Remik: Riset dan E-Jurnal Manajemen Informatika Komputer
Volume 10, Number 3, Agustus 2026
http://doi.org/10.33395/remik.v10i1.16539
e-ISSN : 2541-1330
p-ISSN : 2541-1332

Komparasi CNN dan VLM untuk Diagnosis Penyakit Kulit Tropis pada *Subset Dataset* Derm1M

[1]Fahreza Rizky Pradana, [2] Agung Nilogiri, [3]Qurrota A'yun
[1,2, 3]Universitas Muhammadiyah Jember
Jember, Indonesia

[1]rizkyfahreza937@gmail.com, [2]agungnilogiri@unmuhjember.ac.id,
[3]qurrota.ayun@unmuhjember.ac.id

*Penulis Korespondensi
Diajukan : 07/07/2026
Diterima : 25/07/2026
Dipublikasi : 01/08/2026

ABSTRAK
Penyakit kulit tropis merupakan beban kesehatan yang besar di Indonesia, sementara distribusi dokter spesialis dermatologi yang tidak merata membuat akses diagnosis dini di daerah terpencil masih sangat terbatas. Penelitian ini bertujuan membandingkan performa klasifikasi dan kualitas penjelasan visual antara tiga model kecerdasan buatan, yaitu EfficientNet-B4 sebagai *baseline* CNN, LLaVA-7B *zero-shot*, dan LLaVA-7B dengan *fine-tuning* QLoRA, dalam mengenali lima kelas penyakit kulit tropis pada *subset dataset* Derm1M. Sebanyak 5.500 gambar dipilih secara *stratified* dari lima kelas target. Ketiga model dilatih dan diuji pada pembagian data yang identik, kemudian dibandingkan menggunakan metrik akurasi, presisi, *recall*, *F1-score* makro, serta tiga metrik *explainability* Grad-CAM: *Average Drop*, *Average Increase*, dan *Structural Similarity Index Measure* (SSIM) intra-kelas. LLaVA-7B QLoRA mencapai akurasi tertinggi (77,45%; F1-score makro 77,62%), diikuti EfficientNet-B4 (67,45%) dan LLaVA *zero-shot* (29,82%). Namun, dari sisi kualitas penjelasan visual, EfficientNet-B4 justru mencatat *Average Increase* (23,33%) dan SSIM intra-kelas (0,533) tertinggi di antara ketiga model, sedangkan LLaVA QLoRA tidak unggul pada satu metrik XAI pun secara individual meski menunjukkan profil paling seimbang. Temuan ini menunjukkan bahwa peningkatan akurasi klasifikasi pada VLM yang di-*fine-tune* tidak selalu diikuti oleh peningkatan kualitas *explainability*, sehingga pemilihan model perlu mempertimbangkan *trade-off* antara akurasi, kualitas penjelasan visual, dan kebutuhan komputasi; EfficientNet-B4 tetap menjadi pilihan yang praktis untuk fasilitas kesehatan dengan keterbatasan sumber daya. Hasil ini merupakan temuan pada *subset dataset* internasional Derm1M dan belum tentu merepresentasikan kondisi klinis lokal Indonesia.

Kata Kunci: CNN, Penyakit Kulit Tropis, *Vision Language Model*, QLoRA, Grad-CAM.

I. PENDAHULUAN
Penyakit kulit merupakan salah satu kelompok penyakit yang memberikan beban kesehatan masyarakat yang besar secara global. Data terbaru dari studi *Global Burden of Disease* 2021 mencatat lebih dari 2 miliar kasus prevalensi dan sekitar 4,7 miliar kasus insiden secara global dalam satu tahun (Li et al., 2025). Di Indonesia, kondisi ini diperparah oleh distribusi dokter spesialis kulit dan kelamin yang tidak merata. Rasio dokter kulit terhadap jumlah penduduk dilaporkan sekitar 1:100.000, sementara sebagian besar dokter kulit terkonsentrasi di kota-kota besar (Perdoski, 2023). Penelitian di Klinik Pratama Panti Siwi Jember mencatat bahwa lebih dari separuh kunjungan pasien (51,7%) merupakan penyakit kulit infeksi (Lidjaja, 2022), sedangkan data di RSUD Jagakarsa Jakarta menunjukkan dermatitis sebagai diagnosis non-infeksi terbanyak (34,4%) dari 1.066 kasus yang tercatat pada periode 2023–2024 (Alfadli & Khairunisa, 2024).

Remik: Riset dan E-Jurnal Manajemen Informatika Komputer
Volume 10, Number 3, Agustus 2026 e-ISSN : 2541-1330
http://doi.org/10.33395/remik.v10i1.16539 p-ISSN : 2541-1332

Perkembangan *deep learning*, khususnya *Convolutional Neural Network* (CNN), membuka peluang baru dalam mendukung diagnosis penyakit kulit berbasis citra. Penelitian terdahulu menunjukkan bahwa CNN mampu mencapai performa klasifikasi yang kompetitif dengan dokter spesialis pada tugas klasifikasi *melanoma* dan kanker kulit (Haenssle et al., 2018; Esteva et al., 2017). Namun, CNN konvensional memiliki keterbatasan dalam memberikan penjelasan langsung yang mudah dipahami dan diverifikasi secara medis atas dasar keputusannya. CNN juga umumnya berfokus pada informasi visual dan tidak secara alami mengintegrasikan konteks klinis berbasis teks (Zhang et al., 2022; Daneshjou et al., 2022).

VLM dirancang untuk memahami dan mengintegrasikan informasi dari modalitas visual dan teks secara simultan, sehingga berpotensi menghasilkan keluaran yang tidak hanya berupa prediksi tetapi juga deskripsi berbasis bahasa mengenai informasi visual yang mendasari prediksi tersebut (Liu et al., 2023). Penelitian SkinGPT-4 menunjukkan potensi model *multimodal* yang diadaptasi untuk dermatologi dalam mendukung analisis kondisi kulit melalui keluaran berbasis visual dan bahasa yang lebih kontekstual (Zhou et al., 2024). Selain itu, *dataset* Derm1M yang baru dirilis menyediakan lebih dari satu juta gambar dermatologi berlabel dengan penyelarasan ontologi klinis, termasuk beberapa penyakit infeksi yang juga dapat ditemukan di wilayah tropis. Meskipun demikian, *dataset* ini bersifat internasional dan tidak secara khusus dikumpulkan dari populasi pasien Indonesia (Yan et al., 2025).

