

Perbandingan Strategi Cost-Sensitive Learning Statis dan Dinamis untuk Deteksi Melanoma Menggunakan MobileNetV3-Small pada Dataset HAM10000

Raya Osgibran^a, Resty Wulanningrum^{b*}, Risky Aswi Ramadhani^c

^{a,b,c} Teknik Informatika Universitas Nusantara PGRI Kediri

*restyw@unpkdr.ac.id

Diterima : 21-05-2026

-

Direview : 10-06-2026

-

Diterbitkan : 30-06-2026

Intisari

Penelitian ini membandingkan efektivitas strategi cost-sensitive learning statis dan dinamis untuk deteksi melanoma menggunakan arsitektur MobileNetV3-Small pada dataset HAM10000 yang memiliki ketidakseimbangan kelas sebesar 1:6 antara kelas melanoma dan melanocytic nevi. Ketidakseimbangan data menyebabkan model cenderung bias terhadap kelas mayoritas sehingga meningkatkan risiko false negative yang berbahaya secara klinis. Metode yang dibandingkan meliputi Class Weight, SMOTE, Weighted Random Sampler (WRS), kombinasi SMOTE + Class Weight sebagai pendekatan statis, serta Focal Loss sebagai pendekatan dinamis. Seluruh eksperimen dilakukan menggunakan konfigurasi pelatihan yang identik untuk memastikan perbedaan performa hanya dipengaruhi oleh metode penanganan imbalance. Evaluasi dilakukan menggunakan accuracy, precision, recall, F1-score, AUC-ROC, confusion matrix, dan analisis Grad-CAM. Hasil penelitian menunjukkan bahwa seluruh metode cost-sensitive learning berhasil meningkatkan recall melanoma dibandingkan baseline sebesar 0,5676. Kombinasi SMOTE + Class Weight menghasilkan recall tertinggi sebesar 0,9099, sedangkan SMOTE memperoleh nilai AUC terbaik sebesar 0,9318 dengan keseimbangan performa yang lebih baik. Analisis Grad-CAM menunjukkan bahwa Focal Loss menghasilkan distribusi probabilitas yang lebih terkalibrasi, ditunjukkan oleh probabilitas tertinggi pada kasus false negative dan probabilitas terendah pada false positive. Temuan ini menunjukkan bahwa evaluasi model AI medis tidak cukup hanya menggunakan metrik agregat, tetapi juga memerlukan analisis perilaku model pada kasus-kasus kritis untuk mendukung interpretabilitas dan validitas klinis.

Kata Kunci : Melanoma; Cost-Sensitive Learning; MobileNetV3-Small; Focal Loss; HAM10000; Grad-CAM

Abstract

This study compares the effectiveness of static and dynamic cost-sensitive learning strategies for melanoma detection using the MobileNetV3-Small architecture on the HAM10000 dataset, which has a class imbalance of 1:6 between the melanoma and melanocytic nevi classes. Data imbalance causes the model to be biased toward the majority class, thereby increasing the risk of clinically dangerous false negatives. The methods compared include Class Weight, SMOTE, Weighted Random Sampler (WRS), a combination of SMOTE + Class Weight as a static approach, and Focal Loss as a dynamic approach. All experiments were conducted using identical training configurations to ensure that performance differences were influenced solely by the imbalance handling methods. Evaluation was performed using accuracy, precision, recall, F1-score, AUC-ROC, confusion matrix, and Grad-CAM analysis. The results show that all cost-sensitive learning methods successfully improved melanoma recall compared to the baseline of 0.5676. The combination of SMOTE and Class Weight achieved the highest recall of 0.9099, while SMOTE achieved the best AUC value of 0.9318 with a better balance of performance. Grad-CAM analysis indicates that Focal Loss produces a more calibrated probability distribution, evidenced by the highest probability in false negative cases and the lowest probability in false positive cases. These findings suggest that evaluating medical AI models requires more than just aggregate metrics; it also necessitates analyzing model behavior in critical cases to support interpretability and clinical validity.

Keywords : Melanoma; Cost-Sensitive Learning; MobileNetV3-Small; Focal Loss; HAM10000; Grad-CAM

1. Pendahuluan

Melanoma merupakan salah satu jenis kanker kulit paling berbahaya yang berasal dari sel melanosit. Meskipun hanya menyumbang sekitar 1% dari seluruh kasus kanker kulit, melanoma bertanggung jawab atas sebagian besar kematian akibat kanker kulit (Arnold et al., 2022). Tingkat

kelangsungan hidup sangat bergantung pada stadium deteksi: apabila terdeteksi dini, angka kelangsungan hidup lima tahun mencapai 98%, namun turun drastis menjadi 23% pada stadium lanjut. Deteksi dini menjadi faktor penentu yang kritis dalam meningkatkan prognosis pasien.

Pemeriksaan dermoskopi merupakan teknik non-invasif untuk mengidentifikasi lesi kulit mencurigakan.

Esteva et al. (2017) membuktikan bahwa model deep learning mampu mengklasifikasikan kanker kulit pada tingkat akurasi setara dermatologis berpengalaman, membuka peluang pengembangan sistem computer-aided diagnosis (CAD) yang objektif dan konsisten. Penelitian selanjutnya oleh (Mateen et al. (2024) mengonfirmasi bahwa arsitektur hibrid berbasis CNN mampu meningkatkan performa deteksi melanoma pada dataset HAM10000 secara signifikan.

Dataset HAM10000 (Human Against Machine with 10.000 training images) merupakan benchmark utama dalam penelitian klasifikasi lesi kulit (Tschandl et al., 2018). Dalam konteks deteksi melanoma secara biner, terdapat ketidakseimbangan yang signifikan antara kelas melanoma (mel, 1.113 sampel) dan melanocytic nevi (nv, 6.705 sampel) dengan rasio 1:6. Ketidakseimbangan ini menyebabkan model cenderung bias ke kelas mayoritas sehingga menghasilkan banyak false negative kondisi berbahaya secara klinis (Nguyen et al., 2022).

