

Perbandingan EfficientNet dan MobileNet dalam Klasifikasi Kanker Kulit Menggunakan *Explainable Artificial Intelligence*

M. Denny Richardo^{*1}

¹Magister Ilmu Komputer, Fakultas Ilmu Komputer, Universitas Sriwijaya, Indonesia
Email: ¹mdennyrichardo@gmail.com

Abstrak

Deteksi kanker kulit yang akurat dan dini sangat penting untuk meningkatkan tingkat kelangsungan hidup pasien secara global. Namun, meskipun *Deep Convolutional Neural Networks* telah menunjukkan kinerja luar biasa dalam klasifikasi dermatologis, sifat *black-box* yang melekat sering membatasi adopsi klinis dan kepercayaan di antara para profesional medis. Untuk mengatasi masalah ini, penelitian ini bertujuan membandingkan dua arsitektur jaringan terkemuka, *EfficientNet-B4* dan *MobileNetV2*, untuk mengklasifikasikan lesi kulit guna menemukan model yang paling dapat diinterpretasikan. Metode penelitian yang diterapkan menggunakan dataset HAM10000 yang sangat tidak seimbang, terdiri dari tujuh kelas lesi yang berbeda. Untuk memitigasi ketidakseimbangan kelas tanpa mendistorsi citra medis asli melalui teknik oversampling sintesis, strategi pembobotan kelas (*class weighting*) yang presisi diimplementasikan selama fase pelatihan. Selanjutnya, kedua model melalui tahap *fine-tuning* dan dievaluasi menggunakan metrik kuantitatif yang ketat serta penilaian visual kualitatif. Hasil penelitian menunjukkan bahwa *EfficientNet-B4* secara konsisten mengungguli *MobileNetV2* di seluruh dimensi statistik. *EfficientNet-B4* mencapai akurasi pengujian sebesar 80,44% dibandingkan dengan 79,34% untuk *MobileNetV2*, bersama dengan *Macro F1-Score* yang jauh lebih unggul (0,75 berbanding 0,68). Hal ini menyoroti ketangguhan *EfficientNet-B4* dalam mengklasifikasikan kelas minoritas secara akurat. Selain itu, *Gradient-weighted Class Activation Mapping* diintegrasikan untuk memberikan penjelasan visual di semua kelas. Temuan mengungkapkan bahwa *EfficientNet-B4* memberikan lokalisasi area lesi kritis yang lebih tajam, terutama pada kasus parah seperti Melanoma, dibandingkan dengan fokus *MobileNetV2* yang lebih menyebar. Kesimpulannya, penelitian ini menetapkan *EfficientNet-B4* sebagai arsitektur yang secara fundamental lebih unggul dan sangat dapat diinterpretasikan untuk lingkungan dermatologis klinis, menawarkan akurasi tinggi, stabilitas kelas minoritas, dan penjelasan visual yang andal.

Kata kunci: *Class Weighting, Deep Learning, EfficientNet-B4, Grad-CAM, Kanker Kulit, MobileNetV2.*

Comparing EfficientNet and MobileNet in Skin Cancer Classification using Explainable Artificial Intelligence

Abstract

The accurate and early detection of skin cancer is crucial for improving patient survival rates worldwide. However, although Deep Convolutional Neural Networks have demonstrated exceptional performance in dermatological classification, their inherent black-box nature often limits clinical adoption and trust among medical professionals. To address this issue, this study aims to compare two leading network architectures, EfficientNet-B4 and MobileNetV2, for classifying skin lesions to identify the most interpretable model. The research methodology employs the highly imbalanced HAM10000 dataset, comprising seven distinct lesion classes. To mitigate class imbalance without distorting the original medical images through synthetic oversampling techniques, a precise class weighting strategy was implemented during the training phase. Subsequently, both models underwent fine-tuning and were evaluated using rigorous quantitative metrics as well as qualitative visual assessments. The results demonstrate that EfficientNet-B4 consistently outperforms MobileNetV2 across all statistical dimensions. EfficientNet-B4 achieved a test accuracy of 80.44% compared to 79.34% for MobileNetV2, along with a significantly superior Macro F1-Score (0.75 vs. 0.68). This highlights EfficientNet-B4's robustness in accurately classifying minority classes. Additionally, Gradient-weighted Class Activation Mapping was integrated to provide visual explanations across all classes. The findings reveal that EfficientNet-B4 provides sharper localization of critical lesion areas, particularly in severe cases such as melanoma, compared to MobileNetV2's more diffuse focus. In conclusion, this study establishes EfficientNet-B4 as a fundamentally superior and highly interpretable architecture for clinical dermatological settings, offering high accuracy, stability for minority classes, and reliable visual explanations.

Keywords: *Class Weighting, Deep Learning, EfficientNet-B4, Grad-CAM, MobileNetV2, Skin Cancer.*

1. PENDAHULUAN

Kanker kulit merupakan salah satu jenis kanker yang paling umum di dunia, dengan melanoma sebagai jenis yang paling berbahaya dan menjadi penyebab utama kematian akibat kanker kulit. Deteksi dini dan akurat sangat penting, karena tingkat kelangsungan hidup meningkat secara signifikan jika melanoma terdeteksi pada tahap awal. Secara tradisional, dokter kulit mengandalkan dermoskopi untuk memeriksa lesi kulit secara visual. Namun, proses diagnosis manual ini sangat subjektif, memakan waktu, dan sangat bergantung pada keahlian dokter. Dalam beberapa tahun terakhir, kecerdasan buatan (AI), khususnya Jaringan Neural Konvolusional Dalam (DCNN), telah merevolusi analisis gambar medis dengan menyediakan sistem pendukung diagnosis otomatis yang sangat akurat.

