BCSE497J – Project – I

A MULTIMODAL GENERATIVE AI SYSTEM FOR SKIN LESION DIAGNOSIS AND EXPLANATION

Projects and Internship
Class Id: VL2025260102304

Team

22BCE0476	**Aman Chauhan**
22BCE0830	**Arnav Sinha**
22BCE2218	**Akshat Sinha**

Under the Supervision of

Dr. NAGA PRIYADARSINI R

Assistant Professor Sr. Grade 1
Department of Analytics
School of Computer Science and Engineering (SCOPE)

B.Tech.
in
Computer Science and Engineering (Core)

VIT®
Vellore Institute of Technology
(Deemed to be University under section 3 of UGC Act, 1956)

September 2025

ABSTRACT

Early and reliable differentiation of benign and malignant skin lesions is central to dermatology, yet routine practice can be subjective and time-pressured when visual dermoscopy is not paired consistently with patient context such as age, sex, and lesion location. This work proposes a two-stage clinical assistant that first delivers a calibrated diagnosis from fused multimodal inputs and then produces a concise, clinician-style explanation to improve transparency, documentation efficiency, and trust.

In Stage 1 (diagnostic engine), a modern vision backbone encodes dermoscopic images while structured clinical metadata—rendered as natural language—is embedded by a medical text encoder. The visual and textual representations are concatenated and passed to a softmax classifier to yield class probabilities for common lesion categories. Training uses large, publicly available cohorts and includes cross-dataset checks to assess robustness and generalization beyond the development distribution.

In Stage 2 (generative engine), the predicted class, confidence, and available clinical context are transformed into a controlled prompt for a large language model to generate a focused report. The output states the most likely diagnosis, highlights salient visual and contextual cues, lists plausible differentials, and suggests next steps when appropriate, aligning with common dermatology note structures.

Planned evaluations compare CNN and Vision Transformer encoders for images, and general versus clinically pretrained language models for metadata. Additional ablations test multimodal fusion against single-modality baselines and assess explanation quality for clinical relevance and completeness. By coupling strong discriminative performance with faithful, human-readable rationales, the system aims to serve as a reliable second opinion and support earlier melanoma detection within busy workflows.

TABLE OF CONTENTS

1 INTRODUCTION

Dermatology decisions depend on dermoscopic images and simple patient context, yet image-only or text-only approaches miss complementary signals and black-box outputs hinder trust. We build a two-stage assistant tailored to dermoscopy: (i) a diagnostic engine that fuses embeddings from an image encoder and a metadata text encoder to produce calibrated class probabilities, and (ii) a generative reporter that turns the prediction, confidence, and context into a short clinician-style note with justification, differentials, and next steps. The design aims for accuracy, clarity, and auditability so the tool acts as a reliable second opinion without replacing clinical judgment.

1.1 Background

Dermoscopy exposes morphology (networks, streaks, vessels) that benefits from learned features, while age, sex, and lesion site shift priors and disambiguate similar visuals. Multimodal learning combines these: an image encoder for morphology, a text encoder for context, and a simple fusion for classification. Large public cohorts support robust training and cross-dataset checks. Generative models, when grounded on structured outputs and constrained prompts, can produce concise, reviewable notes that make reasoning explicit.

1.2 Motivations

We target earlier, more consistent triage by pairing calibrated probabilities with clear, reusable wording. A modular two-stage design lets encoders improve independently of reporting, simplifies audits and deployments, and leverages open datasets for validation without changing clinical workflows.

1.3 Scope of the Project

In scope: dermoscopic images plus metadata (age, sex, site); comparison of CNN vs ViT image encoders and general vs clinical text encoders; simple fusion with softmax; discrimination and calibration on internal splits and cross-dataset checks; generation of short, clinician-style reports grounded in structured outputs.

Out of scope: histopathology, non-dermoscopic photos, longitudinal follow-up, or therapy suggestions. Deliverables: trained models, ablations, multimodal vs single-modality evidence, and a compact prompt template for consistent summaries.

2 PROJECT DESCRIPTION AND GOALS

We combine a multimodal diagnostic engine with a grounded generative reporter to deliver calibrated probabilities and concise explanations for common lesion categories, acting as a second opinion that improves transparency and reduces note burden.