Metode *cost-sensitive learning* dirancang untuk menangani masalah ini dengan memperbesar penalti pada kesalahan prediksi kelas minoritas. Pendekatan ini dibedakan menjadi dua paradigma: (1) statis, di mana bobot atau distribusi data ditentukan sebelum pelatihan dan tidak berubah sepanjang proses mencakup Class Weight, SMOTE, WRS, dan Kombinasi; serta (2) dinamis, di mana bobot disesuaikan secara otomatis per sampel berdasarkan tingkat kesulitan klasifikasi diwakili oleh Focal Loss (Lin et al., 2017). Selain evaluasi metrik kuantitatif, analisis Gradient-weighted Class Activation Mapping (Grad-CAM) digunakan untuk mengevaluasi apakah perbedaan strategi tersebut juga tercermin dalam pola perhatian visual model (Selvaraju et al., 2017).

Penelitian ini bertujuan menjawab pertanyaan: "Manakah strategi *cost-sensitive learning* statis vs dinamis yang paling efektif untuk deteksi melanoma pada arsitektur CNN lightweight, dan apakah perbedaan strategi tersebut tercermin dalam pola perhatian visual model?" Kontribusi utama adalah analisis komparatif terstruktur antara metode statis dan dinamis pada MobileNetV3-Small dengan evaluasi kuantitatif (metrik) dan kualitatif (Grad-CAM) secara bersamaan.

2. Kerangka Teori

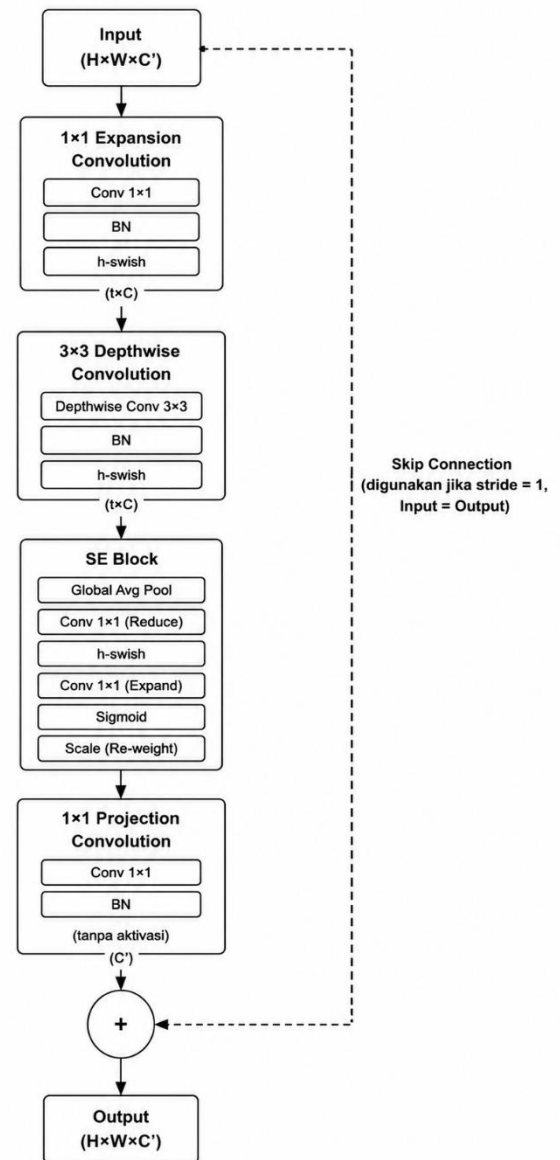
2.1. Melanoma dan Urgensi Deteksi Dini

Melanoma berasal dari transformasi ganas sel melanosit pada lapisan kulit. Dibandingkan kanker kulit non-melanoma, melanoma memiliki agresivitas dan potensi metastasis jauh lebih tinggi (Abohashish et al., 2025). Sistem CAD berbasis deep learning menjadi solusi menjanjikan untuk meningkatkan konsistensi diagnosis. Esteva et al. (2017) membuktikan CNN dapat mencapai akurasi diagnostik setara dermatologis berpengalaman dalam klasifikasi kanker kulit biner, menjadikan pendekatan ini relevan secara klinis.

2.2. Transfer Learning dan MobileNetV3-Small

Transfer learning memanfaatkan bobot model yang dilatih pada dataset besar (ImageNet) sebagai titik awal untuk domain berbeda. Pendekatan ini terbukti efektif ketika data latih terbatas karena fitur-fitur umum yang dipelajari dapat

ditransfer ke domain medis (Pan & Yang, 2010). MobileNetV3-Small dirancang untuk efisiensi komputasi tinggi menggunakan depthwise separable convolution, squeeze-and-excitation blocks, dan aktivasi hard-swish (Howard et al., 2019). Arsitektur ini dipilih karena parameternya minimal namun kompetitif, relevan untuk deployment pada perangkat dengan sumber daya terbatas.



Gambar 1. Inverted Residual Block MobileNetV3-Small

2.3. Ketidakseimbangan Kelas dan Metode Penanganannya

Ketidakseimbangan kelas menyebabkan model memprediksi kelas mayoritas secara dominan, meningkatkan false negative yang berbahaya dalam deteksi penyakit (Nguyen et al., 2022). Ravi (2022) menunjukkan bahwa *cost-sensitive learning* yang memodelkan asimetri biaya kesalahan terbukti meningkatkan kinerja deteksi kelas minoritas. Terdapat dua paradigma pendekatan:

Metode Statis menetapkan bobot atau distribusi data sebelum pelatihan dimulai dan tidak berubah sepanjang proses. Mencakup: (a) Class Weight memberikan bobot lebih besar pada kelas minoritas dalam fungsi loss, dengan formula $w(c) = N/(K \times n(c))$ di mana N adalah total sampel, K jumlah kelas, dan $n(c)$ jumlah sampel kelas c ; (b) SMOTE

menghasilkan sampel sintetis kelas minoritas melalui interpolasi dan diterapkan pada data latih saja (Chawla et al., 2002); (c) Weighted Random Sampler (WRS) mengatur probabilitas pemilihan sampel pada level DataLoader tanpa duplikasi data eksplisit; (d) Kombinasi SMOTE dan Class Weight diterapkan bersamaan, bersifat redundan secara konseptual karena dapat menimbulkan double-penalization pada kelas mayoritas.