Meskipun demikian, penelitian yang secara sistematis membandingkan CNN dengan VLM, khususnya dalam skenario *zero-shot* dan *fine-tuning*, serta mengevaluasi performa klasifikasi dan kualitas penjelasan visual secara kuantitatif pada data dermatologi masih terbatas. Di Indonesia, penelitian dengan pendekatan tersebut juga masih belum banyak ditemukan. penelitian ini menggunakan *subset dataset* publik Derm1M dan bukan data klinis yang dikumpulkan langsung dari pasien di fasilitas kesehatan Indonesia, sehingga hasil penelitian perlu dipahami sebagai temuan awal berbasis *dataset* internasional. Penelitian ini bertujuan membandingkan performa klasifikasi EfficientNet-B4 sebagai *baseline* CNN, LLaVA-7B *zero-shot*, dan LLaVA-7B yang di-*fine-tuning* menggunakan QLoRA pada *subset dataset* Derm1M. Selain performa klasifikasi, penelitian ini juga mengevaluasi kualitas penjelasan visual berbasis Grad-CAM menggunakan metrik *Average Drop*, *Average Increase*, dan SSIM intra-kelas.

II. STUDI LITERATUR

Penelitian Terdahulu

Penelitian klasifikasi penyakit kulit berbasis *deep learning* telah berkembang pesat dalam satu dekade terakhir. Esteva et al. (2017) menunjukkan kemampuan CNN berbasis Inception v3 dalam mengklasifikasikan kanker kulit secara kompetitif dengan dermatologis, yang kemudian diperkuat oleh Haenssle et al. (2018) melalui perbandingan CNN dengan 58 dermatologis pada klasifikasi melanoma. Perkembangan CNN selanjutnya menghasilkan EfficientNet yang diperkenalkan oleh Tan & Le (2019) melalui pendekatan *compound scaling* untuk menyeimbangkan kedalaman, lebar, dan resolusi jaringan.

Perkembangan model *multimodal* turut didorong oleh CLIP (Radford et al., 2021) melalui pembelajaran representasi visual dan bahasa menggunakan *contrastive learning*, yang kemudian menjadi salah satu dasar perkembangan VLM modern seperti LLaVA (Liu et al., 2023). LLaVA mengintegrasikan *vision encoder* dengan *large language model* sehingga mampu memproses informasi visual dan bahasa secara *multimodal*. Untuk mengadaptasi model berparameter besar dengan sumber daya terbatas, (Dettmers et al., 2023) memperkenalkan QLoRA yang menggabungkan LoRA dengan kuantisasi 4-bit.

Pada bidang dermatologi, Zhou et al. (2024) memperkenalkan SkinGPT-4 sebagai VLM yang diadaptasi untuk analisis kondisi kulit dan menghasilkan keluaran berbasis bahasa. Studi tersebut menunjukkan potensi VLM dalam mengintegrasikan informasi visual dan bahasa, tetapi hasilnya tidak dapat dibandingkan secara langsung dengan penelitian ini karena perbedaan dataset, tugas, dan konfigurasi evaluasi. Selain itu, SkinGPT-4 tidak berfokus pada perbandingan *visual*

e-ISSN : 2541-1330
p-ISSN : 2541-1332

explainability antara CNN dan VLM menggunakan Grad-CAM. Meskipun mencakup beberapa kondisi yang relevan dengan wilayah tropis, Derm1M bersifat internasional dan bukan dataset klinis yang dikumpulkan secara khusus dari populasi Indonesia.

Dalam aspek *explainability*, Selvaraju et al. (2017) memperkenalkan Grad-CAM sebagai metode *post-hoc* berbasis gradien yang menghasilkan *heatmap* untuk menunjukkan wilayah citra yang berkontribusi terhadap prediksi model. Pendekatan ini berbeda dari penjelasan tekstual VLM yang disampaikan dalam bahasa alami. Penjelasan tekstual dapat memberikan deskripsi atau interpretasi mengenai kondisi yang diamati, tetapi tidak secara langsung menunjukkan lokasi spasial pada citra yang mendasari prediksi. Sebaliknya, Grad-CAM memvisualisasikan fokus spasial model sehingga dapat dianalisis secara kuantitatif. (Chanda et al., 2025) menunjukkan pemanfaatan Grad-CAM bersama *eye-tracking* dalam konteks klasifikasi melanoma, tetapi belum membandingkan *visual explainability* antara CNN dan VLM. Pentingnya interpretabilitas dalam sistem AI medis juga menjadi perhatian dalam pengembangan teknologi AI yang dapat dipercaya di bidang klinis (Daneshjou et al., 2022).

Berdasarkan penelitian terdahulu tersebut, masih terbatas penelitian yang secara langsung membandingkan CNN dan VLM pada kondisi dermatologi yang sama sekaligus mengevaluasi *visual explainability* keduanya secara kuantitatif. Posisi penelitian ini terletak pada perbandingan EfficientNet-B4 sebagai *baseline* CNN, LLaVA-7B *zero-shot*, dan LLaVA-7B yang di-*fine-tuning* menggunakan QLoRA pada *subset* Derm1M. Selain performa klasifikasi, penelitian ini mengevaluasi *visual explainability* berbasis Grad-CAM menggunakan metrik Average Drop, Average Increase, dan SSIM intra-kelas. Dengan demikian, penelitian ini tidak membandingkan keseluruhan kemampuan *explainability* CNN dan VLM, melainkan secara khusus membandingkan fokus visual pada komponen visual kedua pendekatan menggunakan metode evaluasi yang sama.

Tabel 1 Ringkasan Perbandingan Penelitian Terdahulu yang Relevan

Peneliti (Tahun)	Model/Metode	Fokus	Dataset	Gap terhadap Penelitian Ini
Esteva dkk. (2017)	CNN (*Inception* v3)	Klasifikasi kanker kulit	ISIC, TCGA	Tanpa mekanisme *explainability*; fokus kanker, bukan kulit tropis
Haenssle dkk. (2018)	CNN vs. 58 dermatologis	Klasifikasi *melanoma*	ISIC	Tanpa *explainability*; tidak melibatkan VLM
Zhou dkk. (2024)	SkinGPT-4	VLM untuk analisis dermatologi dan keluaran bahasa	*Dataset* studi	Model besar (13 miliar parameter); tanpa analisis Grad-CAM komparatif
Yan dkk. (2025)	Derm1M	Dataset dermatologi berskala besar	Derm1M (>1 juta citra)	Tanpa evaluasi *head-to-head* untuk kulit tropis Indonesia
Chanda dkk. (2025)	CNN + Grad-CAM + *eye-tracking*	*Explainability* klasifikasi *melanoma*	*Dataset melanoma*	Fokus tunggal pada *melanoma*; tanpa

e-ISSN : 2541-1330
p-ISSN : 2541-1332

				perbandingan CNN vs. VLM
Penelitian ini (2026)	EfficientNet-B4 vs. LLaVA-7B *zero-shot* vs. LLaVA-7B QLoRA + Grad-CAM	Klasifikasi dan *visual explainability*	*Subset* Derm1M, 5 kategori kondisi dermatologi	Membandingkan performa klasifikasi dan *visual explainability* CNN dan VLM

III. METODE

Penelitian ini merupakan penelitian eksperimental komparatif berbasis komputasi. Seluruh model menggunakan *dataset* dan *test set* yang sama dengan pembagian data yang seragam, sehingga perbedaan performa dapat dibandingkan secara lebih terkontrol terhadap perbedaan pendekatan model (Kleinedlerová & Kleinedler, 2022).