2.1 Literature Review

Dermoscopic analysis evolved from hand-crafted features to CNNs and ViTs that capture long-range patterns. Lightweight metadata (age, sex, site) meaningfully shifts priors when fused with image features. Simple concatenation remains a strong, auditable fusion baseline; attention-based methods can help but add complexity. Interpretability and documentation are key for adoption; templated, constrained LLM prompting can turn structured outputs into faithful prose. Open cohorts enable rigorous benchmarking, ablations, and generalization checks.

2.2 Gaps Identified

- Limited explainability in multimodal AI: predictions often lack faithful, human-readable rationales clinicians can review, leading to opacity in internal reasoning when fusing visual and textual data.
- Generalizability across diverse datasets: limited transferability of models to unseen, real-world clinical data.
- Integration of generative AI for reporting: need for clinically relevant, comprehensive reports beyond simple diagnoses.
- Benchmarking of fusion techniques: lack of systematic comparative studies on multimodal fusion methods, making audits and incremental updates difficult in constrained settings.
- Ethical considerations and clinical trust: addressing bias while building clinician trust in AI-driven diagnostic tools.
- Real-time performance challenges: optimizing complex models to ensure smooth clinical workflow integration.
- Image-only systems underuse easily available context near decision boundaries, limiting diagnostic robustness.

2.3 Problem Statement

Given a dermoscopic image and metadata (age, sex, site), map to a calibrated distribution over lesion classes and a faithful, human-readable summary of the decision. The system must stay modular so backbones, fusion, and prompting improve independently under clear, reproducible constraints.

2.4 Project Plan

The work plan is organized into five phases:

Phase 1: Data. Assemble splits from ISIC 2019/2020; standardize images; normalize metadata; handle imbalance and quality gates.

Phase 2: Baselines/Model. Train image-only (CNN, ViT), text-only (BERT), and multimodal concatenation; track discrimination and calibration with fixed seeds.

Phase 3: Ablations/Generalization. Compare encoders, modalities, and calibration; test on ISIC 2018; produce reliability plots and confusion matrices.

Phase 4: Reporting. Design a compact, deterministic prompt with guardrails; validate clarity, faithfulness, and completeness via a small rubric.

Phase 5: Packaging. Provide a lightweight inference utility (image+metadata $rightarrow$probs+report), experiment tables, and labeled limitations.

2.5 Activity Chart and Work Breakdown Structure

This activity view follows the five phases already defined in the plan and maps them to concrete tasks, milestones, and deliverables.

3 REQUIREMENT ANALYSIS

This section specifies the functional and non-functional requirements for the multimodal assistant, defines the datasets and preprocessing pipeline, enumerates model and evaluation requirements, and proposes a novel algorithmic improvement that augments fusion, calibration, and explanation quality. The analysis is organized into clear subparts for direct traceability to experiments and deliverables.