Metode Dinamis menyesuaikan bobot secara otomatis per sampel selama pelatihan berdasarkan confidence model, sehingga model fokus pada sampel yang sulit diklasifikasi. Diwakili oleh Focal Loss pada penelitian ini.

2.4. Focal Loss Dynamic Cost-Sensitive Learning

Focal Loss diusulkan oleh Lin et al. (2017) dalam konteks deteksi objek untuk mengatasi ketidakseimbangan antara foreground dan background. Formula Focal Loss adalah:

$$FL(p_i) = -\alpha_i(1-p_i)^\gamma \log(p_i) \quad (1)$$

di mana p_i adalah probabilitas kelas yang benar; α_i adalah bobot kelas (α untuk kelas positif, $1-\alpha$ untuk negatif); dan γ adalah focusing parameter. Ketika model memprediksi suatu sampel dengan kepercayaan tinggi (p_i tinggi), faktor $(1-p_i)^\gamma$ mendekati nol sehingga kontribusinya terhadap total loss diminimalkan. Sebaliknya, sampel yang sulit (p_i rendah) mendapat penalti lebih besar. Dengan $\gamma = 2$, Focal Loss efektif mengurangi bobot sampel mudah sebesar $100\times$ dibandingkan standar cross-entropy. Dalam penelitian ini, α dihitung dari proporsi kelas n_v di training set ($\alpha \approx 0,858$) sehingga konsisten dengan distribusi ketidakseimbangan aktual.

Perbedaan fundamental dengan metode statis: Class Weight mempertahankan bobot yang sama untuk semua sampel dalam satu kelas sepanjang pelatihan. Focal Loss mengubah bobot setiap sampel secara individual berdasarkan confidence model saat itu, memungkinkan fokus yang adaptif pada hard examples sampel yang secara visual ambigu dan paling berisiko salah klasifikasi.

2.5. Metrik Evaluasi pada Klasifikasi Medis

Akurasi tidak cukup representatif pada dataset tidak seimbang (Nguyen et al., 2022). Tabel 1 merangkum metrik yang digunakan.

Tabel 1 Definisi Metrik Evaluasi pada Klasifikasi Biner

Metrik	Formula	Interpretasi Klinis
Recall / Sensitivity	$TP / (TP + FN)$	Proporsi melanoma yang berhasil terdeteksi
Precision / PPV	$TP / (TP + FP)$	Proporsi prediksi mel yang benar-benar melanoma
Specificity / TNR	$TN / (TN + FP)$	Proporsi n_v yang benar diklasifikasi
F1-Score	$2 \times (P \times R) / (P + R)$	Harmonik rata-rata precision dan recall
NPV	$TN / (TN + FN)$	Kepastian hasil negatif bukan melanoma
AUC-ROC	Area bawah kurva ROC	Kemampuan diskriminasi model secara keseluruhan

Recall menjadi metrik primer karena biaya false negative (melanoma terlewat) jauh lebih besar dibandingkan false positive (biopsi tidak perlu). Prinsip asymmetric misclassification cost ini mendasari seluruh pilihan

metodologis penelitian ini (Hicks et al., 2022). AUC-ROC digunakan sebagai metrik sekunder karena mencerminkan kemampuan diskriminasi terlepas dari threshold.

2.6. Gradient-weighted Class Activation Mapping (Grad-CAM)

Grad-CAM (Selvaraju et al., 2017) adalah teknik explainability yang menghasilkan peta perhatian visual (heatmap) yang menunjukkan area dalam gambar yang paling berkontribusi terhadap prediksi model. Grad-CAM bekerja dengan menghitung gradien dari skor kelas target terhadap aktivasi layer konvolusi terakhir:

$$\alpha_k^c = \frac{1}{Z} \sum_i \sum_j \frac{\partial Y^c}{\partial A_{i,j}^k} \quad (2)$$

$$L_{\text{Grad-CAM}}^c = \sum_k \alpha_k^c A^k \quad (3)$$

di mana α^k adalah bobot kepentingan channel k (rata-rata gradien secara global), A^k adalah peta aktivasi channel k , dan y_c adalah skor kelas target sebelum softmax. ReLU memastikan hanya fitur yang berkontribusi positif terhadap prediksi yang ditampilkan. Pada MobileNetV3-Small, target layer adalah blok features[-1] (blok InvertedResidual terakhir) yang memiliki resolusi spasial 7×7 dan mengandung representasi fitur tingkat tinggi paling dekat dengan output klasifikasi.

Dalam konteks penelitian ini, Grad-CAM digunakan untuk membandingkan apakah model dengan strategi cost-sensitive learning yang berbeda menghasilkan pola perhatian yang berbeda pada jenis kasus yang sama khususnya pada sampel melanoma yang sulit dideteksi (hard negatives) dan kasus n_v yang salah diprediksi sebagai melanoma (false positives). Perbandingan kualitatif ini memberikan dimensi analisis yang tidak dapat ditangkap oleh metrik agregat semata.

3. Metodologi

3.1. Metode Penelitian

Penelitian ini menggunakan pendekatan eksperimen komputasi terstruktur dengan desain controlled experiment. Setiap skenario eksperimen dirancang untuk mengisolasi pengaruh satu variabel terhadap performa model. Variabel tetap mencakup arsitektur model, dataset, pembagian data, ukuran gambar, dan konfigurasi pelatihan. Variabel berubah adalah metode penanganan ketidakseimbangan kelas. Seluruh eksperimen bersifat reproduibel melalui penetapan random seed yang konsisten (seed=42).