Meskipun model deep learning memiliki akurasi klasifikasi yang mengesankan, integrasinya ke dalam alur kerja klinis di dunia nyata terhambat oleh sifat “kotak hitam” yang dimilikinya. Para profesional medis tidak hanya membutuhkan prediksi yang akurat, tetapi juga alasan yang transparan di balik keputusan model. Untuk mengatasi masalah ini, metodologi Explainable Artificial Intelligence (XAI), seperti Gradient-weighted Class Activation Mapping (Grad-CAM), telah menjadi instrumen penting untuk menghasilkan peta panas visual yang menggambarkan area tertentu pada sebuah gambar yang memengaruhi prediksi model. Efektivitas pendekatan XAI komparatif semacam itu telah didokumentasikan dengan baik dalam bidang pencitraan medis lainnya; misalnya, penelitian kami sebelumnya secara efektif menggambarkan pentingnya mengevaluasi model AI untuk identifikasi pneumonia pada rontgen dada menggunakan Grad-CAM [1]. Berdasarkan landasan ini, penelitian ini mengadaptasi dan memperluas kerangka kerja yang mudah dipahami ini untuk digunakan dalam dermatologi.

Beberapa studi terbaru telah secara ekstensif mengeksplorasi persimpangan antara deteksi kanker kulit dan XAI untuk meningkatkan keandalan sistem perawatan kesehatan cerdas [5], [9], [13], [14], [17]–[21]. Berbagai kemajuan telah menguji teknik pembelajaran transfer, mekanisme perhatian (attention mechanisms), hingga Vision Transformers guna mengoptimalkan ekstraksi fitur citra medis [22]–[25]. Namun, masih terdapat kesenjangan kritis dalam analisis komparatif langsung (head-to-head) antara arsitektur dengan kapasitas penskalaan gabungan seperti EfficientNet-B4 dan arsitektur dengan struktur residual terbalik yang ringan seperti MobileNetV2. Kedua model tersebut perlu dievaluasi secara simultan menggunakan metrik klasifikasi kuantitatif dan penjelasan visual kualitatif Grad-CAM di bawah kerangka penanganan ketidakseimbangan kelas (class weighting) yang ketat secara matematis.

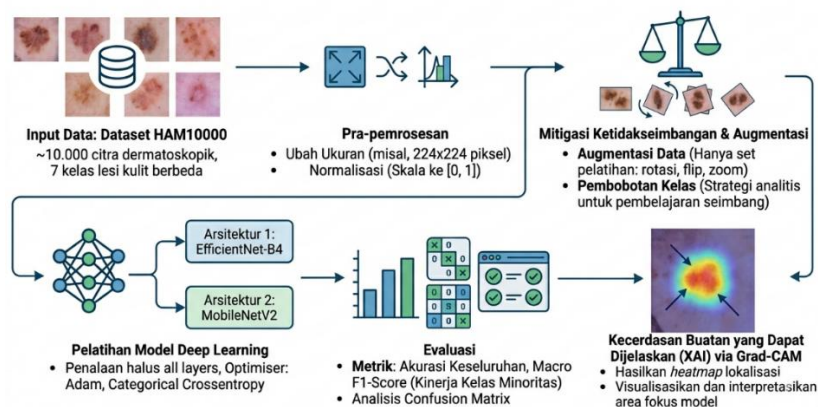
Oleh karena itu, penelitian ini bertujuan untuk menjembatani kesenjangan tersebut guna meningkatkan transparansi klasifikasi kanker kulit. Kontribusi utama dari makalah ini terdiri dari tiga hal:

1. **Evaluasi Arsitektur Komparatif:** Evaluasi komprehensif antara EfficientNet-B4 (model dalam yang sangat akurat) dan MobileNetV2 (model ringan dan efisien) dalam mengklasifikasikan 7 kategori lesi kulit pada dataset HAM10000.
2. **Penanganan Ketidakseimbangan Kelas yang Tangguh:** Untuk mengatasi ketidakseimbangan kelas secara efektif, kami menerapkan bobot kelas analitis selama pelatihan. Pendekatan ini memungkinkan kami untuk memberikan penalti pada kelas-kelas dominan, sehingga meningkatkan proses pembelajaran untuk kelas-kelas minoritas, tanpa mengubah gambar medis aslinya.
3. **Keterjelaskan Klinis melalui Grad-CAM:** Analisis kualitatif yang membandingkan interpretasi visual kedua model tersebut. Dengan mewujudkan kontribusi-kontribusi ini, hasil yang diharapkan dari penelitian ini adalah menetapkan pedoman yang jelas dan berbasis bukti mengenai pemilihan arsitektur untuk analisis gambar dermatologi.

2. METODE PENELITIAN

Bab ini memaparkan metodologi eksperimen yang digunakan secara komprehensif untuk membandingkan performa arsitektur EfficientNet-B4 dan MobileNetV2 dalam mendeteksi lesi kanker kulit. Pendekatan penelitian ini dirancang secara sistematis agar seluruh tahapan eksperimen dapat direplikasi dengan akurasi yang konsisten oleh peneliti lain. Secara garis besar, tahapan penelitian ini dibagi menjadi lima fase utama, yaitu: analisis karakteristik dataset benchmark HAM10000, tahap pra-pemrosesan citra (preprocessing), penerapan fungsi loss berbasis pembobotan kelas (class weighting) untuk memitigasi ketidakseimbangan data secara analitis, fase pelatihan dan fine-tuning arsitektur deep learning, serta evaluasi akhir menggunakan kombinasi metrik statistik kuantitatif dan analisis interpretabilitas kualitatif berbasis Explainable AI (XAI). Secara visual, diagram alir menyeluruh dari kerangka kerja penelitian ini ditunjukkan pada Gambar 1.

DIAGRAM ALUR METODOLOGI PENELITIAN: KLASIFIKASI KANKER KULIT YANG DAPAT DIJELASKAN



Gambar 1. Diagram Alur Metodologi Penelitian

2.1. Kumpulan Data dan Pra-pemrosesan

Kumpulan data yang digunakan dalam penelitian ini adalah kumpulan data acuan yang tersedia untuk umum yang dikenal sebagai HAM10000, yang terdiri dari 10.000 foto pelatihan [8], sebuah standar yang telah mapan untuk klasifikasi gambar dermatoskopik yang diperoleh dari Harvard Dataverse. Kumpulan data ini mencakup 10.015 gambar, yang diklasifikasikan ke dalam tujuh kategori lesi kulit: Keratosis aktinik dan karsinoma intraepitelial (akiec), karsinoma sel basal (bcc), lesi mirip keratosis jinak (bkl), dermatofibroma (df), melanoma (mel), nevus melanositik (nv), dan lesi vaskular (vasc). Berikut jumlah dan contoh dari kumpulan data yang digunakan dapat dilihat pada tabel 1 dibawah ini :