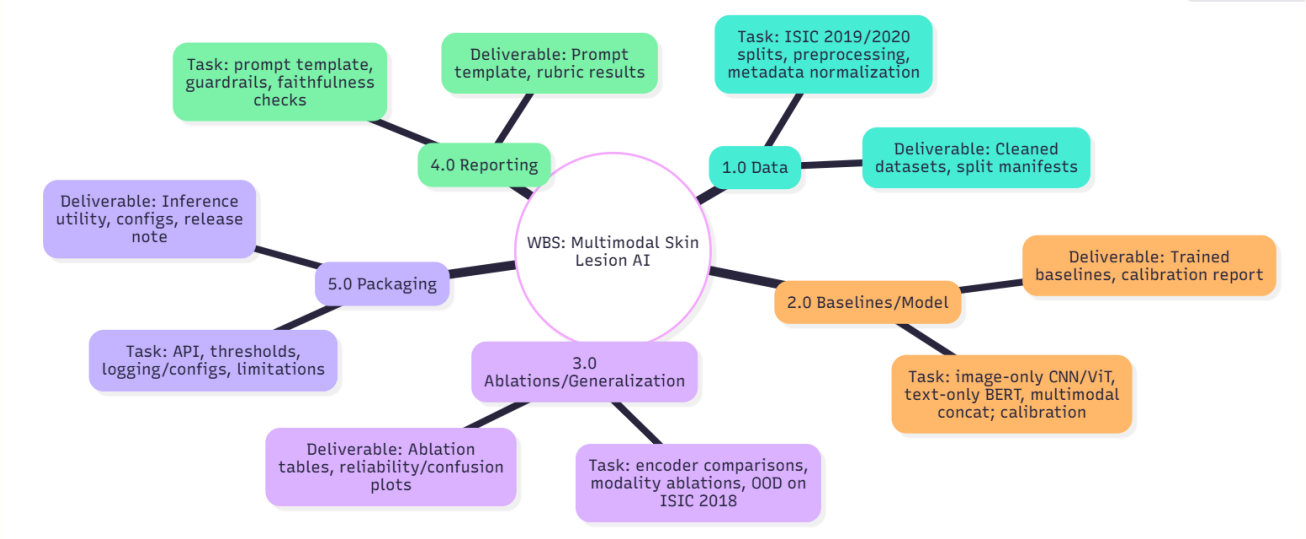

Figure 1: Work Breakdown Structure (WBS).

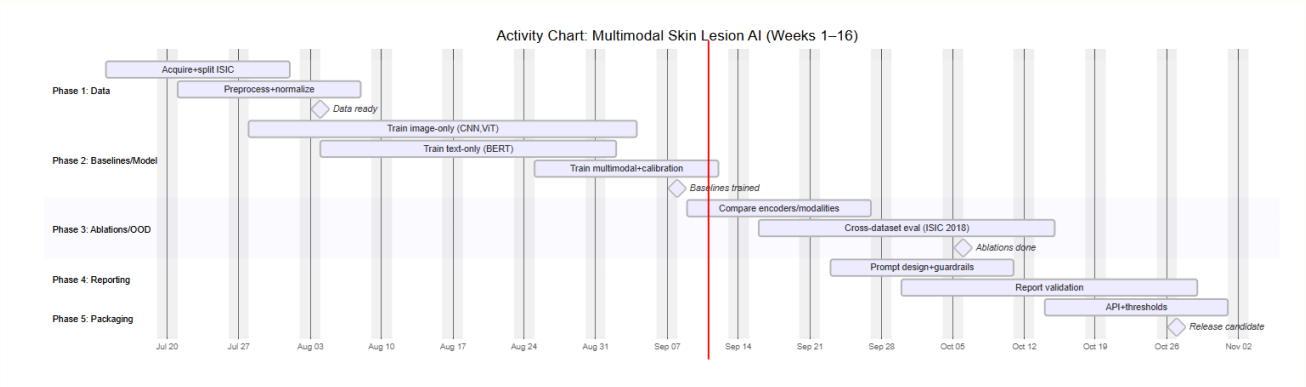

Figure 2: Activity chart (Gantt) aligned to Phases 1–5.

3.1 Datasets and Governance

- Train/Val/Test on ISIC 2019/2020 with patient-level stratification; reserve ISIC 2018 for out-of-distribution tests.
- Metadata in scope: age, sex, lesion site; de-identified; missing fields flagged explicitly.
- Document licenses, splits, transforms, and class definitions; version datasets and configs.

3.2 Functional Requirements

- Input: one dermoscopic image+age, sex, site.
- Output (diagnostic): class probabilities with per-class confidence and calibrated overall score.
- Output (reporter): short note with diagnosis, justification (visual+context), differentials, and next steps.
- Uncertainty/Audit: below-threshold confidence defers to expert; log model/version/seed and preprocessing hashes.

3.3 Data and Preprocessing Requirements

- Imaging: square crop/pad; resize (e.g., 448–512); normalize; light color-preserving augments.
- Metadata: standardize categories; bucketize age if useful; encode as short sentences (e.g., "Male, 62 years, upper back").
- Imbalance/Quality: stratified sampling and/or class weights; exclude corrupted images; represent missing metadata explicitly.

3.4 Model and Baseline Requirements

- Image encoders: one strong CNN and one ViT family model, fine-tuned from public weights.
- Text encoders: compact BERT and a clinical variant for metadata sentences.
- Fusion/Calibration: concatenation+linear softmax as reference; temperature scaling for calibration.
- Practicality: report single-image CPU/GPU latency (mean, p95).