3.2. Objek Penelitian

Objek penelitian adalah dataset HAM10000, kumpulan citra dermoskopi kulit yang dikurasi secara klinis dan tersedia secara publik. Dataset ini terdiri dari 10.015 citra dengan tujuh kelas lesi kulit. Dalam penelitian ini digunakan dua kelas: melanoma (mel) sebagai kelas positif (label 1) dengan 1.113 sampel, dan melanocytic nevi (nv) sebagai kelas negatif (label 0) dengan 6.705 sampel. Rasio ketidakseimbangan kelas adalah 1:6, yang menjadi tantangan utama penelitian ini.

3.3. Instrumen Penelitian

Tabel 2 Perangkat Lunak

Kategori	Spesifikasi
Bahasa Pemrograman	Python 3.11
Framework Deep Learning	PyTorch 2.8.0+cu128 dengan CUDA dan Automatic Mixed Precision (AMP)
Library Utama	torchvision, scikit-learn, imbalanced-learn, Optuna, tqdm, matplotlib, opencv-python
Lingkungan Eksekusi	Jupyter Notebook (.ipynb) satu notebook per skenario eksperimen
Penyimpanan Model	Format .pth (PyTorch state_dict)
Sistem Operasi	Windows 11

Tabel 3 Perangkat Keras

Kategori	Spesifikasi
Jenis Perangkat	Laptop
Nama Perangkat	MSI GF63 THIN 10SC
CPU	Intel Core i5-10500H @ 2,50 GHz
GPU	NVIDIA GTX 1650
RAM	16 GB

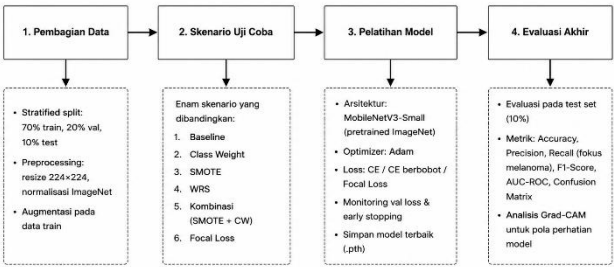
3.4. Data dan Preprocessing

Label kelas diekstrak dari kolom dx pada file HAM10000 metadata.csv. Data dibagi secara stratifikasi untuk mempertahankan proporsi kelas di setiap subset: 70% data latih, 20% data validasi, dan 10% data pengujian, dengan random_state=42. Pemilihan rasio pembagian ini merujuk pada penelitian sebelumnya yang membuktikan bahwa alokasi tersebut memberikan evaluasi kinerja yang komprehensif pada model arsitektur CNN (Andrianto et al., 2026) . Pembagian ini bersifat tetap dan identik di semua skenario untuk memastikan komparabilitas.

Pada data latih, diterapkan augmentasi geometrik: RandomResizedCrop(224), RandomHorizontalFlip(), dan RandomRotation(15°), diikuti normalisasi ImageNet. Pada data validasi dan pengujian, hanya dilakukan Resize(256), CenterCrop(224), dan normalisasi tanpa augmentasi untuk menghindari bias evaluasi.

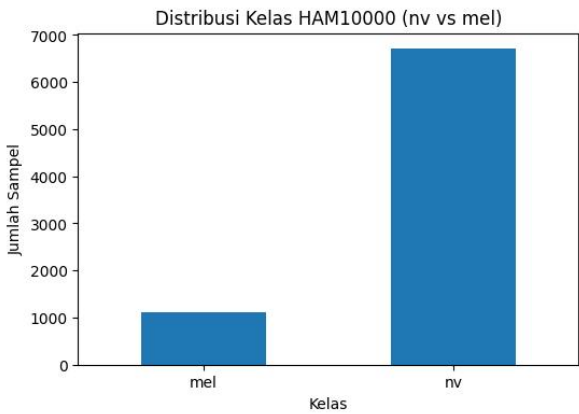
3.5. Metodologi Penelitian

Analisis dilakukan secara bertahap mengikuti alur metodologi yang disajikan pada Gambar 2: mulai dari persiapan dataset, preprocessing, pelatihan model baseline, perbandingan enam metode cost-sensitive learning, evaluasi kuantitatif, hingga analisis visual menggunakan Grad-CAM. Setiap tahap dirancang untuk dapat diisolasi dan dibandingkan secara terkontrol.



Gambar 2. Alur Metodologi Penelitian Klasifikasi Melanoma pada Dataset HAM10000

3.5.1. Pembagian Data



Gambar 3. Distribusi Dataset HAM10000 setelah Stratified Split

Dataset HAM10000 digunakan dalam bentuk klasifikasi biner dengan dua kelas, yaitu melanoma (mel) sebagai kelas positif dan melanocytic nevi (nv) sebagai kelas negatif. Data dibagi secara stratifikasi menjadi data latih (70%), validasi (20%), dan pengujian (10%) menggunakan random_state = 42 untuk menjaga proporsi kelas pada setiap subset. Tahap preprocessing meliputi resize, normalisasi ImageNet, dan augmentasi geometrik pada data latih berupa RandomResizedCrop, RandomHorizontalFlip, dan RandomRotation untuk meningkatkan variasi data dan mengurangi overfitting. Seluruh citra dikonversi ke format RGB dengan ukuran akhir 224×224 piksel sesuai kebutuhan input MobileNetV3-Small.

3.5.2. Skenario Uji Coba

Penelitian ini membandingkan lima metode penanganan ketidakseimbangan kelas secara terkontrol pada arsitektur MobileNetV3-Small. Skenario yang diuji meliputi baseline tanpa penanganan imbalance, Class Weight, SMOTE, Weighted Random Sampler (WRS), kombinasi SMOTE + Class Weight, dan Focal Loss sebagai metode dinamis. Seluruh skenario menggunakan konfigurasi pelatihan yang identik sehingga perbedaan performa yang diperoleh dapat diatribusikan secara langsung pada metode imbalance handling yang digunakan.

