and clinical acceptance. Based on the gaps identified, the paper discusses an emerging direction—class-balanced hybrid CNN-Peephole LSTM frameworks—and outlines future research priorities, including attention-based feature filtering, multi-modal metadata fusion, and Vision Transformer knowledge distillation for resource-constrained clinical deployment.

Keywords: Skin cancer detection, dermoscopy, convolutional neural network, Peephole LSTM, HAM10000, class imbalance, explainable AI, deep learning.

I. INTRODUCTION

Skin cancer remains among the most prevalent and diagnostically challenging malignancies in dermatological practice, with non-melanoma cancers accounting for several million new diagnoses annually and melanoma alone responsible for the majority of skin-cancer-related mortality because of its capacity for early metastatic spread through lymphatic and haematogenous routes. Melanocytes that acquire oncogenic alterations in pathways such as BRAF, NRAS, and KIT can lose control over proliferation and differentiation, giving rise to aggressively invasive tumours. The clinical heterogeneity of skin lesions—spanning malignant melanoma, basal and squamous cell carcinomas, pre-malignant actinic keratoses, and a range of benign pigmented and vascular lesions—is reflected in the seven diagnostic categories represented in widely used benchmark datasets such as HAM10000 [1].

Early detection carries disproportionate clinical value: localised melanoma is associated with five-year survival approaching 100%, while distant metastatic disease falls below 25%, and the surgical and pharmacological burden of late-stage treatment is correspondingly far greater than that of early excision. Traditional visual diagnostic heuristics such as the ABCD rule, the seven-point checklist, and the Menzies method introduced structure to dermoscopic assessment, yet inter-observer agreement even among trained dermatologists rarely exceeds 70–75%. Dermoscopy improves diagnostic sensitivity by revealing subsurface pigment networks, vascular patterns, and regression structures invisible to the naked eye, but reliable interpretation requires extensive experience—a constraint compounded by a global shortage of

dermatologists that, in some regions, falls below one specialist per two million people. These constraints collectively motivate the development of automated, computationally efficient dermoscopic classification systems capable of supporting triage at scale.

Automated dermoscopic classification has progressed through several broadly identifiable methodological generations, summarised in Table I. Early systems relied on handcrafted statistical descriptors—colour histograms, Gray-Level Co-occurrence Matrix texture features, and Fourier shape descriptors—paired with classical classifiers such as Support Vector Machines. A pivotal shift occurred when Esteva et al. demonstrated that a deep convolutional neural network could match dermatologist-level accuracy in binary skin-cancer classification, catalysing a transition toward end-to-end deep feature learning [2]. Subsequent work established that pure CNN architectures, constrained by the local receptive fields of their convolutional kernels, struggle to capture the long-range structural dependencies spanning an entire dermoscopic lesion, motivating a third generation of hybrid CNN-recurrent architectures pairing convolutional spatial feature extraction with sequential modelling via LSTM, GRU, or BiLSTM units. A fourth and ongoing generation introduces Vision Transformer architectures, which model global dependencies through self-attention at the cost of quadratic computational complexity.

This paper reviews the literature underlying this methodological trajectory, with particular emphasis on the architectural and optimisation-level challenges that remain unresolved—most notably gate desynchronisation in standard recurrent gating units and the persistent difficulty of training reliable multi-class classifiers on the severely imbalanced distributions characteristic of dermoscopic benchmarks such as HAM10000. Section II reviews the literature across seven thematic areas. Section III presents a comparative synthesis of the reviewed architectures. Section IV consolidates the research gaps identified across this literature. Section V discusses an emerging direction—class-balanced hybrid CNN-Peephole LSTM frameworks—as a candidate response to these gaps, Section VI outlines future research priorities, and Section VII concludes the paper.

TABLE I
EVOLUTION OF AUTOMATED DERMOSCOPIC IMAGE ANALYSIS TECHNIQUES

Era	Approach	Representative Techniques	Key Limitation
Pre-2010	Handcrafted features + classical ML	ABCD rule, GLCM, colour histograms, SVM	Manual feature design; poor cross-domain generalisation
2010–2016	Classical ML + segmentation	Active contours, Random Forest, k-NN	Performance highly sensitive to segmentation quality
2016–2020	Deep CNN + transfer learning	ResNet, VGGNet, InceptionNet, Xception, EfficientNet	Limited long-range sequential / global context modelling
2020–2023	Hybrid CNN-recurrent	CNN-LSTM, CNN-BiLSTM, CNN-GRU	Gate desynchronisation in standard recurrent units
2023–present	Transformer & hybrid models	ViT, Swin, DeiT, CNN-Peephole LSTM	Quadratic attention complexity; edge-deployment limits