3.5 Novel Algorithmic Improvement: Uncertainty-Guided Cross-Modal Gated Fusion with Prototype Alignment

We introduce UG-CMGF, an uncertainty-aware gate that balances image and metadata features per case and aligns the joint embedding to class prototypes. A selection head defers low-confidence cases

to improve safety. This preserves the simple concatenation baseline while improving robustness and providing grounded signals for the report. See Appendix A for equations, loss terms, and inference flow.

3.6 Evaluation and Quality Requirements

- Metrics: AUROC/AUPRC/Accuracy/F1; ECE and reliability plots; per-class support and confusion matrices.
- Generalization: train/validate on ISIC 2019/2020; evaluate on ISIC 2018; sensitivity analyses by site and sex.
- Safety/Deferral: track deferral rates and error types; require manual review for deferred/low-confidence cases.

3.7 System and Deployment Requirements

- Reproducibility: fixed seeds, deterministic loaders where feasible, exact environment manifests, stored splits.
- Packaging: API takes image+metadata→probabilities+report; CPU/GPU modes; configurable thresholds.
- Monitoring: log hashed inputs, outputs, latency, confidence, and model version; support rollbacks and threshold tuning.

3.8 Risks, Ethics, and Mitigations

- Overconfidence: use temperature scaling and abstention; display calibrated confidence.
- Dataset bias: monitor subgroup metrics; consider re-weighting or thresholds if disparities appear.
- Scope/Privacy: restrict generation to diagnostic justification/differentials; exclude PII from prompts and logs.

3.9 Deliverables

- Trained baselines and UG-CMGF with configs and weights.
- Evaluation report (discrimination, calibration, ablations, OOD).
- Prompt templates and a minimal inference package producing calibrated probabilities and concise reports with deferral.

4 SYSTEM DESIGN

4.1 Architecture Overview

The system has two stages: a multimodal diagnostic engine that fuses image and metadata features into calibrated class probabilities, and a generative reporter that turns structured outputs into a concise clinician-style summary under scope and safety guardrails.

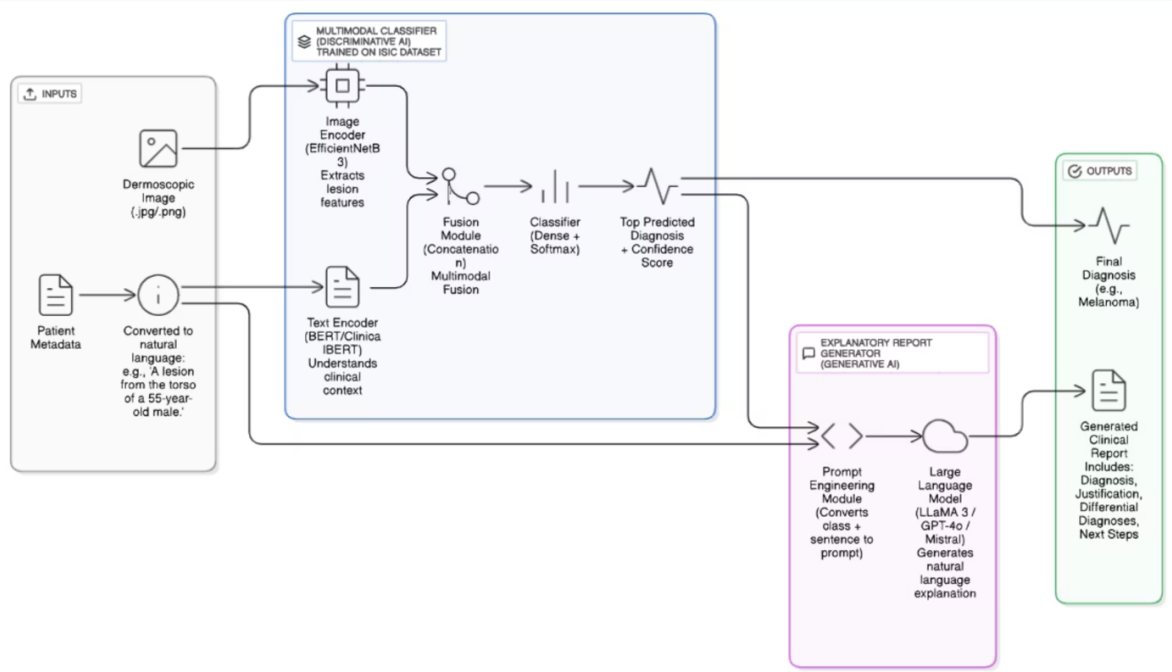

Figure 3: Two-stage system: image/text encoders, gated fusion with prototypes, calibrated softmax, selection head, and controlled prompting.