Tabel 4 Ringkasan Konfigurasi Lima Variasi Metode

Nama Metode	Paradigma	Level	Komponen Berbeda dari S1
Class Weight	Statis	Loss	CrossEntropyLoss(weight=w(c)); w(c) = N / (K × n(c))
SMOTE	Statis	Data	RandomOverSampler pada data latih; rasio 1:1; mencegah data leakage
WRS	Statis	DataLoader	WeightedRandomSampler; bobot=1/n(c); replacement=True
Kombinasi	Statis	Data+Loss	SMOTE + Class Weight secara bersamaan (ganda)
Focal Loss	Dinamis	Loss	$FL(p_i) = -\alpha_i(1-p_i)^{\gamma} \log(p_i)$; $\alpha=0,858$; $\gamma=2$

3.5.3. Pelatihan Model

Proses pelatihan dilakukan menggunakan arsitektur MobileNetV3-Small dengan bobot awal pretrained

ImageNet melalui pendekatan transfer learning. Model dilatih menggunakan optimizer Adam dan fungsi loss sesuai skenario yang diuji, yaitu CrossEntropyLoss standar, CrossEntropyLoss berbobot, atau Focal Loss (Kingma & Ba, 2017). Pelatihan dilakukan secara bertahap menggunakan data latih dan data validasi untuk memantau performa model pada setiap epoch. Model terbaik dipilih berdasarkan nilai validation loss terendah dan disimpan dalam format .pth untuk digunakan pada tahap evaluasi akhir dan analisis Grad-CAM.

3.5.4. Evaluasi Akhir

Evaluasi akhir dilakukan menggunakan test set yang terpisah sepenuhnya dari proses pelatihan dan validasi. Setiap model dievaluasi menggunakan metrik accuracy, precision, recall, F1-score, AUC-ROC, dan confusion matrix dengan fokus utama pada recall melanoma karena false negative memiliki risiko klinis yang tinggi. Selain itu, dilakukan pemeriksaan distribusi prediksi untuk memastikan model tidak mengalami collapse ke satu kelas tertentu. Hasil evaluasi seluruh skenario kemudian dibandingkan untuk menganalisis efektivitas masing-masing metode imbalance handling.

4. Hasil dan Pembahasan

4.1. Performa Model Baseline

Model baseline (Skenario 1) dilatih menggunakan CrossEntropyLoss standar tanpa intervensi apapun terhadap ketidakseimbangan kelas. Hasil evaluasi pada test set menunjukkan bahwa model mencapai akurasi 90,54%, namun recall kelas melanoma hanya sebesar 0,5676. Nilai ini berarti hampir separuh dari seluruh kasus melanoma di test set (48 dari 111 kasus) tidak berhasil terdeteksi (False Negative = 48). Sebaliknya, model menghasilkan precision yang tinggi (0,7079) dengan False Positive hanya sebesar 26 kasus. Kondisi ini mencerminkan bahwa model tanpa penanganan imbalance cenderung konservatif lebih memilih untuk memprediksi nv dan hanya memprediksi melanoma ketika model sangat yakin, sehingga menghasilkan banyak false negative yang berbahaya secara klinis. Nilai AUC sebesar 0,9253 menunjukkan bahwa kemampuan diskriminasi model secara keseluruhan sebetulnya cukup baik, namun threshold keputusan default tidak optimal untuk kelas yang tidak seimbang.

4.2. Perbandingan Metode Cost-Sensitive Learning

Tabel 5 menyajikan perbandingan performa seluruh metode cost-sensitive learning yang diujikan terhadap model baseline. Semua metode diujikan dengan konfigurasi identik kecuali pada komponen imbalance handling, sehingga perbedaan performa dapat sepenuhnya diatribusikan pada perbedaan metode.

Tabel 5 Perbandingan Performa Seluruh Metode pada Test Set

Skenario	Accuracy	Precision	Recall ★	F1	AUC	FN ↓	FP
S1 Baseline	0,9054	0,7079	0,5676	0,6300	0,9253	48	26

S2 Class Weight	0,8504	0,4848	0,8649	0,6214	0,9240	15	102
S3 SMOTE	0,8875	0,5816	0,7387	0,6508	0,9318	29	59
S4 WRS	0,8862	0,5764	0,7477	0,6510	0,9275	28	61
S5 Kombinasi	0,7992	0,4073	0,9099	0,5627	0,9282	10	147
S6 Focal Loss	0,8529	0,4894	0,8288	0,6154	0,9239	19	96

★ = metrik prioritas utama, ↓ = semakin kecil semakin baik, Bold = nilai terbaik per kolom, sedangkan Merah pada S2d = overconfident klinis

4.2.1. Analisis Metode Statis

Keempat metode statis (Class Weight, SMOTE, WRS, dan Kombinasi) seluruhnya meningkatkan recall melanoma secara signifikan dibandingkan baseline. Class Weight menghasilkan recall tertinggi kedua (0,8649) dengan FN terkecil kedua (15 kasus) di antara metode statis. Mekanismenya adalah memperbesar penalti kesalahan prediksi pada kelas melanoma melalui bobot statis di fungsi loss, sehingga model didorong lebih agresif dalam memprediksi kelas minoritas. Dampak sampingnya adalah peningkatan FP yang signifikan (102 kasus) dan penurunan precision menjadi 0,4848.

SMOTE dan WRS menghasilkan performa yang sangat mirip satu sama lain (recall 0,7387 vs 0,7477; F1 0,6508 vs 0,6510; AUC 0,9318 vs 0,9275), yang menunjukkan bahwa oversampling berbasis data dan pembobotan sampler pada level DataLoader memberikan efek yang setara pada dataset ini. SMOTE menghasilkan AUC tertinggi di antara seluruh metode imbalance handling (0,9318), menandakan kemampuan diskriminasi yang paling baik. Keunggulan SMOTE dan WRS terletak pada keseimbangan yang lebih baik antara recall dan precision dibandingkan Class Weight.