Dataset dan Pembagian Data

Data penelitian bersumber dari *dataset* publik Derm1M (Yan et al., 2025), yang terdiri atas lebih dari satu juta citra dermatologi berlabel. Label yang digunakan mengikuti anotasi asli *dataset* dan belum diverifikasi ulang oleh dokter spesialis kulit Indonesia, sehingga kualitas label bergantung pada anotasi sumber. Karena keterbatasan komputasi, penelitian menggunakan 5.500 citra yang dipilih secara acak dengan jumlah seimbang dari lima kategori, yaitu eksim dan dermatitis, infeksi jamur (*tinea*), *urtikaria* (biduran), jerawat (*acne*), serta *nevus* dan *melanoma*, masing-masing sebanyak 1.100 citra. Kategori *nevus* dan *melanoma* digabungkan untuk menjaga keseimbangan jumlah data, meskipun keduanya memiliki perbedaan klinis sehingga sebaiknya dipisahkan pada penelitian selanjutnya. Pemilihan kategori mempertimbangkan data prevalensi Kementerian Kesehatan (2023) serta studi klinis di Jember (Lidjaja, 2022) dan Jakarta (Alfadli & Khairunisa, 2024). serta ketersediaan citra yang memadai pada Derm1M.

Data dibagi dengan metode 5-Fold *Stratified* K-Fold dan 10% data (550 citra) ditetapkan sebagai *test set*. Setiap pembagian menghasilkan 3.960 citra pelatihan dan 990 citra validasi. Distribusi data ditampilkan pada Tabel 2.

Tabel 2 Distribusi *Subset Dataset* Derm1M per Kelas dan Partisi

Kelas Penyakit	*Pool/Per-Fold Train*	*Per-Fold Val*	*Test*	Total
Eksim Dermatitis	792	198	110	1.100
Infeksi Jamur (*Tinea*)	792	198	110	1.100
Biduran (*Urtikaria*)	792	198	110	1.100
Jerawat (*Acne*)	792	198	110	1.100
Tahi Lalat *(Melanoma)*	792	198	110	1.100
Total	**3.960**	**990**	**550**	**5.500**

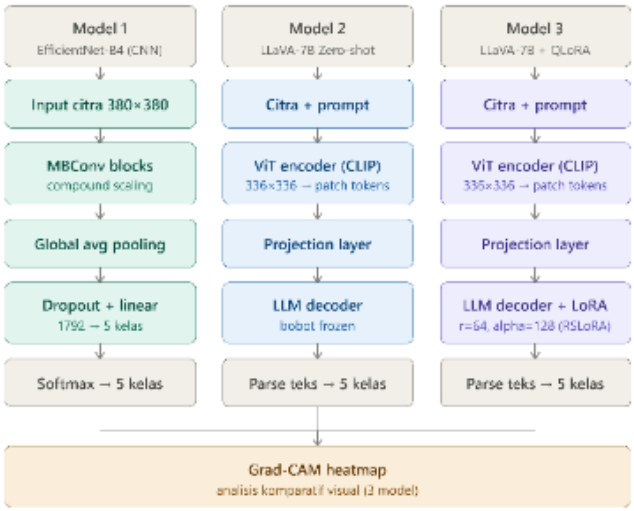
Sumber: *Dataset* Derm1M (Yan et al., 2025), diolah penulis

Preprocessing Data

Sebelum pembagian data, citra diproses untuk mengurangi gangguan visual, meliputi penghapusan rambut, artefak tepi *dermoskopik*, dan penyesuaian pencahayaan. Selanjutnya dilakukan penyaringan kualitas berdasarkan resolusi, karakteristik area kulit, struktur spasial, dan tekstur untuk mengurangi citra yang tidak sesuai. Pemrosesan kemudian disesuaikan dengan model. Pada CNN EfficientNet-B4, citra diubah menjadi ukuran 380×380 *piksel* dan dinormalisasi menggunakan standar ImageNet, sedangkan pada VLM LLaVA-7B citra diproses menggunakan CLIP *ImageProcessor* dengan ukuran 336×336 *piksel*.

Arsitektur dan Implementasi Model

Remik: Riset dan E-Jurnal Manajemen Informatika Komputer
Volume 10, Number 3, Agustus 2026
http://doi.org/10.33395/remik.v10i1.16539
e-ISSN : 2541-1330
p-ISSN : 2541-1332



Gambar 1 Diagram Arsitektur Komparatif Tiga Model
Sumber Gambar : Penulis

Penelitian membandingkan tiga model yang mewakili pendekatan CNN dan *Vision-Language* Model (VLM), yaitu EfficientNet-B4, LLaVA-7B *zero-shot*, dan LLaVA-7B QLoRA.

EfficientNet-B4 digunakan sebagai model CNN dengan bobot awal dari ImageNet dan dilatih melalui dua tahap, yaitu pelatihan bagian klasifikasi dan dilanjutkan dengan pelatihan seluruh model. Bagian klasifikasi disesuaikan untuk mengenali lima kelas penyakit. Pelatihan menggunakan Focal Loss dan Label Smoothing, serta augmentasi MixUp, CutMix, dan CoarseDropout untuk meningkatkan kemampuan model dalam mengenali variasi citra.

LLaVA-7B *zero-shot* digunakan tanpa pelatihan tambahan dengan memanfaatkan bobot *pretrained* dari *Hugging Face*. Model diberikan lima kategori penyakit sebagai pilihan jawaban, kemudian probabilitas masing-masing kategori dihitung berdasarkan tingkat kesesuaian model terhadap label tersebut. Pendekatan ini digunakan untuk menghindari kesalahan ketika model menghasilkan jawaban dalam bentuk teks bebas.