4.2 Components

Image encoder. CNN or ViT backbone yields z_{img} after pooling+projection.

Metadata encoder. Compact BERT-class model yields z_{text} from short sentences.

Fusion and classifier. Concatenation+linear softmax (reference); UG-CMGF adds uncertainty-gated fusion and prototypes.

Calibration and selection. Temperature scaling for probabilities; selection head supports conservative deferral.

Generative reporter. Structured prompt from class, confidence, and cues produces a focused note.

4.3 Data Flow

1. Validate and normalize image+metadata.
2. Extract z_{img} and z_{text}.
3. Fuse (concatenation or UG-CMGF) and classify; calibrate probabilities.
4. If selected, generate the report; else return a defer message with probability summary.

4.4 Prompting Template (Report Skeleton)

- **Diagnosis**: <top class> (confidence: <value>).
- **Justification**: salient morphology and context summarized from image cues and metadata.
- **Differentials**: 2–3 plausible alternatives with brief rationale.
- **Next steps**: dermoscopy follow-up or escalation guidance consistent with scope.
- **Note**: this summary supports—not replaces—clinical judgment.

4.5 Deployment Considerations

- Stateless inference service exposing a simple API (image + metadata $\rightarrow$ probabilities + report).
- CPU and GPU targets; configurable thresholds for deferral and report length.
- Logging for inputs (hashed), outputs, latency, confidence, and model version for audit.

4.6 Assumptions and Limitations

- Scope limited to dermoscopy and the specified metadata fields; no treatment recommendations.
- Reports remain decision support and require clinician review, especially on deferred or low-confidence cases.

=

5 References

Datasets

- **ISIC 2020:** Contains over 33,000 images and metadata. Focuses on melanoma detection. https://challenge2020.isic-archive.com/
- **ISIC 2019:** Contains over 25,000 images with 8 diagnostic categories. https://challenge2019.isic-archive.com/

- **ISIC 2018:** Contains 10,000 images for lesion classification into 7 categories. https://challenge2018.isic-archive.com/
- **Kaggle Resources:**
  https://www.kaggle.com/datasets/kmader/skin-cancer-mnist-ham10000/code
  https://www.kaggle.com/code/sujitmishra64/melanoma-detection
  https://www.kaggle.com/datasets/fanconic/skin-cancer-malignant-vs-benign/code
- **ISIC Archive Main Page:** https://www.isic-archive.com/
- **NIH Open-i Medical Image Archive:** https://openi.nlm.nih.gov/

References

[1] Chatterjee, S., Fruhling, A., Kotiadis, K., & Gartner, D. (2024). *Towards new frontiers of healthcare systems research using artificial intelligence and generative AI*. Health Systems, 13(4), 263–273. DOI: 10.1080/20476965.2024.2402128

[2] Reddy, S. (2024). Generative AI in healthcare: an implementation science informed translational path on application, integration and governance. Implementation Science, 19:27. https://doi.org/10.1186/s13012-024-01357-9

[3] Saeed, M., Naseer, A., Masood, H., Rehman, S. U., & Gruhn, V. (2023). *The Power of Generative AI to Augment for Enhanced Skin Cancer Classification: A Deep Learning Approach*. IEEE Access. DOI: 10.1109/ACCESS.2023.3332628

[4] La Salvia, M., Torti, E., Leon, R., Fabelo, H., Ortega, S., Martinez-Vega, B., Callico, G. M., & Leporati, F. (2022). *Deep Convolutional Generative Adversarial Networks to Enhance Artificial Intelligence in Healthcare: A Skin Cancer Application*. *Sensors*, 22(16), Article 6145. https://doi.org/10.3390/s22166145

[5] Jütte, L., González-Villà, S., Quintana, J., Steven, M., Garcia, R., & Roth, B. (2024). *Integrating generative AI with ABCDE rule analysis for enhanced skin cancer diagnosis, dermatologist training and patient education*. Frontiers in Medicine, 11, Article 1445318. doi:10.3389/fmed.2024.1445318