Kombinasi (SMOTE + Class Weight) menghasilkan recall tertinggi dari seluruh metode (0,9099) dengan FN hanya 10 kasus, namun pada saat yang bersamaan menghasilkan FP tertinggi (147 kasus) dan precision terendah (0,4073). Secara klinis, Kombinasi berhasil mendeteksi melanoma paling banyak, namun lebih dari 59% dari semua prediksi melanoma ternyata salah. Kondisi ini terjadi karena penerapan dua mekanisme imbalance handling sekaligus bersifat redundan SMOTE sudah menyeimbangkan distribusi data, sehingga penambahan class weight menjadi double-penalizing kelas mayoritas dan mendorong model ke perilaku yang terlalu agresif.

4.2.2. Analisis Metode Dinamis Focal Loss

Focal Loss (Skenario 6) merupakan satu-satunya metode dinamis yang diujikan. Secara metrik agregat, Focal Loss menghasilkan recall 0,8288 dengan FN = 19, precision 0,4894, F1 0,6154, dan AUC 0,9239. Nilai-nilai ini menempatkan Focal Loss di posisi tengah di antara metode statis lebih tinggi recall-nya dari SMOTE dan WRS, namun lebih rendah dari Class Weight. Jika evaluasi berhenti pada metrik agregat, Focal Loss tampak tidak unggul secara signifikan.

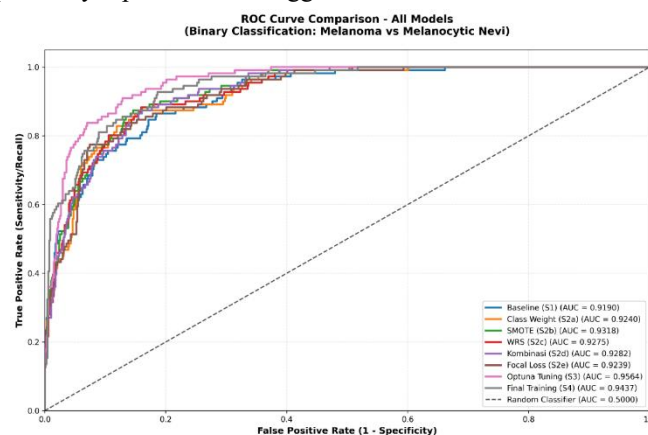
Namun demikian, analisis Grad-CAM yang diuraikan pada Subbab 4.4 mengungkap perbedaan kualitatif yang tidak tampak dari metrik agregat khususnya pada bagaimana model menangani sampel hard cases (melanoma yang sulit diklasifikasi) dan distribusi kepercayaan model pada kasus false positive. Ini merupakan implikasi langsung dari mekanisme focusing $\gamma=2$ yang membuat model secara dinamis memperbobot sampel berdasarkan tingkat kesulitannya, berbeda dari metode statis yang mempertahankan bobot yang sama sepanjang proses pelatihan.

4.3. Analisis ROC Curve

Gambar 4 menyajikan perbandingan ROC curve dari seluruh model pada test set. Kurva ROC memvisualisasikan hubungan antara True Positive Rate (recall/sensitivity) dan False Positive Rate ($1 - \text{specificity}$) pada seluruh rentang threshold, sehingga memberikan gambaran kemampuan diskriminasi model yang lebih komprehensif.

Seluruh metode imbalance handling menghasilkan kurva ROC di atas baseline, yang mengonfirmasi bahwa semua intervensi memberikan peningkatan kemampuan diskriminasi model. SMOTE menghasilkan AUC tertinggi (0,9318), diikuti oleh Kombinasi (0,9282), WRS (0,9275), Class Weight (0,9240), dan Focal Loss (0,9239). Namun perbedaan AUC antar metode sangat kecil (rentang hanya 0,0079 di antara S2a–S2e), mengindikasikan bahwa semua metode memiliki kemampuan diskriminasi yang setara secara statistik.

Yang menarik dari kurva ROC adalah perbedaan bentuk kurva di rentang FPR rendah (0–0,2), yang relevan secara klinis karena sistem skrining medis idealnya beroperasi pada titik ini. Pada rentang tersebut, metode berbasis loss-level (Class Weight dan Focal Loss) menunjukkan TPR yang relatif lebih tinggi dibandingkan SMOTE dan WRS, mengindikasikan bahwa pendekatan berbasis fungsi loss lebih efektif dalam mengidentifikasi melanoma ketika specificity dipertahankan tinggi.



Gambar 4. ROC Curve Comparison Semua Model pada Test Set

4.4. Analisis Grad-CAM Pola Perhatian Visual

Untuk menjawab komponen kedua dari research question apakah perbedaan strategi cost-sensitive learning tercermin dalam pola perhatian visual model dilakukan analisis Grad-CAM pada enam model (S1, S2a–S2e) terhadap dua sampel per kategori: True Positive (TP), False Negative (FN), dan False Positive (FP). Gambar 5–7 menyajikan visualisasi Grad-CAM untuk masing-masing kategori.

4.4.1. True Positive Melanoma Terdeteksi dengan Benar

Pada sampel TP (Gambar 5), semua model berhasil memprediksi melanoma dengan benar dan dengan kepercayaan tinggi ($P(\text{mel}) > 0,75$ untuk semua model). Secara visual, perhatian model-model dengan metode statis (Class Weight, SMOTE, WRS, Kombinasi) cenderung lebih terfokus pada area sentral lesi yang menampilkan fitur dermoskopi khas melanoma seperti variasi warna dan pola

tidak reguler. Model Baseline menunjukkan perhatian yang lebih tersebar ke tepi gambar, meskipun tetap berhasil memprediksi dengan benar. Focal Loss memiliki nilai $P(\text{mel})$ sedikit lebih rendah dibandingkan metode statis pada sampel TP (0,899 dan 0,765 vs $>0,87$ pada metode lain), mengindikasikan bahwa mekanisme focusing Focal Loss membuat model lebih "berhati-hati" bahkan pada kasus yang relatif jelas.