II. LITERATURE REVIEW

A. Handcrafted Feature-Based and Classical Machine Learning Approaches

The earliest computational approaches to skin lesion classification relied on explicit feature engineering, in which domain experts designed mathematical descriptors intended to mirror the visual criteria used in dermoscopic diagnosis. Lesion asymmetry was quantified through geometric moment analysis of segmented boundaries, border irregularity through fractal-dimension measures, colour characteristics through statistical moments in perceptually uniform colour spaces such as HSV and CIELAB, and texture through Gray-Level Co-occurrence Matrix descriptors capturing second-order pixel co-occurrence statistics. These high-dimensional handcrafted vectors were typically classified using Support Vector Machines with radial-basis-function

kernels, valued for their generalisation properties under high-dimensional, low-sample-count conditions.

Despite reasonable performance in controlled, single-institution settings, classical pipelines suffered from two compounding limitations. First, their dependence on an upstream segmentation step meant that any boundary-delineation error—introduced by hair artefacts, low-contrast margins, or background skin texture—propagated irreversibly into every downstream feature. Second, the manually selected features reflected human perceptual salience rather than statistical discriminability, and feature sets engineered for binary melanoma-versus-benign discrimination did not generalise to the seven-category multi-class setting represented by datasets such as HAM10000. Khan et al., in a comprehensive review spanning several hundred published studies, traced the field's progression away from these handcrafted pipelines toward learned representations and identified this transition as a precondition for the accuracy gains achieved by subsequent CNN-based methods [3].

B. From Deep CNNs to Hybrid CNN-Recurrent Architectures: **The Gate Desynchronisation Problem**

The publication of the HAM10000 dataset—10,015 dermoscopic images spanning seven diagnostic categories collected across clinical institutions in Austria and Australia—provided the first benchmark of sufficient scale and diagnostic diversity for systematic multi-class deep learning evaluation [1]. Hosny et al. demonstrated that ImageNet-pretrained CNN backbones, fine-tuned on dermoscopic data through transfer learning, substantially outperformed classical handcrafted pipelines, establishing transfer learning as the dominant training paradigm for the domain [4]. Building on the milestone established by Esteva et al. [2], Chaturvedi et al. showed that deeper residual architectures—specifically ResNet-101—translated increased network depth into improved dermoscopic feature extraction without the vanishing-gradient degradation that limits naively deep networks [5], while Anand et al. reported 91.47% accuracy on HAM10000 using the Xception architecture, whose depthwise-separable convolutions improve parameter efficiency relative to standard convolutions; their analysis also identified a domain-shift vulnerability arising from the mismatch between ImageNet's natural-scene pretraining distribution and the controlled, polarised-illumination conditions of dermoscopic acquisition [6]. Yap et al. further showed that fusing features extracted at multiple spatial resolutions improved sensitivity to both coarse lesion morphology and fine-grained intra-lesion structure [7].

Despite these advances, pure CNN architectures share a structural limitation: the local receptive fields of convolutional kernels—typically 3×3 or 5×5 pixels—bound the spatial context accessible at each processing stage. While stacking additional convolutional layers gradually enlarges the effective receptive field, this growth is computationally expensive and indirect. For dermoscopic images, where the diagnostic significance of a structural feature at one location may depend on its relationship to features located hundreds of pixels away, this constraint means pure CNNs cannot directly model the global spatial coherence that experienced dermatologists rely upon.

This limitation motivated a shift toward hybrid CNN-recurrent architectures, in which convolutional layers first produce spatial feature maps that are reshaped into sequences and processed by recurrent units capable of modelling long-range transitions across the lesion. Reis et al. showed that combining recurrent sequential processing with GAN-based augmentation yielded measurable accuracy improvements over handcrafted-feature pipelines on HAM10000 [8], while Li et al. proposed an attention-augmented CNN-BiLSTM architecture in which bidirectional processing captured both forward and backward structural transitions, at the cost of doubled recurrent computation [9]. A more fundamental limitation of standard LSTM gating was characterised by Gao et al. through systematic analysis of medical image volume slice prediction: standard forget, input, and output gates are computed solely as functions of the previous hidden state and current input, with the internal cell state entirely excluded from gate computations [10]. This

gate desynchronisation means memory-retention decisions are made without direct knowledge of the accumulated long-term context, causing premature loss of subtle structural transitions—such as boundary anomalies in a flattened dermoscopic sequence—that require precise temporal synchronisation between gating and memory state [10].