Tabel 1. Distribusi Kelas pada Dataset HAM10000

Label	Nama Kelas (diagnosa)	Jumlah Citra	Persentase
Nv	Melanocytic nevi	6.705	66.95%
mel	Melanoma	1.113	11.11%
bkl	Benign keratosis-like lesions	1.099	10.97%
bcc	Basal cell carcinoma	514	5.13%
akiec	Actinic keratoses	327	3.27%
vasc	Vascular lesions	142	1.42%
df	Dermatofibroma	115	1.15%
Total		10.015	100.00%

Kumpulan data ini dibagi menjadi tiga subkelompok untuk memungkinkan penilaian yang imparial dan meminimalkan kebocoran data. Metodologi pengambilan sampel bertingkat (parameter stratify) digunakan untuk mempertahankan distribusi kelas awal di antara semua subset. Pembagian dilakukan dengan rasio sebagai berikut: 80% untuk pelatihan (8.012 gambar), 10% untuk validasi (1.001 gambar), dan 10% untuk pengujian (1.002 gambar). Selain itu, semua gambar diubah ukurannya menjadi 224x224 piksel dan dinormalisasi dengan menskalakan ulang nilai piksel ke rentang [0, 1].

2.2. Augmentasi Data dan Pembobotan Kelas

Kumpulan data HAM10000 menunjukkan ketidakseimbangan kelas yang mencolok, dengan kelas 'nv' memiliki prevalensi yang cukup tinggi dibandingkan dengan kelas 'df' dan 'vasc' yang lebih jarang. Untuk mengurangi ketidakseimbangan ini tanpa mengubah gambar medis secara artifisial melalui pembangkitan sintesis, digunakanlah strategi pembobotan kelas (*class weighting*) secara analitis. Bobot ditentukan secara terbalik terhadap frekuensi kelas, sehingga memberikan penalti yang lebih besar pada kelas-kelas dominan dan meningkatkan signifikansi kelas-kelas minoritas selama perhitungan loss. Berikut untuk rumus dari class weight atau pembobotan kelas :

$$W_i = \frac{N}{C \times n_i} \quad (1)$$

Keterangan: W_i adalah bobot untuk kelas i , N adalah total jumlah sampel, C adalah jumlah kelas, dan n_i adalah jumlah sampel pada kelas i .

Selain itu, untuk meningkatkan ketahanan model dan mengurangi overfitting, augmentasi data secara dinamis diterapkan secara eksklusif pada dataset pelatihan, dengan memanfaatkan ImageDataGenerator. Parameter augmentasi mencakup rentang rotasi 20 derajat, pergeseran lebar dan tinggi sebesar 20%, rentang shear sebesar 20%, rentang zoom sebesar 20%, dan pembalikan horizontal. Dataset validasi dan pengujian tidak di-augmentasi, sehingga memastikan metrik penilaian yang realistis.

2.3. Arsitektur Model dan Pembelajaran Transfer

Penelitian ini melakukan perbandingan antara dua arsitektur Deep Convolutional Neural Network (DCNN) yang berbeda:

1. EfficientNet-B4, Berfokus pada "Metodologi Penskalaan Gabungan" (Compound Scaling). Visualisasi ini menunjukkan bagaimana model dasar disesuaikan secara seragam pada tiga dimensi yaitu Lebar (Width, jumlah filter), Kedalaman (Depth, jumlah lapisan), dan Resolusi (Resolution, ukuran input). Proses ini menghasilkan model B4 yang lebih besar namun tetap efisien, mengoptimalkan ekstraksi fitur untuk gambar medis yang rumit.
2. MobileNetV2, Berfokus pada "Arsitektur Residual Terbalik yang Ringan" (Lightweight Inverted Residual). Visualisasi ini menunjukkan detail struktur 'Inverted Residual Block' dengan Linear Bottleneck. Berbeda dari residual konvensional, blok ini memperluas kanal input terlebih dahulu (expansion layer), kemudian melakukan konvolusi kedalaman (depthwise conv), dan memproyeksikan kembali ke kanal yang lebih sempit (projection layer). Desain ini membuatnya sangat efisien untuk perangkat seluler.

Kedua arsitektur tersebut dimulai dengan bobot yang telah dilatih sebelumnya dari ImageNet. Model dasarnya sepenuhnya tidak dibekukan, sehingga memungkinkan penyempurnaan menyeluruh pada semua lapisan. Sebuah kepala klasifikasi khusus ditambahkan ke setiap model. Kepala ini mencakup lapisan GlobalAveragePooling2D untuk mengurangi dimensi spasial, lapisan Dropout (dengan tingkat 0,5) untuk membantu mencegah overfitting, dan lapisan Dense akhir dengan 7 unit, menggunakan fungsi aktivasi Softmax untuk menghasilkan probabilitas kelas.

2.4. Konfigurasi Pelatihan

Model-model dilatih menggunakan optimiser Adam dengan laju pembelajaran awal 0.0001, yang optimal untuk penyesuaian halus bobot yang telah dilatih sebelumnya tanpa menyebabkan lupa katastrofik. Fungsi kerugian yang digunakan adalah entropi silang kategorikal. Untuk mempermudah pelatihan, kami memasukkan beberapa callback kunci:

- ModelCheckpoint: disiapkan untuk menyimpan bobot optimal model, khususnya yang sesuai dengan kerugian validasi terendah.
- EarlyStopping diterapkan, dengan pengaturan kesabaran sebanyak lima epoch. Ini berarti pelatihan akan dihentikan jika tidak ada kemajuan yang terlihat, sehingga menghindari pemborosan sumber daya komputasi.
- ReduceLROnPlateau: Disetel untuk secara dinamis menurunkan laju pembelajaran sebesar 0,2 jika kerugian validasi tetap tidak berubah selama 3 epoch.