LLaVA-7B QLoRA merupakan versi LLaVA yang disesuaikan melalui pelatihan tambahan dengan metode QLoRA yang lebih hemat sumber daya komputasi. Model dilatih selama lima *epoch* menggunakan data berupa pasangan citra dan label penyakit. Pelatihan dihentikan berdasarkan kinerja validasi untuk mencegah penurunan kemampuan model. Evaluasi menggunakan *test set* dan metode penilaian probabilitas yang sama dengan model *zero-shot*, sehingga hasil ketiga model dapat dibandingkan secara konsisten. Ringkasan konfigurasi utama model disajikan pada Tabel 3.

Tabel 3 Konfigurasi *Hyperparameter* Pelatihan Ketiga Model

Parameter	**EfficientNet-B4**	**LLaVA-7B *Zero-shot***	**LLaVA-7B + *QLoRA***
Batch size	32	1	2
Learning rate	3-tier: 1e-6 / 1e-5 / 3e-5	—	1e-4 (LoRA)
Epoch	3 + 35	—	5
Optimizer	AdamW	—	PagedAdamW
Quantization	—	4-bit	4-bit NF4
LoRA r / α	—	—	64 / 128
Ukuran *Input*	380×380 px	336×336 px	336×336 px
Loss Function	*Focal Loss* (γ=2,0)	—	Cross-Entropy

Sumber: Konfigurasi eksperimen penelitian

Remik: Riset dan E-Jurnal Manajemen Informatika Komputer
Volume 10, Number 3, Agustus 2026
http://doi.org/10.33395/remik.v10i1.16539
e-ISSN : 2541-1330
p-ISSN : 2541-1332

Analisis *Explainability* dengan Grad-CAM

Analisis *explainability* menggunakan *Gradient-weighted Class Activation Mapping* (Grad-CAM) diterapkan pada ketiga model dengan menyesuaikan arsitektur masing-masing. Pada EfficientNet-B4, Grad-CAM diterapkan pada lapisan konvolusi terakhir sebelum proses *pooling*. Pada LLaVA-7B *zero-shot* dan QLoRA, analisis difokuskan pada bagian *vision encoder* untuk mengisolasi kontribusi citra tanpa melibatkan *decoder* bahasa. Gradien dihitung berdasarkan skor kemungkinan (*likelihood*) kelas yang diprediksi, sehingga *heatmap* menunjukkan area citra yang berkontribusi terhadap keputusan klasifikasi. Analisis dilakukan secara kualitatif pada 30 sampel representatif dan secara kuantitatif menggunakan *Average Drop*, *Average Increase*, dan SSIM intra-kelas.

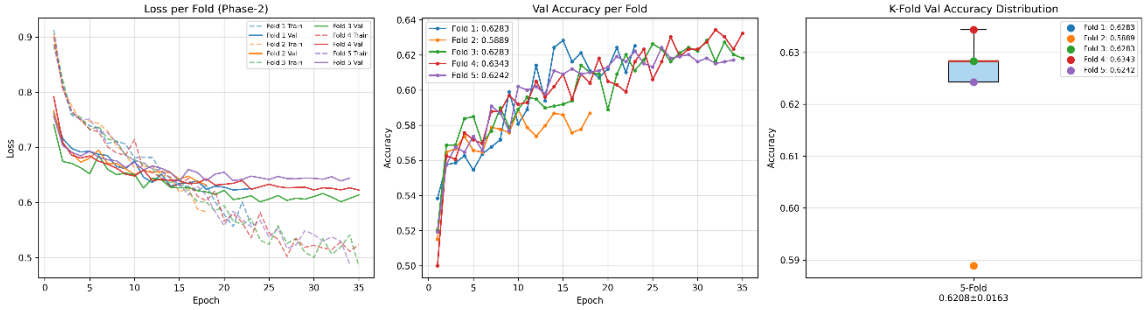
Metrik Evaluasi

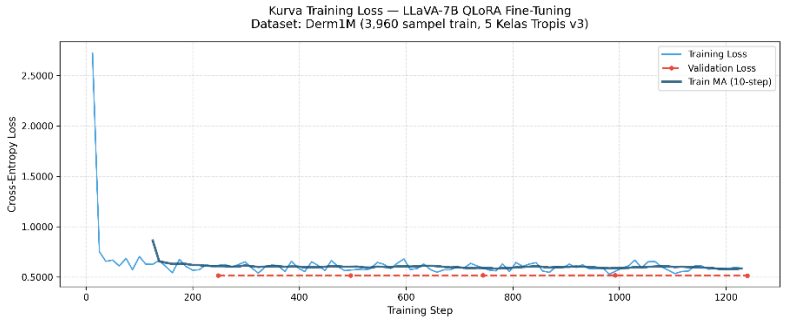
Evaluasi klasifikasi dilakukan menggunakan akurasi, presisi makro, *recall* makro, dan *F1-score* makro berdasarkan *confusion matrix* pada *test set*. Evaluasi *explainability* menggunakan *Average Drop* untuk mengukur perubahan kepercayaan model ketika hanya area yang dianggap penting dipertahankan, *Average Increase* untuk mengukur peningkatan kepercayaan setelah latar belakang dihilangkan, dan SSIM intra-kelas untuk menilai konsistensi pola *heatmap* antar gambar dalam kelas yang sama. Nilai *Average Drop* dan *Average Increase* yang lebih rendah menunjukkan interpretasi yang lebih baik, sedangkan SSIM yang lebih tinggi menunjukkan konsistensi yang lebih baik.

Perbedaan performa klasifikasi antar model diuji menggunakan uji McNemar berpasangan dengan koreksi kontinuitas pada tingkat signifikansi α=0,05, menggunakan *test set* yang sama (n=550). Sementara itu, perbandingan metrik *explainability* disajikan sebagai titik estimasi tanpa uji statistik formal dan menjadi salah satu keterbatasan penelitian yang dibahas pada bagian Pembahasan.

IV. HASIL DAN PEMBAHASAN

Pada proses pelatihan, EfficientNet-B4 menunjukkan stabilitas yang cukup baik di seluruh *fold* dengan akurasi validasi berkisar 58,89%–63,43%. *Fold* 4 mencatat akurasi validasi tertinggi (63,43%) dan dipilih sebagai *checkpoint* terbaik. Ketika diuji pada *test set*, model ini menghasilkan akurasi 67,45% – lebih tinggi dari nilai validasinya sendiri yang mengindikasikan kemampuan generalisasi yang baik terhadap data *unseen*. Proses pelatihan LLaVA-7B QLoRA memakan waktu sekitar 25,8 jam untuk 5 *epoch*, dengan *cross-entropy loss* yang konvergen dari nilai awal 2,7 ke 0,55 dalam 1.250 langkah.