[6] Tsai, A.-C., Huang, P.-H., Wu, Z.-C., Wang, J.-F. (2024). *Advanced Pigmented Facial Skin Analysis Using Conditional Generative Adversarial Networks*. 12, 46646–46656. doi:10.1109/ACCESS.2024.3381535

[7] Thoviti, S. H., Varma, B. K., Sai, S. N., & Prasanna, B. L. (2024). *Generative AI Empowered Skin Cancer Diagnosis: Advancing Classification Through Deep Learning*. In 2024 International Conference on IoT Based Control Networks and Intelligent Systems (ICICNIS) (pp. —). IEEE. DOI:10.1109/ICICNIS64247.2024.10823133

[8] Reddy, N. N., & Agarwal, P. (2025). *Diagnosis and Classification of Skin Cancer Using Generative Artificial Intelligence (Gen AI)*. In Generative Artificial Intelligence for Biomedical and Smart Health Informatics (pp. 591–605). Wiley. DOI:10.1002/9781394280735.ch28

[9] Garcia-Espinosa, E., Ruiz-Castilla, J. S., & Garcia-Lamont, F. (2025). *Generative AI and Transformers in Advanced Skin Lesion Classification applied on a mobile device*. International Journal of Combinatorial Optimization Problems and Informatics, 16(2), 158–175. https://doi.org/10.61467/2007.1558.2025.v16i2.1078

[10] Amgothu, S., Lokesh, A., Kumar, S. S., Devipriyanka, S., & Chandu, R. (2025). *Enhanced Skin Lesion Analysis using Generative AI for Cancer Diagnosis*. In 2025 International Conference on Sensors and Related Networks (SENNET) – Special Focus on Digital Healthcare (SENNET 64220), Bengaluru, India, July 24–27, 2025. IEEE. DOI:10.1109/SENNET64220.2025.11136018

[11] Jütte, L., González-Villà, S., Quintana, J., Steven, M., Garcia, R., & Roth, B. (2025). *Generative AI for enhanced skin cancer diagnosis, dermatologist training, and patient education*. In Proceedings of SPIE—International Society for Optics and Photonics (Vol. 13292, p. 132920F), Photonics in Dermatology and Plastic Surgery, BiOS 2025, San Francisco, CA, USA, March 19, 2025. https://doi.org/10.1117/12.3042664

[12] Udrea, A., & Mitra, G. D. (2017). *Generative Adversarial Neural Networks for Pigmented and Non-Pigmented Skin Lesions Detection in Clinical Images*. In 2017 21st International Conference on Control Systems and Computer Science (CSCS), Bucharest, Romania, May 29–31, 2017. IEEE. DOI:10.1109/CSCS.2017.56

[13] Kalaivani, A., Sangeetha Devi, A., & Shanmugapriya, A. (2024). *Generative Models and Diffusion Models for Skin Sore Detection and Treatment*. In 2024 4th International Conference on Ubiquitous Computing and Intelligent Information Systems (ICUIS), Gobichettipalayam, India, December 12–13, 2024. IEEE. DOI:10.1109/ICUIS64676.2024.10866246

[14] Mutepfe, F., Kalejahi, B. K., Meshgini, S., & Danishvar, S. (2021). *Generative Adversarial Network Image Synthesis Method for Skin Lesion Generation and Classification*. Journal of Medical Signals Sensors, 11(4), 237–252. doi:10.4103/jmss.JMSS5320

[15] Innani, S., Dutande, P., Baid, U., Pokuri, V., Bakas, S., Talbar, S., Baheti, B., & Guntuku, S. C. (2023). *Generative adversarial networks based skin lesion segmentation*. Scientific Reports, 13, Article 13467. doi:10.1038/s41598-023-39648-8

[16] Masood, H., Naseer, A., & Saeed, M. (2024). *Optimized Skin Lesion Segmentation: Analysing DeepLabV3+ and ASSP Against Generative AI-Based Deep Learning Approach*. Foundations of Science. Advance online publication. https://doi.org/10.1007/s10699-024-09957-w