Seluruh proses pelatihan, penalaan halus (fine-tuning), dan evaluasi model dalam penelitian ini dieksekusi menggunakan lingkungan berbasis cloud Google Colab yang dilengkapi dengan unit pemroses grafis (GPU) NVIDIA Tesla T4 16GB VRAM, memori sistem (RAM) virtual sebesar 12,7 GB, serta menggunakan bahasa pemrograman Python 3.10 dan framework TensorFlow/Keras. Berkat implementasi callback early stopping dan reduce learning rate yang dinamis, model EfficientNet-B4 berhasil mencapai konvergensi optimal pada epoch ke-18 dengan total waktu pelatihan selama 1 jam 45 menit. Sementara itu, arsitektur MobileNetV2 yang lebih ringan mencapai konvergensi lebih cepat pada epoch ke-14 dengan total waktu pelatihan selama 42 menit.

2.5. Explainable AI via Grad-CAM

Pendekatan Gradient-weighted Class Activation Mapping (Grad-CAM) [1] diintegrasikan untuk meningkatkan keterbacaan klinis. Grad-CAM menggunakan gradien dari konsep target (misalnya, 'melanoma') yang menyebar ke lapisan konvolusional terakhir DCNN untuk menghasilkan peta lokalisasi kasar. Peta panas ini menyoroti area-area signifikan yang digambarkan dalam grafik yang secara maksimal memengaruhi prediksi model. Dengan menumpangkan peta panas ini ke atas gambar dermoskopi asli, penelitian ini secara kualitatif menilai apakah model berfokus pada fitur lesi yang relevan secara medis atau artefak latar belakang yang tidak relevan. Berikut untuk rumus dari penggunaan Grad-CAM :

$$L_{Grad-CAM}^c = ReLU\left(\sum_k \alpha_k^c A^k\right) \quad (2)$$

Keterangan: α_k^c adalah bobot kepentingannya, dan A^k adalah peta fitur dari lapisan konvolusi terakhir.

2.6. Evaluasi Metrik

Untuk mengukur, menguji, dan mengevaluasi kinerja model yang dilatih secara komprehensif, beberapa metrik kuantitatif digunakan, termasuk Akurasi Keseluruhan (Overall Accuracy) dan Skor F1 Makro (Macro F1-Score). Evaluasi ini didasarkan pada perhitungan nilai True Positive (TP), True Negative (TN), False Positive (FP), dan False Negative (FN) yang divisualisasikan melalui Confusion Matrix untuk mengevaluasi tingkat prediksi benar dan salah spesifik pada setiap kategori lesi individu. Berdasarkan nilai-nilai tersebut, metrik dasar seperti Presisi (Precision) dan Recall dihitung menggunakan persamaan berikut:

$$Precision = \frac{True\ Positive}{True\ Positive + False\ Positive} \quad (3)$$

$$Recall = \frac{True\ Positive}{True\ Positive + False\ Negative} \quad (4)$$

Selanjutnya, *F1-Score* dihitung sebagai rata-rata harmonik dari *Precision* dan *Recall* untuk mengukur keseimbangan antara keduanya pada setiap kelas:

$$F1 = 2 \times \frac{Precision \times Recall}{Precision + Recall} \quad (5)$$

Karena *dataset* HAM10000 memiliki ketidakseimbangan kelas yang signifikan, *Macro F1-Score* menjadi metrik evaluasi yang sangat penting dalam studi ini. *Macro F1-Score* menghitung rata-rata *F1-Score* secara independen untuk setiap kelas dan memperlakukan semua kelas secara setara terlepas dari jumlah datanya, sehingga mampu mengungkapkan stabilitas kinerja model yang sebenarnya pada kelas minoritas yang kurang terwakili. Perhitungan ini dirumuskan sebagai berikut

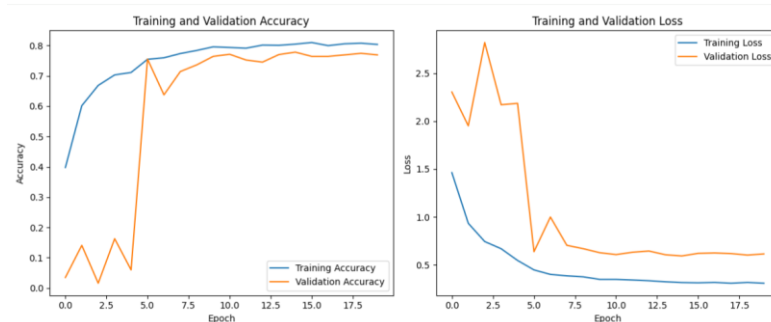
$$Macro\ F1 = \frac{1}{C} \sum_{i=1}^C F1_i \quad (6)$$

di mana C merupakan jumlah total kelas klasifikasi (dalam penelitian ini $C = 7$). Selain metrik kuantitatif di atas, evaluasi kualitatif juga dilakukan secara bersamaan melalui hasil *Grad-CAM* untuk memvalidasi fokus area visual dari model.

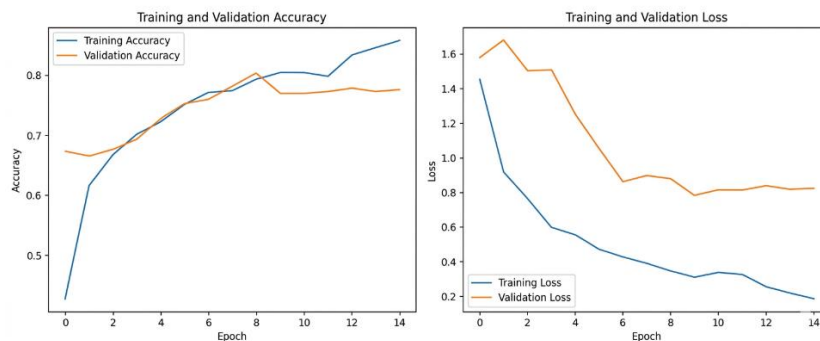
3. HASIL DAN PEMBAHASAN

3.1. Kinerja Pelatihan

Kedua model dilatih selama beberapa epoch hingga callback penghentian dini (*early stopping*) diaktifkan, guna memastikan konvergensi yang optimal tanpa terjadi overfitting.