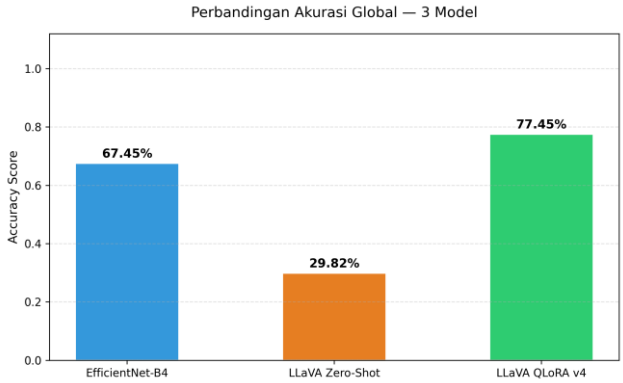




Remik: Riset dan E-Jurnal Manajemen Informatika Komputer
Volume 10, Number 3, Agustus 2026
http://doi.org/10.33395/remik.v10i1.16539
e-ISSN : 2541-1330
p-ISSN : 2541-1332

Gambar 2 Kurva Pelatihan
Sumber : Hasil eksperimen penelitian

Perbandingan performa ketiga model pada *test set* mengungkapkan hierarki yang jelas. LLaVA QLoRA mencapai akurasi tertinggi 77,45%. EfficientNet-B4 berada di posisi kedua dengan akurasi 67,45%. LLaVA *zero-shot* menempati posisi terakhir dengan akurasi hanya 29,82%, sebagaimana dirangkum pada Tabel 4 dan divisualisasikan pada Gambar 3.



Gambar 3 Perbandingan Akurasi Global
Sumber: Hasil eksperimen penelitian

Tabel 4 Ringkasan Metrik Evaluasi *Test Set*

Model	*Accuracy*	*Precision (Macro)*	*Recall (Macro)*	*F1-Score (Macro)*
EfficientNet-B4	67,45%	67,47%	67,45%	66,82%
LLaVA *Zero-Shot*	29,82%	33,51%	29,82%	22,99%
LLaVA QLoRA	77,45%	77,97%	77,45%	77,62%

Hasil uji *McNemar* (metodologi pada bagian Metode) menunjukkan ketiga perbandingan signifikan secara statistik ($p < 0,001$), sebagaimana Tabel 5 mengonfirmasi urutan performa LLaVA QLoRA > EfficientNet-B4 > LLaVA *zero-shot* bukan variasi acak, melainkan perbedaan konsisten pada pola kebenaran prediksi masing-masing model.

Tabel 5 Hasil Uji *McNemar* untuk Perbandingan Berpasangan Ketiga Model

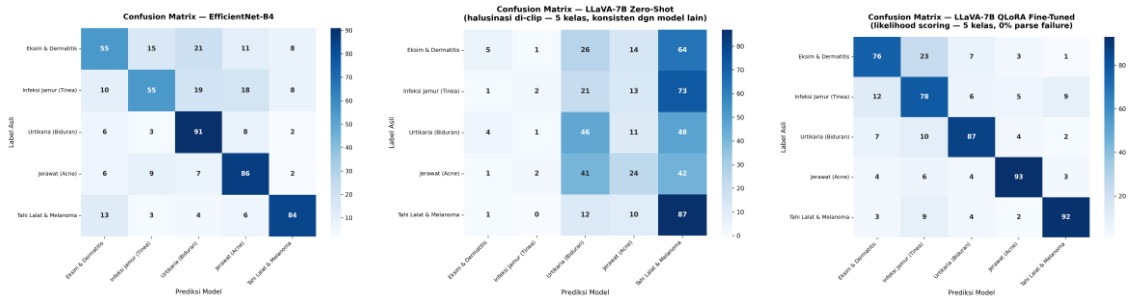
Pasangan Model	χ^2	*p-value*	Keputusan
EfficientNet-B4 vs. LLaVA *Zero-Shot*	149,95	< 0,001	Signifikan
EfficientNet-B4 vs. LLaVA QLoRA	20,11	< 0,001	Signifikan
LLaVA *Zero-Shot* vs. LLaVA QLoRA	208,96	< 0,001	Signifikan

Sumber: Hasil eksperimen penelitian

Pada konfigurasi dan *dataset* yang digunakan, temuan ini mengindikasikan bahwa VLM generalis tanpa adaptasi domain berisiko tinggi menghasilkan diagnosis yang tidak akurat untuk penyakit kulit tropis. Kegagalan LLaVA *zero-shot* bukan sekadar masalah akurasi rendah, melainkan indikasi bias prediksi yang ekstrem: 314 dari 550 gambar uji (57%) diprediksi sebagai kelas Tahi Lalat & *Melanoma*, jauh melampaui proporsi sebenarnya (20%), sementara kelas Eksim hampir diabaikan total (hanya 5 dari 110 terdeteksi benar, dengan 105 di antaranya salah diprediksi sebagai kelas lain) dan Infeksi Jamur bahkan lebih parah (hanya 2 dari 110 benar, 108 salah). Pola

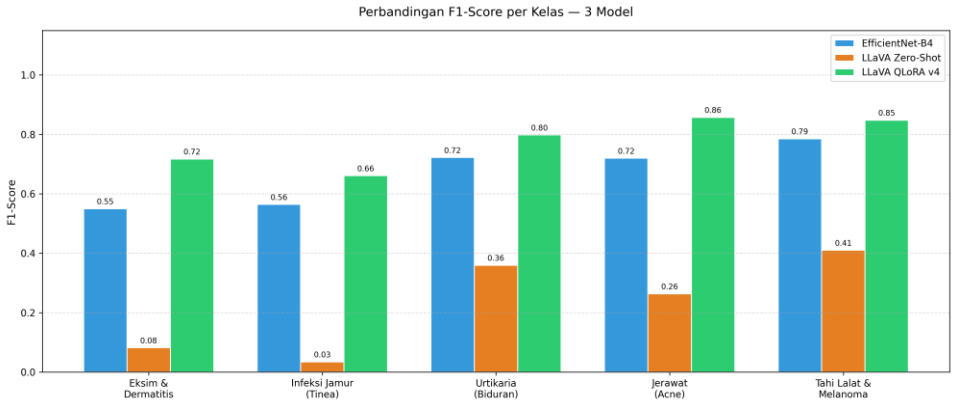
Remik: Riset dan E-Jurnal Manajemen Informatika Komputer
Volume 10, Number 3, Agustus 2026
http://doi.org/10.33395/remik.v10i1.16539
e-ISSN : 2541-1330
p-ISSN : 2541-1332

ini konsisten dengan fenomena halusinasi kelas, di mana model cenderung merespons dengan konsep yang paling dominan dalam data *pre-training* globalnya alih-alih mempelajari distribusi kelas pada tugas spesifik yang diberikan.