[17] Wen, D., Soltan, A. A., Trucco, E., & Matin, R. N. (2024). *From data to diagnosis: skin cancer image datasets for artificial intelligence*. Clinical and Experimental Dermatology, 49(7), 675–685. doi:10.1093/ced/llae112

[18] Mallikharjuna Rao, K., Ghanta Sai Krishna, Supriya, K., & Meetiksha Sorgile. (2025). *LesionAid: vision transformers-based skin lesion generation and classification – A practical review*. Multimedia Tools and Applications. Advance online publication. doi:10.1007/s11042-025-20797-z

[19] Bissoto, A., & Avila, S. (2020). *Improving Skin Lesion Analysis with Generative Adversarial Networks*. In Anais Estendidos da XXXIII Conference on Graphics, Patterns and Images, Workshop de Teses e Dissertações. DOI:10.5753/sibgrapi.est.2020.12986

[20] Bissoto, A., Perez, F., Valle, E., & Avila, S. (2018). *Skin Lesion Synthesis with Generative Adversarial Networks*. In OR 2.0 Context-Aware Operating Theaters, Computer Assisted Robotic Endoscopy, Clinical Image-Based Procedures, and Skin Image Analysis (Lecture Notes in Computer Science, Vol. 11041, pp. 294–302). Springer. https://doi.org/10.1007/978-3-030-01201-432

[21] Marques, A. G., de Figueiredo, M. V. C., Nascimento, J. J. d. C., de Souza, C. T., de Mattos Dourado Júnior, C. M. J., & de Albuquerque, V. H. C. (2024). *New Approach Generative AI Melanoma Data Fusion for Classification in Dermoscopic Images with Large Language Model*. In 2024 37th SIBGRAPI Conference on Graphics, Patterns and Images (SIBGRAPI), Manaus, Brazil, September 30–October 3, 2024. IEEE. DOI:10.1109/SIBGRAPI62404.2024.10716298

[22] Salvi, M., Branciforti, F., Veronese, F., Zavattaro, E., Tarantino, V., Savoia, P., & Meiburger, K. M. (2022). *DermoCC-GAN: A new approach for standardizing dermatological images using generative adversarial networks*. Computer Methods and Programs in Biomedicine, 225, Article 107040. doi:10.1016/j.cmpb.2022.107040

[23] Veeramani, N., & Jayaraman, P. (2025). *A promising AI based super resolution image reconstruction technique for early diagnosis of skin cancer*. Scientific Reports, 15, Article 5084. doi:10.1038/s41598-025-89693-8

[24] Wang, H., Qi, Q., Sun, W., Li, X., Dong, B., & Yao, C. (2023). *Classification of skin lesions with generative adversarial networks and improved MobileNetV2*. International Journal of Imaging Systems and Technology, advance online publication. https://doi.org/10.1002/ima.22880

[25] Ravindranath, R. C., Vikas, K. R., Chandramma, R., Sheela, S., Ruhin Kouser, R., & Dhiraj, C. (2025). *DermaGAN: Enhancing Skin Lesion Classification with Generative Adversarial Networks*. In 2025 International Conference on Emerging Technologies in Computing and Communication (ETCC), June 26–27, 2025. IEEE. DOI:10.1109/ETCC65847.2025.11108424

[26] Ravindranath, R. C., Vikas, K. R., Chandramma, R., Sheela, S., Ruhin Kouser, R., & Dhiraj, C. (2025). *DermaGAN: Enhancing Skin Lesion Classification with Generative Adversarial Networks*. In 2025 International Conference on Emerging Technologies in Computing and Communication (ETCC), June 26–27, 2025. IEEE. DOI:10.1109/ETCC65847.2025.11108424

[27] Al-Rasheed, A., Ksibi, A., Ayadi, M., Alzahrani, A. I. A., Zakariah, M., Ali Hakami, N. (2022). *An Ensemble of Transfer Learning Models for the Prediction of Skin Lesions with Conditional Generative Adversarial Networks*. Diagnostics, 12(12), Article 3145. doi:10.3390/diagnostics12123145

[28] S. Abbasi, M. B. Farooq, T. Mukherjee, J. Churm, O. Pournik, G. Epiphaniou, and T. N. Arvanitis, "Deep learning-based synthetic skin lesion image classification," in *Proc. 34th Medical Informatics Europe Conf. (MIE)*, pp. 1145–1150, IOS Press, 2024.