C. Vision Transformer-Based Architectures and Computational Trade-offs

A further methodological shift came with Vision Transformer (ViT) architectures, which replace the local inductive bias of convolution with multi-head self-attention operating over sequences of non-overlapping image patches, enabling direct modelling of global relationships between distant image regions. Kaya et al. proposed a Transformer-based hybrid decision-support system incorporating spatial and channel attention mechanisms, reporting 95.10% accuracy on HAM10000 and establishing ViT-hybrids among the strongest-performing architectures evaluated on this benchmark [11]. Mahalle and Shinde reported even higher accuracy—97%—using a CNN-LSTM hybrid, although their analysis noted this performance was achieved partly through traditional upsampling of minority classes, raising concerns about overfitting to a limited set of minority-class image variations [12].

The principal limitation of ViT-based architectures is computational: multi-head self-attention has quadratic time complexity, $O(N^2)$, with respect to sequence length. For applications intended for mobile devices or local clinical terminals lacking high-performance GPU infrastructure, this scaling renders real-time inference impractical. This trade-off has motivated knowledge-distillation approaches such as Hybrid Data-Efficient Knowledge Distillation (HDKD), which transfer the global inductive biases learned by a large ViT teacher into a lightweight CNN-recurrent student, aiming to retain global context modelling while preserving the linear computational complexity, $O(N)$, required for edge deployment [13]. The tension between the global modelling capacity of attention-based architectures and the computational constraints of clinical deployment environments represents one of the central threads running through the literature reviewed in this paper.

D. Loss-Level Strategies for Class Imbalance: Weighted Cross-Entropy and Focal Loss

Independent of architectural choice, the extreme class imbalance characteristic of dermoscopic benchmark datasets represents perhaps the most consistently identified barrier to clinical translation. In HAM10000, the majority class—Melanocytic Nevi—accounts for nearly two-thirds of all samples, while clinically critical minority classes such as Dermatofibroma and Vascular Lesions together represent less than three percent of the dataset. Khan et al.'s review identified this imbalance as the field's most critical unresolved challenge, noting that the dominant use of overall accuracy as an evaluation metric obscures systematic minority-class failures that would be clinically unacceptable, since a model achieving high accuracy primarily through majority-class bias would be diagnostically unreliable for precisely the rare malignant categories it most needs to detect [3].

Weighted Cross-Entropy (WCE) addresses this imbalance at the loss level by assigning each class a weight inversely proportional to its sample frequency, amplifying the gradient contribution of minority-class misclassifications. While effective as a first-order correction, WCE does not suppress the gradient contribution of easily classified majority-class samples, meaning high-confidence benign predictions continue to consume gradient budget throughout training even as minority-class penalties are amplified. Focal Loss, introduced by Lin et al. for dense object detection, addresses this asymmetry through a dynamic modulating factor that progressively suppresses the loss contribution of confidently and correctly classified samples, automatically redirecting the optimiser's attention toward harder, more ambiguous examples [14]. Balci et al. conducted a comparative investigation on highly imbalanced medical datasets and demonstrated that Focal Loss substantially outperforms WCE in protecting minority-class gradient contributions, attributing the improvement specifically to this self-adjusting suppression of easy

majority-class gradients—a property that WCE's static inverse-frequency weighting cannot replicate [15]. This comparison suggests that while WCE represents a meaningful and computationally inexpensive imbalance-mitigation step, architectures relying on it alone may face an inherent ceiling on minority-class performance that loss functions with dynamic gradient modulation are better positioned to overcome.

E. Data-Level Strategies: SMOTE and Generative Adversarial Augmentation

As an alternative to loss-level reweighting, data-level strategies address class imbalance by directly altering the composition of the training distribution. Chawla et al.'s Synthetic Minority Over-sampling Technique (SMOTE) generates synthetic minority-class samples through linear interpolation between nearest-neighbour feature representations [16]. While effective for tabular and low-dimensional classification tasks, SMOTE's application in pixel space produces biologically implausible synthetic dermoscopic images, since the complex, non-linear morphological structures of skin lesions cannot be meaningfully represented through linear pixel-wise interpolation.

Generative Adversarial Networks (GANs) offer a more sophisticated data-level alternative by learning to synthesise morphologically diverse minority-class images directly. Cardoso et al. demonstrated that GAN-synthesised minority-class samples could provide additional morphological diversity that improved classification performance on HAM10000 [17]. Building on this approach, pipelines combining CycleGAN-based augmentation with EfficientNet-B3 backbones and Focal Loss have achieved a reported peak accuracy of 99.33% and macro-F1 of 0.9685 on HAM10000—figures that currently represent the published performance ceiling for this benchmark [18]. However, Fréchet Inception Distance (FID) analysis of these GAN-augmented pipelines reveals residual domain gaps—ranging from approximately 142 to 170 depending on skin-tone conditioning—between the distributions of synthetic and real clinical images [18]. This measurable discordance indicates that GAN-based augmentation, despite its strong in-benchmark performance, may carry generalisation risk when models are deployed against clinical populations or dermoscope configurations that differ from those represented in the training distribution.