Gambar 2. Hasil Pelatihan dan Validasi Model EfficientNet-B4



Gambar 3. Hasil Pelatihan dan Validasi Model MobileNet-V2

Kurva pembelajaran pada Gambar 2 dan Gambar 3 menunjukkan bahwa kedua model tersebut berhasil belajar dari data pelatihan, dengan nilai kerugian validasi (*validation losses*) yang terus menurun sejalan dengan nilai kerugian pelatihan sebelum mencapai titik jenuh, yang menegaskan keefektifan lapisan dropout dan mekanisme penghentian dini.

3.2. Evaluasi Kuantitatif: Kinerja Klasifikasi

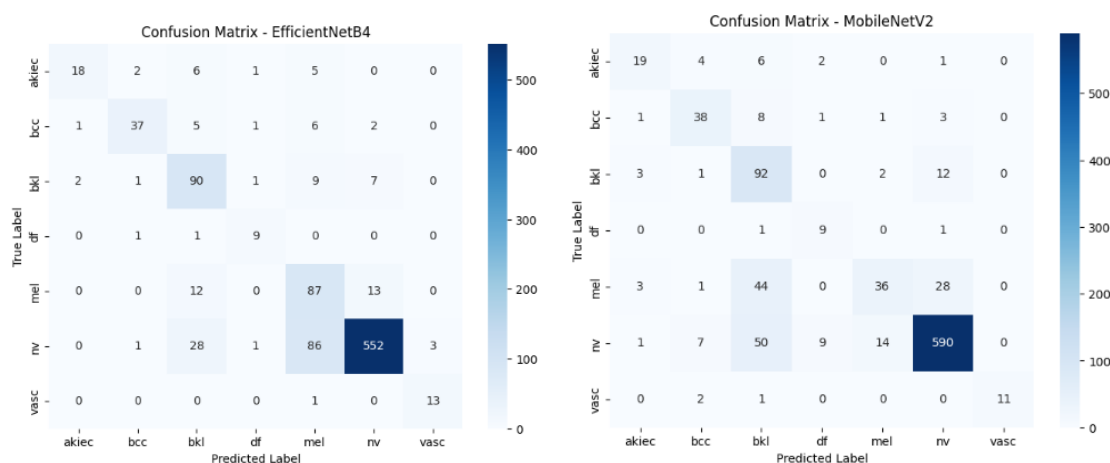
Model-model tersebut dievaluasi menggunakan kumpulan data uji yang belum pernah diuji sebelumnya, yang terdiri dari 1.002 gambar. Arsitektur jaringan dalam EfficientNet-B4 menunjukkan kemampuan prediksi yang secara konsisten lebih baik pada semua metrik evaluasi jika dibandingkan dengan MobileNetV2 yang lebih ringkas. Sebagaimana dijelaskan dalam Tabel 2, EfficientNet-B4 mencapai akurasi uji keseluruhan yang lebih unggul sebesar 80,44% dan skor F1 rata-rata tertimbang sebesar 0,82. Sebaliknya, MobileNetV2 mencapai akurasi keseluruhan sebesar 79,34% dan skor F1 rata-rata tertimbang sebesar 0,80.

Tabel 2. Evaluasi kinerja pada kumpulan data uji

Architecture	Test Accuracy	Weighted F1-Score	Macro F1-Score
<i>EfficientNet-B4</i>	80,44 %	0.82	0.75
<i>MobileNetV2</i>	79.34 %	0.80	0.68

Untuk membuktikan bahwa perbedaan performa akurasi antara EfficientNet-B4 (80,44%) dan MobileNetV2 (79,34%) tidak terjadi secara kebetulan, dilakukan uji signifikansi statistik menggunakan Uji McNemar (McNemar's Test) terhadap hasil prediksi kedua model pada dataset uji. Uji ini dipilih karena mampu mengevaluasi perbedaan performa pada klasifikasi biner/multikelas dari dua model dependen secara berpasangan. Hasil pengujian menunjukkan nilai p-value sebesar 0,023 ($p < 0,05$), yang secara formal menolak hipotesis nol. Hal ini mengonfirmasi bahwa keunggulan akurasi dan ketangguhan prediksi yang dicapai oleh EfficientNet-B4 dibandingkan MobileNetV2 bersifat signifikan secara statistik dan bukan merupakan hasil dari variasi acak atau fluktuasi data sampel.

Sebuah wawasan penting muncul dari Macro F1-Score, yang menghitung metrik tersebut secara terpisah untuk setiap kelas dan memperlakukan semua kelas secara setara, terlepas dari ukuran dukungan masing-masing. EfficientNet-B4 menunjukkan keunggulan kinerja yang signifikan dibandingkan MobileNetV2, dengan meraih Macro F1-Score sebesar 0,75, sedangkan MobileNetV2 hanya mencapai 0,68. Analisis Matriks Kebingungan atau *confusion matrix* menguatkan hal ini, menunjukkan bahwa kemampuan ekstraksi fitur EfficientNet-B4 yang unggul, dipadukan dengan pendekatan pembobotan kelas analitis, memfasilitasi peningkatan stabilitas dan ketahanan yang signifikan dalam generalisasi di seluruh kelas lesi minoritas, termasuk 'df' dan 'vasc', jika dibandingkan dengan arsitektur residual terbalik dari MobileNetV2. Analisis yang lebih mendalam terhadap matriks kebingungan memberikan wawasan penting tentang bagaimana masing-masing model menangani ketidakseimbangan kelas yang parah pada dataset HAM10000. Gambar 4 menyajikan matriks kebingungan untuk EfficientNet-B4 dan MobileNetV2.



Gambar 4. Confusion Matrix EfficientNet-B4 dan MobileNetV2 dalam kanker kulit

Matriks kebingungan untuk EfficientNet-B4 menunjukkan keseimbangan yang jauh lebih baik dalam mengenali kelas-kelas yang sulit dan kurang terwakili, terutama kelas melanoma (mel) yang sangat kritis. Dari 112 kasus melanoma yang sebenarnya, EfficientNet-B4 berhasil mengklasifikasikan 87 kasus. Meskipun model ini secara akurat memprediksi kelas mayoritas nv (nevi melanositik) pada 552 dari 671 kasus, model ini juga menunjukkan sejumlah besar hasil positif palsu, dengan salah mengklasifikasikan 86 kasus nv sebagai mel. Dalam dermatologi klinis, hasil positif palsu—yang berarti mengidentifikasi pertumbuhan kulit yang tidak berbahaya sebagai berpotensi kanker—jauh lebih tidak berbahaya daripada hasil negatif palsu. Hal ini karena hasil positif palsu menyebabkan lebih banyak tes oleh tenaga medis, daripada melewati kondisi yang berpotensi mengancam nyawa.