Gambar 4 *Confusion Matrix* Ketiga Model

Sebaliknya, LLaVA QLoRA menampilkan *confusion matrix* jauh lebih rapi: *False Positive* tertinggi hanya 48 sampel (Infeksi Jamur), *False Negative* tertinggi 34 sampel (Eksim). Jerawat (84,5%) dan *Melanoma* (83,6%) berperforma terbaik; Eksim paling sulit (69,1%), sebagian besar salah ke Infeksi Jamur dapat dimaklumi karena keduanya menampilkan lesi kemerahan bersisik yang morfologis serupa. EfficientNet-B4 unggul pada *Urtikaria* (82,7%) dan Jerawat (78,2%), namun kesulitan pada Eksim (*recall* 50%, *False Negative* 55 sampel). Analisis F1-score (Tabel 6, Gambar 5) memperlihatkan LLaVA QLoRA mendominasi kelima kelas, sementara Infeksi Jamur konsisten paling sulit bagi semua model.



Gambar 5 Perbandingan F1-Score

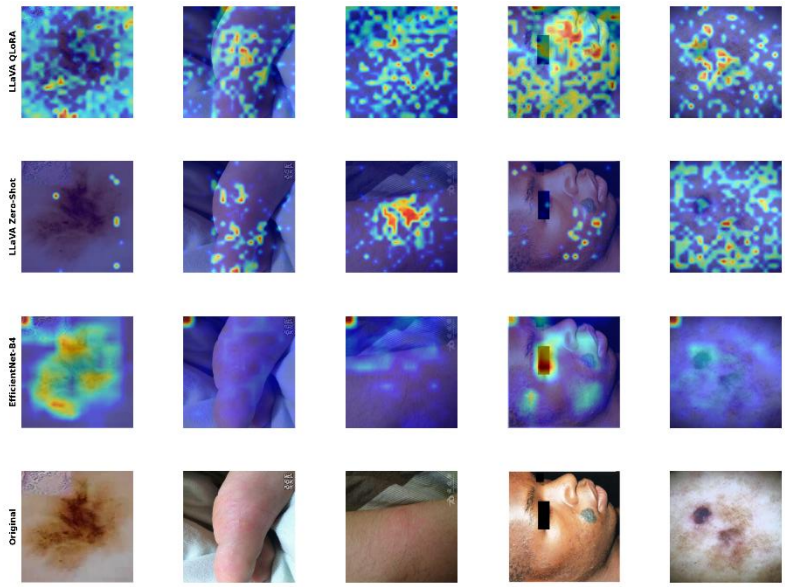
Tabel 6 F1-Score per Kelas Penyakit

Kelas Penyakit	EfficientNet-B4	LLaVA *Zero-Shot*	LLaVA QLoRA
Eksim & Dermatitis	55%	8%	72%
Infeksi Jamur (*Tinea*)	56%	3%	66%
***Urtikaria* (Biduran)**	72%	36%	80%
Jerawat (*Acne*)	72%	26%	86%
Tahi Lalat & *Melanoma*	79%	41%	85%
Rata-rata *Macro*	67%	23%	78%

Evaluasi Grad-CAM mengungkapkan perbedaan karakter yang mendasar antar arsitektur. EfficientNet-B4 menghasilkan *heatmap* berbentuk area kontinu yang cenderung mengarah ke

Remik: Riset dan E-Jurnal Manajemen Informatika Komputer
Volume 10, Number 3, Agustus 2026
http://doi.org/10.33395/remik.v10i1.16539
e-ISSN : 2541-1330
p-ISSN : 2541-1332

region lesi, konsisten dengan sifat CNN yang membangun representasi spasial secara hierarkis dari lapisan ke lapisan. LLaVA *zero-shot* menghasilkan *heatmap* berpola *grid* 14×14 *pixel* yang tersebar acak tanpa fokus yang jelas pada area lesi. LLaVA QLoRA juga berbasis ViT dengan pola *grid* yang masih terlihat, namun area yang menyala tampak lebih mengarah ke bagian lesi dibandingkan LLaVA *zero-shot*, sebagaimana terlihat pada Gambar 6.



Gambar 6 Visualisasi Grad-CAM

Secara kuantitatif, gambaran yang muncul lebih beragam dari dugaan awal. EfficientNet-B4 mencatat *Average Increase* tertinggi (23,33%) di antara ketiga model, mengindikasikan bahwa proporsi gambar yang kepercayaannya justru meningkat ketika hanya area *heatmap* yang dipertahankan paling banyak terjadi pada model ini region yang disorot merupakan representasi paling diskriminatif dari citra tersebut. LLaVA QLoRA mencatat kombinasi *Average Drop* dan *Average Increase* yang paling seimbang (25,56% dan 16,67%), tanpa menjadi yang tertinggi pada metrik manapun secara individual. LLaVA *zero-shot* mencatat *Average Drop* terendah secara numerik (10,75%), namun angka ini perlu dibaca secara hati-hati: karena *confidence* model pada citra asli memang sudah rendah dan tidak stabil (akurasi klasifikasi hanya 29,82%, dengan prediksi yang condong berat ke kelas Tahi Lalat & *Melanoma*), *Average Drop* yang kecil kemungkinan besar hanya mencerminkan bahwa memang tidak banyak keyakinan yang bisa turun sejak awal, bukan indikasi *heatmap*-nya menangkap area yang relevan. Dugaan ini didukung oleh *Average Increase* LLaVA *zero-shot* yang jatuh ke 0,00%, sebagaimana ditampilkan pada Tabel 7 dan Gambar 7.