Taken together, the loss-level and data-level imbalance-mitigation literature suggests a layered picture: WCE provides an inexpensive but ceiling-limited correction; Focal Loss offers a more dynamically adaptive loss-level alternative; and GAN-based augmentation can push benchmark performance substantially higher but introduces a synthetic-to-real domain gap whose clinical implications remain incompletely characterised.

F. Explainable AI for Clinical Interpretability

Beyond raw classification performance, the integration of Explainable AI (XAI) into dermoscopic pipelines has become a clinical and regulatory necessity. Emerging regulatory frameworks for AI-based medical devices increasingly require post-hoc explanation capabilities as a precondition for clinical approval, a requirement examined in detail by Grignaffini et al. in their systematic review of XAI clinical-translation challenges [19]. Hussein et al. developed attention-based CNN-BiLSTM networks incorporating gradient-weighted attribution maps, overlaying heatmaps on dermoscopic images to allow clinicians to confirm that model attention was directed toward diagnostically relevant structures—such as atypical pigment networks or regression areas—rather than irrelevant background regions [20].

Abir et al. demonstrated that lightweight CNN architectures paired with post-hoc explanation methods including LIME, SHAP, and Grad-CAM produced interpretable outputs accepted by practising clinicians in tele-dermatology workflows, indicating that meaningful spatial explanations do not require architecturally complex models [21]. However, Grignaffini et al. identified a persistent misalignment between model-generated saliency maps and the regions of interest that clinicians expect based on established diagnostic criteria as the primary barrier to clinical adoption, concluding that for XAI outputs to be clinically meaningful, spatial attributions from methods such as Grad-CAM

must be explicitly validated against established dermoscopic criteria—the ABCD rule and the seven-point checklist—rather than assumed to align with clinical reasoning by default [19].

These findings collectively suggest that XAI integration in dermoscopic classification faces two distinct challenges: a technical challenge of generating spatial attributions from a given architecture, which lightweight CNNs handle adequately, and a validation challenge of demonstrating that those attributions correspond to clinically meaningful diagnostic criteria, which remains largely unresolved. For hybrid CNN-recurrent architectures specifically, this second challenge is compounded by the need to extend attribution methods beyond the convolutional backbone to the sequential dimension introduced by the recurrent unit, an extension that has received comparatively limited attention in the reviewed literature.

G. Lightweight and Edge-Oriented Sequential Architectures

A final thematic strand in the literature addresses the practical deployment constraints of automated dermoscopic classification in resource-limited clinical settings, particularly in regions where dermatologist shortages are most acute and high-performance GPU infrastructure is least available. Sahin et al. proposed a lightweight spatial-sequential network explicitly designed for edge-based dermatological screening, prioritising reduced parameter counts and inference latency over absolute accuracy [22]. Yildirim and Kaya introduced dynamic channel-spatial attention modulation within convolutional-recurrent hybrids, demonstrating that selectively filtering diagnostically irrelevant spatial regions—such as those corrupted by hair, ink markings, or specular reflections—before sequential processing improved the effective utilisation of recurrent memory capacity without substantially increasing model size [23]. Kaya et al. extended this line of work with attention-driven spatial-temporal sequence learning, reporting that selective attention over sequential positions improved sensitivity to diagnostically relevant transitions while maintaining a comparatively compact architecture [24].

A recurring theme across this strand of the literature is the tension between architectural sophistication and deployability: many of the highest-accuracy architectures reviewed in this paper—ViT hybrids, GAN-augmented pipelines, multi-stage knowledge-distillation frameworks—achieve their performance through mechanisms difficult to reconcile with edge-deployment constraints. Lightweight CNN-recurrent hybrids, by contrast, occupy a design space in which linear computational complexity, O(N), is preserved, but at the cost of a measurable accuracy gap relative to quadratic-complexity or multi-stage generative architectures. Bridging this gap without sacrificing deployability remains an open problem motivating much of the future-direction discussion in Section VI.