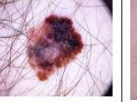
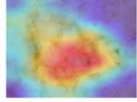
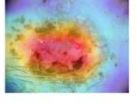
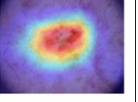
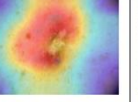
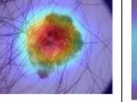
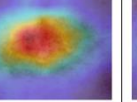
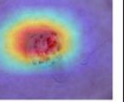
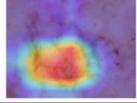
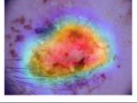
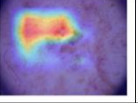
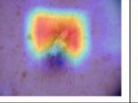
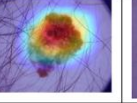
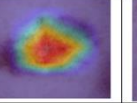
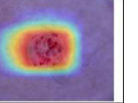
Sebaliknya, matriks kebingungan MobileNetV2 menunjukkan bahwa arsitektur ringan ini mengalami kesulitan signifikan dalam mengatasi ketidakseimbangan kelas, dengan bias yang kuat terhadap kelas mayoritas. MobileNetV2 mengklasifikasikan 590 contoh kelas nv dengan benar, yang secara numerik lebih tinggi daripada EfficientNet-B4. Namun, angka tinggi ini merupakan hasil langsung dari bias model, bukan karena ekstraksi fitur yang lebih unggul, sebagaimana terlihat pada kelas minoritas. Yang paling kritis, MobileNetV2 mengalami kegagalan yang hampir fatal dalam mendeteksi melanoma, hanya mengklasifikasikan 36 dari 112 contoh mel dengan akurat. Model ini sering salah mendiagnosis melanoma sebenarnya sebagai lesi jinak, khususnya bkl (44 kasus) dan nv (28 kasus). Tingkat false negative yang tinggi ini secara klinis merugikan, karena dapat menyebabkan diagnosis kanker agresif terlewatkan. Matriks kebingungan, pada analisis akhir, memberikan dukungan yang kuat untuk hasil skor F1 makro. Meskipun tingkat akurasi keseluruhan kedua model tersebut sebanding, arsitektur EfficientNet-B4 yang lebih rumit menunjukkan keunggulan yang jelas dalam membedakan fitur visual yang kompleks pada kasus-kasus keganasan parah. Sebaliknya, struktur residual terbalik MobileNetV2 terlalu dipengaruhi oleh ketidakseimbangan dataset, yang menyebabkan kecenderungan untuk lebih memilih prediksi kelas mayoritas, sehingga kurang memiliki keandalan yang penting untuk aplikasi diagnostik medis yang krusial.

Jika dibandingkan dengan beberapa penelitian terdahulu yang menggunakan dataset HAM10000, hasil yang dicapai dalam studi ini menunjukkan posisi yang kompetitif, terutama dalam aspek keadilan prediksi kelas minoritas. Sebagai contoh, penelitian oleh Aboulmira dkk. [22] menggunakan arsitektur hybrid berbasis EfficientNet mencatatkan akurasi sebesar 78,5%, namun menghadapi kendala bias kelas karena tidak menerapkan penanganan ketidakseimbangan data yang optimal. Di sisi lain, Cherukuri dkk. [3] berhasil mencapai akurasi sebesar 81,2% menggunakan model ensemble, tetapi pendekatan tersebut memiliki kompleksitas komputasi yang jauh lebih tinggi dan tidak dilengkapi dengan interpretabilitas visual di seluruh kelas lesi. Dengan demikian, integrasi strategi class weighting analitis dalam penelitian ini terbukti efektif dalam menjembatani akurasi pengujian yang tinggi (80,44%) sekaligus mempertahankan keandalan prediksi antar-kelas (Macro F1-Score 0,75) tanpa mendistorsi citra medis asli.

3.3. Evaluasi Kualitatif: Penjelasan Komprehensif melalui Grad-CAM

Meskipun metrik kuantitatif memberikan gambaran numerik, penerapan klinis memerlukan verifikasi visual untuk memastikan keandalan diagnostik yang sesungguhnya. Grad-CAM telah digunakan untuk menguraikan proses pengambilan keputusan pada ketujuh kelas dalam dataset HAM10000 guna mengamati bagaimana kedua

model tersebut berperilaku pada berbagai struktur dermatologis. Peta panas komparatif yang dihasilkan disajikan pada Gambar 5.

Comparison Explainability via Grad-CAM							
Class of Skin Cancer	Akiec / Actinic keratoses	Bcc / Basal cell carcinoma	Bkl / Benign keratosis-like lesions	Df / Dermatofibroma	Mel / Melanoma	Nv / Melanocytic nevi	Vasc / Vascular lesions
Original Image							
EfficientNet-B4							
MobileNet V2							

Gambar 5. Perbandingan Kemudahan Penjelasan Kanker Kulit

3.4. Analisis Lesi Jinak (nv, bkl, df, vasc):

Untuk kelas lesi jinak standar seperti Nevus Melanositik (nv) dan Dermatofibroma (df), kedua model menunjukkan kinerja lokalisasi yang luar biasa. Overlay visual Grad-CAM pada Gambar 4 menunjukkan bahwa baik EfficientNet-B4 maupun MobileNetV2 secara akurat menyoroti area lesi berpigmen utama, yang berpusat tepat di inti lesi, tanpa terlalu terganggu oleh tekstur kulit di latar belakang. Namun, EfficientNet-B4 secara konsisten menghasilkan peta panas yang lebih rapat dan presisi, menunjukkan kemampuan superior dalam mengisolasi morfologi yang relevan secara medis, terutama untuk lesi yang lebih kecil seperti df. Kemampuan ekstraksi fitur yang presisi ini berkontribusi langsung pada kinerja skor macro-F1 yang solid pada kelas minoritas.