4.4.2. False Negative Melanoma yang Terlewat (Temuan Utama)

Analisis sampel FN merupakan bagian paling kritis dari evaluasi klinis dan sekaligus menghasilkan temuan paling signifikan penelitian ini. Kedua sampel FN yang dianalisis (ISIC_0029013 dan ISIC_0030034) merupakan kasus melanoma yang tidak berhasil dideteksi oleh satupun model, namun terdapat perbedaan tingkat kepercayaan yang substansial antar metode. Tabel 6 menyajikan nilai $P(\text{mel})$ setiap model pada kedua sampel FN tersebut.

Tabel 6 Nilai $P(\text{mel})$ pada Sampel False Negative Semakin Tinggi Berarti Model Lebih Dekat Mendeteksi

Model	FN Sampel 1 (ISIC_0029013)	FN Sampel 2 (ISIC_0030034)	Rata-rata $P(\text{mel})$
S1 Baseline	0,002	0,010	0,006
S2 Class Weight	0,036	0,220	0,128
S3 SMOTE	0,018	0,079	0,049
S4 WRS	0,031	0,075	0,053
S5 Kombinasi	0,088	0,257	0,173
S6 Focal Loss	0,200	0,390	0,295

Baseline menunjukkan $P(\text{mel})$ yang sangat rendah pada kedua sampel (0,002 dan 0,010), yang berarti model hampir sepenuhnya yakin bahwa sampel ini bukan melanoma. Analisis Grad-CAM memperlihatkan bahwa perhatian Baseline pada sampel FN justru tertuju pada sudut gambar dan area kulit non-lesi, bukan pada lesi itu sendiri mengindikasikan kegagalan representasi fitur yang fundamental pada model tanpa penanganan imbalance.

Focal Loss secara konsisten menghasilkan $P(\text{mel})$ tertinggi pada kedua sampel FN (0,200 dan 0,390), dengan rata-rata 0,295 jauh di atas metode lain. Ini berarti meskipun Focal Loss tetap memprediksi "nv" (karena threshold 0,5 belum terpenuhi), modelnya paling dekat dengan ambang

deteksi dibandingkan semua metode lain. Temuan ini sesuai dengan mekanisme focusing parameter γ pada Focal Loss: melanoma yang sulit diklasifikasi merupakan tepat target yang dioptimalkan oleh Focal Loss, sehingga representasi fiturnya berkembang lebih baik untuk kasus-kasus ambigu.

4.4.3. False Positive Kepercayaan Model pada Prediksi yang Salah

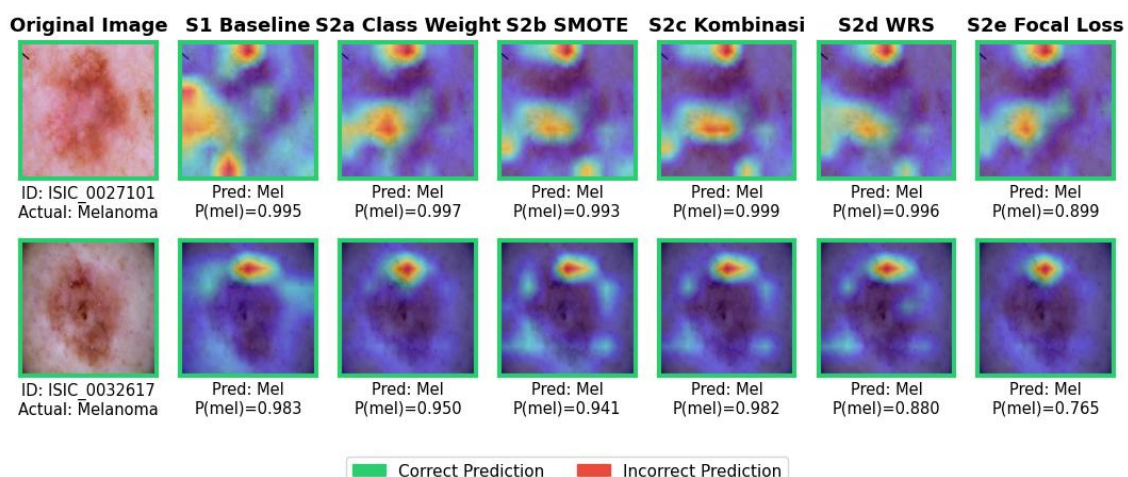
Analisis sampel FP mengungkap dimensi lain dari perbedaan perilaku antar metode. Tabel 7 menampilkan nilai $P(\text{mel})$ dari setiap model pada dua sampel nv yang secara salah diprediksi sebagai melanoma.

Tabel 7 Nilai $P(\text{mel})$ pada Sampel False Positive Semakin Rendah Berarti Model Kurang Overconfident