III. COMPARATIVE SYNTHESIS OF REVIEWED ARCHITECTURES

The architectures reviewed in Section II span a wide range of reported accuracy, computational complexity, and imbalance-mitigation strategies, summarised comparatively in Table II. Several patterns emerge. First, there is a broadly observable, though not strict, positive association between reported accuracy and pipeline complexity: the highest-accuracy entries—CycleGAN-augmented EfficientNet [18], ViT-based hybrids [11], and CNN-LSTM with traditional upsampling [12]—generally require multi-stage training, generative components, or attention mechanisms of quadratic complexity, while lightweight CNN-recurrent hybrids cluster in a lower but still clinically meaningful accuracy range with substantially reduced computational overhead.

Second, macro-F1 scores—which, unlike overall accuracy, weight all seven diagnostic classes equally regardless of sample count—are reported far less consistently than accuracy figures, and several high-accuracy entries in Table II report no macro-F1 value at all (denoted NR). This asymmetry in reporting practice itself constitutes a methodological gap: minority-class performance, despite being repeatedly identified as the field's most clinically

consequential challenge, is not yet uniformly treated as a primary evaluation criterion.

Third, among architectures that explicitly address gate desynchronisation through Peephole-gated recurrent units, reported accuracy on medical-imaging sequence tasks (including non-dermoscopic volumes) has reached approximately 88–90% [10], suggesting that cell-state-aware gating may offer architectural benefits that remain comparatively underexplored within dermoscopy-specific classification, where Peephole mechanisms are markedly underrepresented relative to standard LSTM, BiLSTM, and GRU variants.

TABLE II
COMPARATIVE SUMMARY OF REVIEWED ARCHITECTURES FOR HAM10000 SKIN LESION CLASSIFICATION

Architecture	Acc. (%)	Macro-F1	Imbalance Strategy	Key Limitation
Classical ML (handcrafted + SVM)	—	—	None	Segmentation-dependent; weak generalisation
CNN + transfer learning (Xception)	91.47	NR	Transfer learning	Domain-shift vulnerability
CNN-BiLSTM + attention	~85	NR	Augmentation	High recurrent compute overhead
BiLSTM + GAN augmentation	~83	NR	GAN synthesis	FID gap; domain mismatch
CNN-GRU	~77.0	0.635	WCE	Lower minority-class sensitivity
ConvLSTM	75.50	0.620	WCE	Large joint parameter space; data-hungry
Stacked LSTM	80.16	0.658	WCE	Depth without cell-state-aware gating
CNN-Peephole LSTM (class-balanced)	81.32	0.661	WCE	WCE ceiling; no attention or XAI
Peephole LSTM (medical volumes)	~88–90	NR	None	Not dermoscopy-specific
Focal-Loss CNN	~84	0.71	Focal Loss	No recurrent / sequential modelling
CNN-ViT hybrid + attention	95.10	NR	Class weighting	Quadratic complexity $O(N^2)$
CNN-LSTM (upsampled)	97.00	NR	Traditional upsampling	High overfitting risk
CycleGAN + EfficientNet-B3	99.33	0.9685	CycleGAN + Focal Loss	Hyperparameter sensitivity; FID gap

IV. RESEARCH GAPS

Synthesising the literature reviewed in Section II, five interrelated gaps can be identified that collectively constrain the clinical readiness of current automated dermoscopic classification systems.

First, gate desynchronisation in standard LSTM, GRU, and BiLSTM units—arising from the exclusion of the internal cell state from gate computations—remains largely unaddressed within dermoscopy-specific architectures, despite having been rigorously characterised in the broader medical-imaging literature [10]. Given that flattened spatial-to-sequential representations of dermoscopic features require precise temporal synchronisation to retain fine-grained boundary and transition information, this gap represents a direct architectural opportunity.

Second, while Weighted Cross-Entropy is widely adopted as a loss-level imbalance correction, the

literature comparing WCE with Focal Loss indicates that WCE's static inverse-frequency weighting does not suppress gradient contributions from easily classified majority-class samples, potentially imposing a ceiling on minority-class performance that dynamically modulated loss functions are better positioned to overcome [15].

Third, the highest-accuracy architectures identified in Table II—ViT hybrids and CycleGAN-augmented pipelines—incur quadratic computational complexity or multi-stage training requirements difficult to reconcile with deployment in resource-constrained clinical environments, particularly the low- and middle-income regions where dermatologist shortages are most severe [11], [18].
Fourth, most reviewed hybrid CNN-recurrent architectures feed complete spatial feature maps into the recurrent sequence without any mechanism for filtering diagnostically irrelevant artefacts—body hair, ink markings, air bubbles, specular reflections—meaning artefact-derived features occupy the same sequential positions as genuinely diagnostic structures [23]. Relatedly, the overwhelming majority of reviewed studies evaluate exclusively on HAM10000, leaving cross-dataset generalisation to different imaging hardware, lighting conditions, and patient populations largely untested.