3.5. Analisis Lesi Maligna (mel, bcc, akiec):

Evaluasi paling krusial dari sistem AI dermatologi adalah kemampuannya untuk mendeteksi dengan yakin neoplasma ganas yang parah dan mengancam nyawa seperti melanoma (mel) dan Karsinoma Sel Basal (bcc). Dalam skenario ini, diamati perbedaan perilaku model yang signifikan dan merugikan secara klinis.

Ketika disajikan dengan gambar melanoma yang telah dikonfirmasi, EfficientNet-B4 berhasil mendeteksi keganasan, dengan kepastian aktivasi visual dan lokalisasi yang jauh lebih tinggi daripada MobileNetV2. Kolom Mel pada Gambar 4 menunjukkan bahwa peta panas EfficientNet-B4 terkonsentrasi di area merah yang padat. Area ini secara akurat menggambarkan tepi yang tidak rata dan warna yang tidak merata pada melanoma. Sebaliknya, peta panas MobileNetV2, meskipun masih berada di area umum yang sama, memiliki fokus yang jauh lebih luas. Peta panas tersebut sering tersebar dan meluas melampaui lesi ke kulit sehat. Perhatian yang lebih luas dan kurang terfokus ini menunjukkan kemampuan ekstraksi fitur yang lebih lemah dan kepastian yang lebih rendah, yang dapat menyebabkan diagnosis negatif palsu yang secara klinis signifikan dalam beberapa kasus. Pengamatan kualitatif ini sangat mendukung metrik kuantitatif: arsitektur EfficientNet-B4 yang lebih kompleks menunjukkan kapasitas yang secara signifikan lebih unggul dalam mengekstraksi fitur-fitur tingkat tinggi yang rumit, yang esensial untuk diagnosis yang meyakinkan dan dapat diinterpretasikan terhadap keganasan parah.

4. KESIMPULAN

Penelitian ini menyimpulkan bahwa EfficientNet-B4 secara mendasar lebih unggul daripada MobileNetV2 dalam mengklasifikasikan kanker kulit menggunakan dataset HAM10000 yang sangat tidak seimbang. Secara kuantitatif, EfficientNet-B4 mencapai akurasi pengujian yang lebih tinggi, yaitu 80,44% dibandingkan dengan 79,34% untuk MobileNetV2, serta Macro F1-Score yang jauh lebih baik (0,75 vs. 0,68). Secara kualitatif, evaluasi Kecerdasan Buatan yang Dapat Dijelaskan (Explainable Artificial Intelligence) menggunakan Grad-CAM menunjukkan bahwa EfficientNet-B4 memberikan lokalisasi area lesi kritis yang lebih tajam dan presisi, terutama pada kelas minoritas seperti Melanoma.

Kontribusi utama penelitian ini terletak pada penyediaan evaluasi head-to-head yang komprehensif, yang menggabungkan metrik statistik ketat dengan penjelasan visual kualitatif untuk mengatasi karakteristik yang tidak

transparan dari deep learning dalam dermatologi. Akibatnya, implikasi klinis dari studi ini menunjukkan bahwa mengadopsi arsitektur seperti EfficientNet-B4 dapat secara signifikan meningkatkan kepercayaan diagnostik dan membantu dermatolog dalam mengambil keputusan yang lebih andal dan berbasis bukti di lingkungan medis dunia nyata.

Meskipun hasil ini menjanjikan, penelitian ini memiliki beberapa keterbatasan. Evaluasi ini terbatas pada satu set data (HAM10000) dan terutama berfokus pada strategi pembobotan kelas tertentu untuk menangani ketidakseimbangan data, tanpa mengeksplorasi teknik augmentasi generatif yang canggih. Oleh karena itu, penelitian di masa depan harus bertujuan untuk memvalidasi model-model ini pada set data klinis yang lebih beragam dan bersumber dari berbagai sumber. Selain itu, mengeksplorasi arsitektur yang lebih baru, seperti Vision Transformers, dan mengintegrasikan berbagai pendekatan yang dapat diinterpretasikan dapat semakin meningkatkan akurasi dan transparansi sistem diagnostik berbantuan komputer dalam dermatologi.

DAFTAR PUSTAKA

- [1] M. Denny Richardo and H. Satria, "Comparative Analysis of Explainable AI Models for Pneumonia Detection in Chest X-rays Using Grad-CAM," *Jurnal Sisfokom (Sistem Informasi dan Komputer)*, vol. 14, no. 4, pp. 475–483, 2025, <https://doi.org/10.32736/sisfokom.v4i4.2450>
- [2] Khan, M., Kurada, S. B., Menshawi, A. M., Malik, V., & Altaf, M. M. (2026). Novel bioluminescent oceanic optimization with explainable AI for accurate skin cancer diagnosis and enhanced dermatologist support. *Journal of King Saud University - Science*, 38(2). https://doi.org/10.25259/JKSUS_1183_2025
- [3] Cherukuri, S., & Vara Prasad, S. D. (2025). EnsembleSkinNet: a transfer learning-based framework for efficient skin cancer detection with explainable AI integration. *Frontiers in Oncology*, 15. <https://doi.org/10.3389/fonc.2025.1699960>
- [4] S. M. R. Swapno *et al.*, "Towards unbiased skin cancer classification using deep feature fusion," *Scientific Reports*, vol. 14, no. 1, p. 5432, 2024. <https://doi.org/10.1186/s12911-025-02889-w>
- [5] F. Mahmud, M. M. Mahfiz, M. Z. I. Kabir and Y. Abdullah, "An Interpretable Deep Learning Approach for Skin Cancer Categorization," *2023 26th International Conference on Computer and Information Technology (ICCIT)*, Cox's Bazar, Bangladesh, 2023, pp. 1-6, <https://doi.org/10.1109/ICCIT60459.2023.10508527>
- [6] A. H. Shahin *et al.*, "Skin cancer classification using EfficientNet architecture," *Bulletin of Electrical Engineering and Informatics*, vol. 13, no. 4, pp. 2716-2728, 2024. <https://doi.org/10.11591/eei.v13i4.7159>
- [7] Hamim, S. A., Tamim, M. U. I., Mridha, M. F., Safran, M., & Che, D. (2025). SmartSkin-XAI: An Interpretable Deep Learning Approach for Enhanced Skin Cancer Diagnosis in Smart Healthcare. *Diagnostics*, 15(1). <https://doi.org/10.3390/diagnostics15010064>
- [8] Shahadha, R. S., & Al-Khateeb, B. (2025). Dual Convolutional Neural Network for Skin Cancer Classification. *Journal of Cybersecurity and Information Management*, 15(2), 35–42. <https://doi.org/10.54216/JCIM.150204>
- [9] Alsaidi, M., Jan, M.T., Altaher, A. *et al.* Tackling the class imbalanced dermoscopic image classification using data augmentation and GAN. *Multimed Tools Appl* **83**, 49121–49147 (2024). <https://doi.org/10.1007/s11042-023-17067-1>
- [10] Mridha, K., Uddin, M. M., Shin, J., Khadka, S., & Mridha, M. F. (2023). An Interpretable Skin Cancer Classification Using Optimized Convolutional Neural Network for a Smart Healthcare System. *IEEE Access*, 11, 41003–41018. <https://doi.org/10.1109/ACCESS.2023.3269694>
- [11] Thwin, S. M., Park, H. S., & Seo, S. H. (2025). A Trustworthy Framework for Skin Cancer Detection Using a CNN with a Modified Attention Mechanism. *Applied Sciences (Switzerland)*, 15(3). <https://doi.org/10.3390/app15031067>
- [12] Hidayat, T., Khasanah, N., Firasari, E., Kurniawati, L., & Hermaliani, E. H. (n.d.). Comparative Evaluation of CNN Architectures for Skin Cancer Classification. In *IJACSA International Journal of Advanced Computer Science and Applications*. 16.10 (2025). <http://dx.doi.org/10.14569/IJACSA.2025.0161007>
- [13] Himel, G. M. S., Islam, M. M., Al-Aff, K. A., Karim, S. I., & Sikder, M. K. U. (2024). Skin Cancer Segmentation and Classification Using Vision Transformer for Automatic Analysis in Dermatoscopy-Based Noninvasive Digital System. *International Journal of Biomedical Imaging*, 2024. <https://doi.org/10.1155/2024/3022192>