[29] P. R. Medi, P. Nemani, V. R. Pitta, V. Udutalapally, D. Das, and S. P. Mohanty, "Skinaid: A GAN-based automatic skin lesion monitoring method for IoMT frameworks," in *Proc. 2021 19th OITS Int. Conf. Inf. Technol. (OCIT)*, pp. 200–205, IEEE, 2021.

[30] M. A. Farooq, Y. Wang, M. Schukat, M. A. Little, and P. Corcoran, "Derm-T2IM: Harnessing synthetic skin lesion data via stable diffusion models for enhanced skin disease classification using ViT and CNN," in *Proc. 2024 46th Annu. Int. Conf. IEEE Eng. Med. Biol. Soc. (EMBC)*, pp. 1–5, IEEE, 2024.

[31] A. S. Rao, J. Kim, A. Mu, C. C. Young, E. Kalmowitz, M. Senter-Zapata, D. C. Whitehead, L. Garibyan, A. B. Landman, and M. D. Succi, "Synthetic medical education in dermatology leveraging generative artificial intelligence," *npj Digit. Med.*, vol. 8, no. 1, p. 247, 2025.

[32] P. M. Burlina, W. Paul, P. A. Mathew, N. J. Joshi, A. W. Rebman, and J. N. Aucott, "AI progress in skin lesion analysis," *arXiv preprint arXiv:2009.13323*, 2020.

A UG-CMGF: Method Details

Design overview. We propose **UG-CMGF**, an uncertainty-aware fusion mechanism that learns to gate the contributions of image and metadata features on a per-sample basis, while aligning the joint embedding to class prototypes for stability and interpretability.

- *Uncertainty heads*: attach lightweight evidential heads to both image and text encoders to estimate per-sample uncertainty from intermediate features.
- *Gated fusion*: compute gates g_{img} and g_{text} from uncertainty scores using a small MLP with sigmoid outputs and a soft penalty encouraging $g_{img} + g_{text} \approx 1$. Form the fused embedding:

$$z = g_{img} \cdot z_{img} \; + \; g_{text} \cdot z_{text}.$$

- *Prototype alignment*: maintain class prototypes $\{\mu_c\}$ in the joint space and add a prototypical contrastive loss that pulls samples toward the correct prototype and pushes away from others.
- *Selective prediction*: a selection head $s(z)$ estimates whether to auto-report or defer; low $s(z)$ triggers a "review required" path and conservative prompting.
- *Grounded explanation*: expose top prototypes and gate values to the reporting prompt so rationales emphasize morphology when g_{img} is high and contextual priors when g_{text} dominates.

Training objective.

$$\mathcal{L} = \mathcal{L}cls \; + \; \lambda_1 \mathcal{L}proto \; + \; \lambda_2 \mathcal{L}gate \; + \; \lambda_3 \mathcal{L}sel \; + \; \lambda_4 \mathcal{L}_{cal},$$

where $\mathcal{L}cls$ is cross-entropy, $\mathcal{L}proto$ is the prototypical contrastive term, $\mathcal{L}gate$ regularizes complementary gates and robustness to missing metadata, $\mathcal{L}sel$ trains the selection head using confident-correct targets, and $\mathcal{L}_{cal}$ captures calibration (or a temperature-scaling proxy).

Inference flow. Encode image and metadata, estimate uncertainty, compute gates, form z, and output probabilities. If $s(z)$ is below threshold or the maximum probability is low, return a defer message. Otherwise, compose a structured prompt with class, confidence, salient visual tokens, metadata cues, gate values, and nearest prototypes to generate the concise report.

Expected benefits. UG-CMGF down-weights noisy metadata when it conflicts with strong visual evidence and elevates contextual priors when images are ambiguous. Prototype alignment stabilizes boundaries and supports semantically grounded justifications. The selection head provides principled abstention for safer deployment.

A.1 Ablation Protocols

Compare: (i) concatenation baseline vs UG-CMGF, (ii) with/without prototype loss, (iii) with/without selection head, (iv) uncertainty-free gates vs uncertainty-guided gates, and (v) image-only and text-only controls. Report discrimination, calibration, and deferral-quality metrics.