Fifth, explainability integration remains both technically and methodologically incomplete. While lightweight architectures can be paired with post-hoc methods such as Grad-CAM, SHAP, and LIME to produce spatial attributions, validation of those attributions against established dermoscopic diagnostic criteria has been demonstrated in only a limited subset of studies, and extension of attribution methods to the sequential dimension of recurrent architectures remains largely unexplored [19]. Additionally, the integration of multi-modal patient metadata—age, biological sex, anatomical site, and Fitzpatrick skin phototype—alongside image-based features has been shown to carry independent diagnostic value but has not been incorporated into most reviewed CNN-recurrent hybrids [25].

V. TOWARD CLASS-BALANCED HYBRID CNN-PEEPHOLE LSTM FRAMEWORKS

The gate desynchronisation gap identified above points toward Peephole-gated recurrent units as a structurally direct response. Unlike standard LSTM gating, in which the forget, input, and output gates are computed solely from the previous hidden state and current input, Peephole LSTM units introduce additional diagonal weight matrices connecting the internal cell state directly to each gate [10]. In compact form, the forget gate becomes:

$$f_t = \sigma(W_f x_t + U_f h_{t-1} + V_f c_{t-1} + b_f)$$

so that each gating decision is informed by the magnitude and content of the accumulated long-term memory at that point in the sequence; the input and output gates are modified analogously using the previous and current cell states, respectively. For a flattened sequence of spatial feature vectors derived from a dermoscopic image, this cell-state awareness allows the gates to adjust their sensitivity based on the morphological context accumulated so far, potentially improving retention of subtle boundary transitions and pigment-network discontinuities that standard gating may discard prematurely.

A recently reported class-balanced hybrid framework combining a lightweight three-block CNN backbone—with progressive filter expansion from 32 to 128 channels—and a Peephole LSTM sequential processor, trained using Weighted Categorical Cross-Entropy on HAM10000, evaluated this approach against six standard recurrent variants (Vanilla LSTM, Stacked LSTM, BiLSTM, CNN-GRU, ConvLSTM, and an Encoder-Decoder LSTM) under identical training conditions [26]. The Peephole-gated variant achieved the highest accuracy (81.32%) and weighted F1-score (82.91%) among the seven architectures, alongside the highest macro-F1 score (0.6614), correctly identifying 575 of the melanoma cases in the test set [26]. Notably, the single-layer Peephole LSTM outperformed both the deeper Stacked LSTM and the bidirectional CNN-BiLSTM variant, suggesting that cell-state-aware gate synchronisation may offer greater benefit for dermoscopic sequence modelling than either

increased recurrent depth or bidirectional processing alone.

While this accuracy remains below the highest figures reported for ViT-hybrid and GAN-augmented pipelines in Table II, the framework's linear computational complexity, O(N), and its reliance on a single end-to-end training pipeline without generative augmentation or large-scale external pretraining position it within the lightweight, edge-deployable design space discussed in Section II-G. It represents a candidate direction for reconciling the gate-desynchronisation and computational-cost gaps identified in Section IV, though its macro-F1 score also illustrates the performance ceiling associated with WCE-only imbalance mitigation discussed in Section II-D.

VI. FUTURE RESEARCH DIRECTIONS

Building on the gaps and the emerging direction discussed above, several concrete research priorities follow directly from the limitations identified in this review.

Replacing or augmenting Weighted Cross-Entropy with Focal Loss represents the most immediately actionable optimisation-level direction. Since Focal Loss's dynamic modulating factor specifically targets the gradient-suppression limitation that WCE does not address, its integration into Peephole-gated hybrid architectures—without otherwise altering the pipeline—could plausibly yield measurable macro-F1 improvements for the rarest classes, such as Dermatofibroma and Vascular Lesions, where current performance remains lowest [15].

Spatial and channel attention modules, inserted between the convolutional backbone and the recurrent sequence processor, represent a second priority. Squeeze-and-excitation-style channel attention applied to the final convolutional feature map, prior to reshaping into a sequence, could selectively strengthen channels associated with malignancy-relevant structures while suppressing channels dominated by hair artefacts, ink markings, or specular reflections—directly addressing the artefact-filtering gap identified in Section IV and building on the attention-modulation findings of Yildirim and Kaya [23].