Tabel 7 Metrik Kuantitatif Grad-CAM: *Average Drop* dan *Average Increase*

Model	*Average Drop*	*Average Increase*	Interpretasi
EfficientNet-B4	30,81%	23,33%	*Average Drop* tertinggi
LLaVA *Zero-Shot*	10,75%	0,00%	*Average Increase* terendah
LLaVA QLoRA	25,56%	16,67%	Nilai *Average Drop* dan *Average Increase* berada di antara kedua model

Konsistensi visual diukur melalui SSIM intra-kelas, di mana EfficientNet-B4 unggul di hampir semua kelas dengan rata-rata 0,533 jauh di atas LLaVA QLoRA (0,261) dan LLaVA *zero-shot*

Remik: Riset dan E-Jurnal Manajemen Informatika Komputer
Volume 10, Number 3, Agustus 2026
http://doi.org/10.33395/remik.v10i1.16539
e-ISSN : 2541-1330
p-ISSN : 2541-1332

(0,258). SSIM tinggi mengindikasikan model secara konsisten memperhatikan region morfologis yang sama untuk setiap jenis penyakit. SSIM yang relatif rendah pada kedua varian LLaVA tidak serta-merta berarti *heatmap* yang buruk, melainkan dapat mengindikasikan bahwa model VLM ini menggunakan strategi visual yang lebih bervariasi meski akurasi klasifikasinya tetap kompetitif. Detail per kelas disajikan pada Tabel 8.

Tabel 8 SSIM Intra-Kelas

Kelas Penyakit	EfficientNet-B4	LLaVA *Zero-Shot*	LLaVA QLoRA
Eksim & Dermatitis	0,540	0,325	0,162
Infeksi Jamur (*Tinea*)	0,545	0,260	0,174
***Urtikaria* (Biduran)**	0,541	0,237	0,140
Jerawat (*Acne*)	0,515	0,306	0,354
Tahi Lalat & *Melanoma*	0,524	0,162	0,476
Rata-rata	0,533	0,258	0,261

Dari sisi efisiensi komputasi, EfficientNet-B4 memiliki keunggulan yang besar: latensi inferensi hanya 45 ms per gambar (sekitar 71 kali lebih cepat dari LLaVA), kebutuhan VRAM inferensi hanya 2 GB, dan waktu pelatihan jauh lebih singkat (204 menit versus 1.547 menit untuk QLoRA). Hal ini menjadikan EfficientNet-B4 sebagai kandidat yang lebih realistis untuk aplikasi *real-time* di fasilitas kesehatan primer seperti puskesmas. LLaVA QLoRA membutuhkan waktu pelatihan paling lama namun berhasil menekan kebutuhan VRAM inferensi dari 14 GB (versi *full*) menjadi 10 GB berkat *quantization* 4-bit; profil lengkap disajikan pada Tabel 9.

Tabel 9 *Perbandingan Profil Komputasi*

Metrik Komputasi	EfficientNet-B4	LLaVA *Zero-Shot*	LLaVA QLoRA
Waktu Pelatihan	204 menit	—	1.547 menit (25,8 jam)
Latensi Inferensi	45 ms	3.200 ms	3.500 ms
VRAM Inferensi	2 GB	14 GB	10 GB (4-bit)
VRAM *Training*	11 GB	—	11 GB (QLoRA)
Parameter Total	19 juta	7,06 miliar	7,06 miliar
Parameter *Trainable*	19 juta	—	135 juta (LoRA)

Temuan ini sejalan dengan kesimpulan Zhou et al. (2024) bahwa VLM domain-spesifik mampu melampaui CNN dari sisi akurasi, namun keunggulan tersebut tidak serta-merta diikuti keunggulan *explainability*. *Trade-off* yang muncul berdimensi tiga: LLaVA QLoRA unggul akurasi, EfficientNet-B4 unggul konsistensi lokalisasi visual (*Average Increase*, SSIM) dan efisiensi komputasi (latensi ~71× lebih cepat, VRAM 2 GB berbanding 10 GB), sementara *LLaVA zero-shot* tidak unggul pada dimensi manapun; pemilihan arsitektur untuk implementasi nyata dengan demikian bergantung pada kebutuhan dan infrastruktur fasilitas kesehatan yang dituju, bukan satu model "terbaik". Perlu ditekankan bahwa *Average Drop*, *Average Increase*, dan SSIM mengukur konsistensi statistik *heatmap* terhadap prediksi model, bukan relevansi klinisnya secara langsung; validasi oleh dokter spesialis kulit terhadap kesesuaian *heatmap* dengan penilaian klinis tetap diperlukan sebelum kualitas *explainability* ini dapat diklaim bermakna secara medis.

Risiko bias *dataset* juga perlu diperhitungkan: kelima kelas penyakit diambil dari *subset* Derm1M yang, meski mencakup kondisi kulit tropis, tetap merupakan data internasional dengan

e-ISSN : 2541-1330
p-ISSN : 2541-1332

karakteristik yang mungkin berbeda dari kondisi lapangan di Indonesia, sehingga hasil ini belum tentu bertransfer langsung ke data klinis lokal. Kegagalan LLaVA *zero-shot* (29,82%) menggarisbawahi pentingnya *fine-tuning* domain-spesifik, selaras dengan argumen Yan et al. (2025) bahwa kondisi kulit yang tidak terwakili dalam *pre-training* data global menghasilkan performa *zero-shot* jauh di bawah ekspektasi. Sebaliknya, keberhasilan melatih model 7 miliar parameter pada GPU kelas konsumen melalui QLoRA membuka peluang bagi peneliti dengan sumber daya terbatas untuk turut mengeksplorasi adaptasi VLM pada domain medis lokal.

Sistem ini pada tahap ini paling tepat diposisikan sebagai alat bantu skrining awal bagi tenaga medis, bukan pengganti diagnosis klinis oleh dokter spesialis sehingga penerapan langsung pada praktik klinis memerlukan validasi lebih lanjut pada data pasien Indonesia.

V. KESIMPULAN

Pada konfigurasi dan *subset* Derm1M yang digunakan, LLaVA-7B QLoRA mencapai akurasi klasifikasi tertinggi (77,45%; F1 makro 77,62%) dibanding EfficientNet-B4 (67,45%) dan LLaVA *zero-shot* (29,82%) uji McNemar mengonfirmasi ketiga perbedaan signifikan secara statistik ($p < 0{,}001$). Namun dari sisi *explainability*, temuan lebih beragam: EfficientNet-B4 justru mencatat *Average Increase* (23,33%) dan SSIM (0,533) tertinggi, sedangkan LLaVA QLoRA tidak unggul satu metrik XAI pun meski paling seimbang (*Average Drop* 25,56%; *Average Increase* 16,67%) – akurasi tertinggi tidak otomatis berarti penjelasan visual paling konsisten. EfficientNet-B4 juga unggul komputasi (latensi ~71× lebih cepat, VRAM jauh lebih rendah), menjadikannya pilihan realistis untuk fasilitas *berinfrastruktur* terbatas. Keterbatasan penelitian: *dataset* adalah *subset* internasional Derm1M, bukan data klinis Indonesia; label mengikuti anotasi asli tanpa verifikasi dokter; metrik *explainability* belum diuji statistik formal; dan kelas Tahi Lalat & *Melanoma* masih digabung meski implikasi klinisnya berbeda. Hasil ini adalah temuan awal pada konfigurasi dan *dataset* ini, dan sistem yang dihasilkan belum dapat menggantikan diagnosis klinis tanpa validasi lebih lanjut oleh dokter spesialis pada data pasien Indonesia.