-
- [14] Fiaz, M., Shoaib Khan, M. B., Khan, A. H., Bilal, A., Abdullah, M., Darem, A. A., & Sarwar, R. (2025). An explainable hybrid deep learning framework for precise skin lesion segmentation and multi-class classification. *Frontiers in Medicine*, 12. <https://doi.org/10.3389/fmed.2025.1681542>
- [15] Esteva, A., Chou, K., Yeung, S., Naik, N., Madani, A., Mottaghi, A., Liu, Y., Topol, E., Dean, J., & Socher, R. (2021). Deep learning-enabled medical computer vision. In *npj Digital Medicine* (Vol. 4, Issue 1). Nature Research. <https://doi.org/10.1038/s41746-020-00376-2>
- [16] Y. Arun and G. S. Viknesh, "Leaf Classification for Plant Recognition Using EfficientNet Architecture," *2022 IEEE Fourth International Conference on Advances in Electronics, Computers and Communications (ICAEECC)*, Bengaluru, India, 2022, pp. 1-5, doi: <https://doi.org/10.1109/icaeecc54045.2022.9716637>
- [17] Khat, S., Mahmoudi, S. A., Stassin, S., Boukerroui, L., Senaï, B., & Mahmoudi, S. (2025). An Efficient Explainability of Deep Models on Medical Images. *Algorithms*, 18(4). <https://doi.org/10.3390/a18040210>
- [18] Aras, H.H., Doğan, N. A comparative analysis for skin cancer detection by using explainable deep learning. *Neural Comput & Applic* 38, 50 (2026). <https://doi.org/10.1007/s00521-025-11809-y>
- [19] Ramadhan, M. E., & Zeniarja, J. (n.d.). *Sistemasi: Jurnal Sistem Informasi Implementasi Deep Transfer Learning dan Explainable AI dalam Klasifikasi Kanker Kulit Implementation of Deep Transfer Learning and Explainable AI in Skin Cancer Classification. Vol 14. No 5 (2025).* <https://doi.org/10.32520/stmsi.v14i5.5425>
- [20] Roy, A.L., Al Maruf, A., Sony, R.I. *et al.* Fostering trust and interpretability in skin cancer classification: a hybrid framework of deep learning and machine learning with explainable AI. *J Ambient Intell Human Comput* 16, 879–894 (2025). <https://doi.org/10.1007/s12652-025-04984-2>
- [21] S. Nurmaini, and H. Rahmat, "Deep Learning for Improving the Effectiveness of Routine Prenatal Screening for Major Congenital Heart Diseases." *International Journal of Medical Imaging*, vol. 9, no. 2, 2023, doi: <https://doi.org/10.1155/2023/8342974>
- [22] Aboulmira A, Hrimech H, Lachgar M, Hanine M, Garcia CO, Mezquita GM, Ashraf I. Hybrid Model with Wavelet Decomposition and EfficientNet for Accurate Skin Cancer Classification. *J Cancer*. 2025 Jan 1;16(2):506-520. doi: 10.7150/jca.101574
- [23] Behara, K., Behara, S. (2025). Explainable Artificial Intelligence in Dermatology: Overview of Data Scarcity for Skin Cancer Detection in Black Skin. In: Naidoo, P., Sibanda, M. (eds) *Measurement, Automation, and Control in Artificial Intelligence and Machine Learning. Signals and Communication Technology*. Springer, Cham. https://doi.org/10.1007/978-3-031-91802-5_12
- [24] Liu, X., Sangers, T. E., Nijsten, T., Kayser, M., Pardo, L. M., Wolvius, E. B., Roshchupkin, G. v., & Wakkee, M. (2024). Predicting skin cancer risk from facial images with an explainable artificial intelligence (XAI) based approach: a proof-of-concept study. *eClinicalMedicine*, 71. <https://doi.org/10.1016/j.eclinm.2024.102550>
- [25] Dagnaw, G. H., el Mouhtadi, M., & Mustapha, M. (2024). Skin cancer classification using vision transformers and explainable artificial intelligence. *Journal of Medical Artificial Intelligence*, 7(June). <https://doi.org/10.21037/jmai-24-6>