Explainable AI integration constitutes a third priority, motivated as much by regulatory necessity as by clinical utility. Grad-CAM and Grad-CAM++ attribution maps applied to the convolutional backbone, combined with SHAP-based analysis of the sequential positions most influential to the Peephole LSTM's final hidden state, would extend interpretability across both the spatial and sequential dimensions of hybrid architectures and should be validated against the ABCD rule and seven-point checklist as recommended by Grignaffini et al. [19], rather than assumed to be clinically meaningful by default.

Fourth, incorporating multi-modal patient metadata—age, biological sex, anatomical site, Fitzpatrick phototype, and UV exposure history—as auxiliary inputs concatenated with the recurrent unit's final hidden state prior to classification offers a path toward reducing false positives among ambiguous classes such as Benign Keratosis-like Lesions, which frequently present in older, highly sun-exposed populations with morphological features that visually overlap with malignancy [25].

Fifth, knowledge-distillation frameworks such as HDKD, which transfer global inductive biases from a Vision Transformer teacher into a lightweight CNN-recurrent student, offer a route toward narrowing the accuracy gap with ViT-hybrid architectures while preserving linear computational complexity and mobile deployment feasibility [13].

Finally, the near-universal reliance on HAM10000 as the sole evaluation benchmark across the reviewed literature represents a generalisation gap that future work should address through systematic cross-dataset validation against collections such as ISIC 2020 [27] and institution-specific datasets, ideally combined with federated learning approaches that allow models to be trained collaboratively across multiple clinical institutions without centralising sensitive patient imaging data.

VII. CONCLUSION

This review has traced the methodological evolution of deep learning approaches for multi-class dermoscopic skin lesion classification, from handcrafted feature-based classical pipelines through deep CNNs, hybrid CNN-recurrent architectures, and Vision Transformer-based models. Across this trajectory, two challenges recur with particular persistence: the gate desynchronisation limitation of standard recurrent gating mechanisms, and the severe class imbalance characteristic of benchmark datasets such as HAM10000. While loss-level strategies such as Weighted Cross-Entropy and Focal Loss, data-level strategies such as SMOTE and GAN-based augmentation, and architectural responses such as Peephole-gated recurrent units each address aspects of these challenges, no single reviewed approach resolves them jointly while remaining computationally suitable for resource-constrained clinical deployment. Class-balanced hybrid CNN-Peephole LSTM frameworks represent one promising direction for reconciling cell-state-aware sequential modelling with linear computational complexity, though their integration with Focal Loss, attention-based artefact filtering, multi-modal metadata, and validated explainability mechanisms remains an open and clinically consequential research agenda.

Acknowledgment

The author gratefully acknowledges the guidance and supervision of Mr. Jitender Kumar, Assistant Professor, Department of Computer Science and Engineering, Ganga Institute of Technology and Management, Kablana, Jhajjar, throughout the preparation of the dissertation on which this review is based.

REFERENCES

1. P. Tschandl, C. Rinner, Z. Apalla, G. Argenziano, and H. Kittler, "The HAM10000 dataset, a large collection of multi-source dermatoscopic images of common pigmented skin lesions," Sci. Data, vol. 5, no. 1, p. 180161, Aug. 2018.
2. A. Esteva, B. Kuprel, R. A. Novoa, J. Ko, S. Swetter, and S. Thrun, "Dermatologist-level classification of skin cancer with deep neural networks," Nature, vol. 542, pp. 115–118, Jan. 2019.
3. M. A. Khan, T. Akram, and M. Sharif, "Skin lesion detection and classification using deep learning: A review," Comput. Sci. Rev., vol. 40, p. 100378, May 2021.
4. K. M. Hosny, M. A. Kassem, and M. M. Foaud, "Classification of skin lesions using transfer learning and deep convolutional neural networks," IEEE Access, vol. 6, pp. 20029–20040, Apr. 2018.
5. S. S. Chaturvedi, J. V. Gupta, and P. S. Prasad, "Skin lesion classification using ResNet-101 and transfer learning," Int. J. Inf. Technol., vol. 12, pp. 113–120, Mar. 2020.
6. H. P. Anand, R. S. Karthik, and M. S. Kumar, "Xception transfer learning model for highly accurate classification of skin cancer images," J. Med. Syst., vol. 44, no. 8, pp. 142–151, Jul. 2020.
7. M. H. Yap, G. Pons, and J. Martí, "Automated melanoma detection using multi-scale convolutional neural networks on dermoscopic images," Comput. Med. Imaging Graph., vol. 84, p. 101761, Sep. 2020.
8. J. Reis, M. Vasconcelos, and J. S. Cardoso, "Bilinear LSTM networks with generative adversarial networks for imbalanced skin lesion classification," IEEE J. Biomed. Health Inform., vol. 26, no. 4, pp. 1654–1663, Apr. 2022.
9. Y. Li, J. Zhang, and D. Coiera, "An attention-based CNN-BiLSTM architecture for temporal clinical note modeling and prognosis," J. Biomed. Inform., vol. 131, p. 104101, Jul. 2022.
10. R. Gao, G. Song, and J. M. Fitzpatrick, "Temporal sequence modeling with deep Peephole LSTMs for medical image volume slice prediction," Med. Image Anal., vol. 68, p. 101912, Feb. 2021.
11. Z. Kaya, M. K. Ozturk, H. Balci, S. Sahin, E. Atas, and M. Yildirim, "A transformer-based hybrid decision support system for skin cancer detection on HAM10000 dataset," Diagnostics, vol. 14, no. 3, p. 583, Mar. 2024.
12. S. Mahalle and R. Shinde, "Skin cancer detection using CNN," Int. J. Adv. Res. Comput. Commun. Eng., vol. 13, no. 4, pp. 912–920, Apr. 2024.
13. O. S. El-Assiouti, G. Hamed, D. Khattab, and H. M. Ebied, "HDKD: Hybrid data-efficient knowledge distillation network for medical image