Penelitian selanjutnya disarankan mengumpulkan *dataset* kulit lokal Indonesia, melibatkan dokter spesialis dalam validasi klinis dan penilaian relevansi Grad-CAM, memisahkan kelas tahi lalat dan *melanoma*, melengkapi metrik *explainability* dengan uji statistik formal, memperluas cakupan kelas penyakit, serta mengeksplorasi arsitektur VLM yang lebih ringan agar akurasi VLM dapat dipertemukan dengan efisiensi komputasi CNN dalam satu sistem siap-validasi untuk lapangan.

VI. REFERENSI

Alfadli, R., & Khairunisa, S. (2024). Prevalensi Penyakit Kulit Infeksi dan Non-infeksi di Poliklinik Kulit dan Kelamin RSUD Jagakarsa Periode Februari 2023 - Januari 2024. *Jurnal Kedokteran Meditek*, *30*(3), 151–156. https://doi.org/10.36452/jkdoktmeditek.v30i3.3254

Chanda, T., Haggenmueller, S., Bucher, T. C., Holland-Letz, T., Kittler, H., Tschandl, P., Heppt, M. V., Berking, C., Utikal, J. S., Schilling, B., Buerger, C., Navarrete-Dechent, C., Goebeler, M., Kather, J. N., Schneider, C. V., Durani, B., Durani, H., Jansen, M., Wacker, J., Wacker, J., & Brinker, T. J. (2025). Dermatologist-like explainable AI enhances melanoma diagnosis accuracy: eye-tracking study. *Nature Communications*, 16(1), 4739. https://doi.org/10.1038/s41467-025-59532-5

Daneshjou, R., Yuksekgonul, M., Cai, Z. R., Novoa, R., & Zou, J. (2022). SkinCon: A skin disease dataset densely annotated by domain experts for fine-grained model debugging and analysis. *Advances in Neural Information Processing Systems*, *35*(NeurIPS), 1–13.

Dettmers, T., Pagnoni, A., Holtzman, A., & Zettlemoyer, L. (2023). QLORA: Efficient Finetuning of Quantized LLMs. *Advances in Neural Information Processing Systems*, *36*.

Esteva, A., Kuprel, B., Novoa, R. A., Ko, J., Swetter, S. M., Blau, H. M., & Thrun, S. (2017). Dermatologist-level classification of skin cancer with deep neural networks. *Nature*, *542*(7639), 115–118. https://doi.org/10.1038/nature21056

Haenssle, H. A., Fink, C., Schneiderbauer, R., Toberer, F., Buhl, T., Blum, A., … Zalaudek, I. (2018). Man against Machine: Diagnostic performance of a deep learning convolutional

e-ISSN : 2541-1330
p-ISSN : 2541-1332

neural network for dermoscopic melanoma recognition in comparison to 58 dermatologists. *Annals of Oncology*, *29*(8), 1836–1842. https://doi.org/10.1093/annonc/mdy166

Kementrian Kesehatan. (2023). *Profil Kesehatan Indonesia 2023*.

Kleinedlerová, I., & Kleinedler, P. (2022). Experimental Research. *SpringerBriefs in Applied Sciences and Technology*, *16*(December), 25–46. https://doi.org/10.1007/978-3-030-92130-9_4

Li, D., Fan, S., Zhao, H., Song, J., Li, W., & Xu, X. (2025). Global, regional and national burden of skin and subcutaneous diseases: a systematic analysis of the Global Burden of Disease Study 2021. *International Health*, *0*, 1–14. https://doi.org/10.1093/inthealth/ihaf070

Lidjaja, L. N. (2022). Karakteristik Penyakit Infeksi Kulit di Poliklinik Klinik Pratama Pantisiwi Jember, Januari 2018 – Desember 2020. *Cermin Dunia Kedokteran*, *49*(8), 423. https://doi.org/10.55175/cdk.v49i8.1984

Liu, H., Li, C., Wu, Q., & Lee, Y. J. (2023). Visual Instruction Tuning. *Advances in Neural Information Processing Systems*, 36.

Perhimpunan Dokter Spesialis Kulit dan Kelamin Indonesia (Perdoski). (2023). Rasio dokter spesialis kulit dan kelamin terhadap jumlah penduduk Indonesia. Dikutip dalam Media Indonesia, 25 Agustus 2023, https://mediaindonesia.com/humaniora/607880/penyebaran-dokter-kulit-di-daerah-belum-merata

Radford, A., Kim, J. W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., … Sutskever, I. (2021). Learning Transferable Visual Models From Natural Language Supervision. *Proceedings of Machine Learning Research*, *139*, 8748–8763.

Selvaraju, R. R., Cogswell, M., Das, A., Vedantam, R., Parikh, D., & Batra, D. (2017). Grad-CAM: Visual Explanations from Deep Networks via Gradient-Based Localization. *Proceedings of the IEEE International Conference on Computer Vision*, *2017-Octob*, 618–626. https://doi.org/10.1109/ICCV.2017.74

Tan, M., & Le, Q. V. (2019). *EfficientNet : Rethinking Model Scaling for Convolutional Neural Networks*.

Yan, S., Hu, M., Jiang, Y., & Li, X. (2025). *Derm1M: A Million-scale Vision-Language Dataset Aligned with Clinical Ontology Knowledge for Dermatology*.

Zhang, Y., Jiang, H., Miura, Y., Manning, C. D., & Langlotz, C. P. (2022). Contrastive Learning of Medical Visual Representations from Paired Images and Text. *Proceedings of Machine Learning Research*, *182*, 2–25.

Zhou, J., He, X., Sun, L., Xu, J., Chen, X., Chu, Y., … Gao, X. (2024). Pre-trained multimodal large language model enhances dermatological diagnosis using SkinGPT-4. *Nature Communications*. https://doi.org/10.1038/s41467-024-50043-3