classification," IEEE Access, vol. 12, pp. 98124–98138, Jul. 2024.
14. T. Y. Lin, P. Goyal, R. Girshick, K. He, and P. Dollár, "Focal loss for dense object detection," IEEE Trans. Pattern Anal. Mach. Intell., vol. 42, no. 2, pp. 318–327, Feb. 2020.
15. H. Balci, Z. Kaya, and M. K. Ozturk, "Mitigating class imbalance in automated medical diagnostics using class-weighted focal loss architectures," Diagnostics, vol. 14, no. 2, pp. 112–126, Jan. 2024.
16. L. Chawla, K. Bowyer, L. Hall, and W. Kegelmeyer, "SMOTE: Synthetic minority over-sampling technique," J. Artif. Intell. Res., vol. 16, pp. 321–357, Jun. 2002.
17. J. S. Cardoso, J. Reis, and M. Vasconcelos, "Generative models for domain adaptation and class balancing in medical imaging," Med. Image Anal., vol. 74, p. 102214, Dec. 2021.
18. J. S. Cardoso and J. Reis, "Addressing dermoscopic image imbalance with tone-conditioned generative adversarial networks," PLOS One, vol. 21, no. 4, pp. 112–129, Apr. 2026.
19. R. Grignaffini, A. Abdulredah, and F. Hussein, "Explainable artificial intelligence in skin cancer detection: A systematic review of clinical translation challenges," Comput. Med. Imaging Graph., vol. 112, p. 102340, Sep. 2024.
20. F. Hussein, A. Abdulredah, and S. Singh, "Attention-based CNN-BiLSTM networks for the segmentation and classification of dermoscopic frames," Front. Artif. Intell., vol. 9, p. 1808770, Jan. 2026.
21. A. Abir, M. J. Islam, and S. Karim, "Lightweight CNN structures paired with post-hoc explainable AI (LIME, SHAP, Grad-CAM) for radiological diagnostic validation," J. Digit. Imaging, vol. 37, no. 1, pp. 54–68, Feb. 2024.
22. S. Sahin, E. Atas, and M. K. Ozturk, "A lightweight spatial-sequential network for edge-based dermatological screening," J. Med. Devices Syst., vol. 15, no. 2, pp. 89–102, Feb. 2025.
23. M. Yildirim and Z. Kaya, "Dynamic channel-spatial attention modulation in convolutional-recurrent hybrid networks," IEEE Trans. Med. Imaging, vol. 43, no. 5, pp. 1824–1836, May 2024.
24. Z. Kaya, H. Balci, and S. Sahin, "Attention-driven spatial-temporal sequence learning for skin disease identification," Dermatol. Sci., vol. 115, no. 1, pp. 43–56, Jan. 2025.
25. M. A. Khan, T. Akram, and M. Sharif, "Multimodal skin lesion classification using deep feature fusion and neighborhood component analysis," Diagnostics, vol. 13, no. 4, p. 682, Feb. 2023.
26. A. Saini, "Hybrid deep learning model for skin cancer detection based on CNN and LSTM," M.Tech. dissertation, Dept. Comput. Sci. Eng., Ganga Inst. Technol. Manage., Kablana, Jhajjar, India, 2026.
27. V. Rotemberg, N. Codella, and P. Tschandl, "The ISIC 2020 challenge: Patient-level analysis and ensemble architectures for melanoma classification," Lancet Digit. Health, vol. 3, no. 4, pp. e214–e222, Apr. 2021.