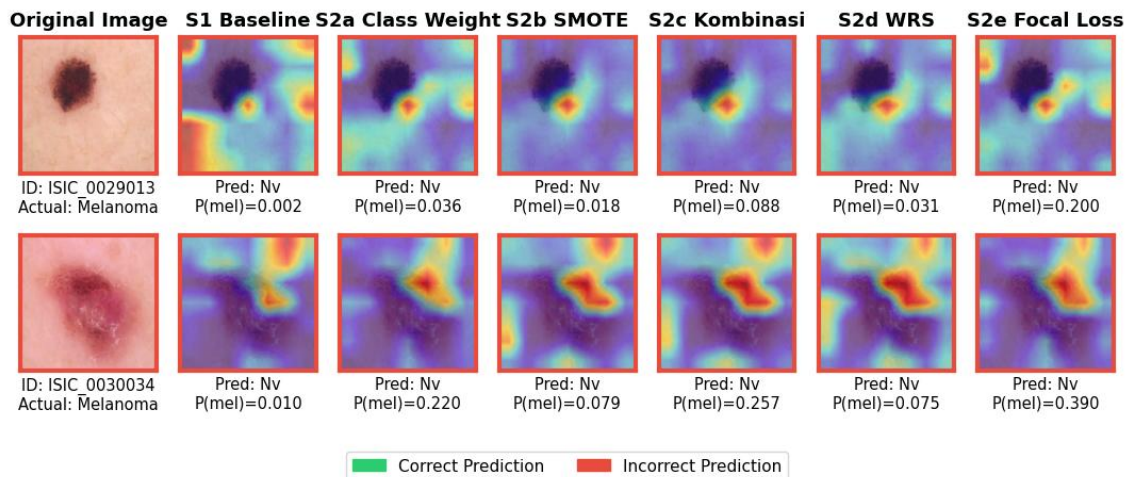
Model	FP Sampel 1 (ISIC_0033222)	FP Sampel 2 (ISIC_0027568)	Rata-rata $P(\text{mel})$
S1 Baseline	0,929	0,901	0,915
S2 Class Weight	0,949	0,870	0,910
S3 SMOTE	0,864	0,772	0,818
S4 WRS	0,936	0,719	0,828
S5 Kombinasi	0,985	0,930	0,958
S6 Focal Loss	0,752	0,675	0,714

Kombinasi menghasilkan $P(\text{mel})$ tertinggi pada kedua sampel FP (0,985 dan 0,930) perilaku overconfident yang berbahaya dalam konteks klinis. Sebaliknya, Focal Loss menghasilkan $P(\text{mel})$ paling rendah pada kedua sampel FP (0,752 dan 0,675), menandakan bahwa model masih salah prediksi namun dengan tingkat kepercayaan yang paling rendah. Dalam praktik klinis, prediksi FP dengan kepercayaan rendah lebih mudah dikelola dokter dapat memprioritaskan konfirmasi biopsi berdasarkan skor kepercayaan model.

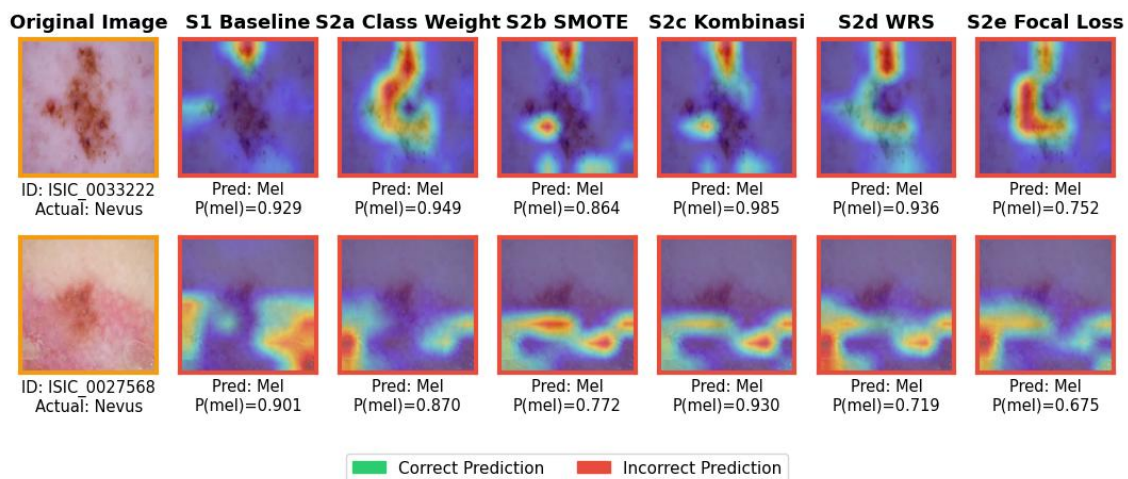
Pola ini $P(\text{mel})$ tertinggi pada FN dan terendah pada FP mengindikasikan bahwa Focal Loss menghasilkan distribusi probabilitas yang lebih terkalibrasi. Model tidak sekadar menggeser threshold secara uniform seperti metode statis, melainkan belajar membedakan sampel yang sulit dari yang mudah secara lebih halus. Hal ini merupakan manifestasi dari tujuan desain Focal Loss yang secara eksplisit menargetkan hard negatives dan hard positives dalam proses pembelajaran (Hicks et al., 2022).



Gambar 5. Grad-CAM True Positive Semua Model



Gambar 6. Grad-CAM False Negative Semua Model



Gambar 7. Grad-CAM False Positive Semua Model

4.5. Perbandingan dengan Penelitian Terdahulu

Tabel 8 membandingkan hasil penelitian ini dengan penelitian terkait yang menggunakan dataset HAM10000 untuk klasifikasi melanoma.

Tabel 8 Perbandingan dengan Penelitian Terdahulu pada Dataset HAM10000

Penelitian	Arsitektur	Metode Imbalance	Recall Mel	AUC
Mateen et al. (2024)	Hybrid CNN-LSTM	Augmentasi + resampling	-	0,9480
Nguyen et al. (2022)	DenseNet + Attention	SMOTE + oversampling	-	0,9310
Areej Alotaibi (2025)	Xception + Attention	Augmentasi	95,47%	0,987
Kumar Lilhore et al. (2024)	U-Net + MobileNetV3 + Bayesian	Augmentasi	96,45%	0,9845

Penelitian ini SMOTE	MobileNetV3-Small	SMOTE	0,7387	0,9318
Penelitian ini Class Weight	MobileNetV3-Small	Class Weight	0,8649	0,9240
Penelitian ini Focal Loss	MobileNetV3-Small	Focal Loss	0,8288	0,9239

Penelitian Mateen et al. (2024) menggunakan arsitektur hybrid CNN-LSTM yang jauh lebih kompleks dan mencapai AUC 0,9480. Meskipun AUC ini lebih tinggi dari metode terbaik dalam penelitian ini (SMOTE, AUC 0,9318), arsitektur hybrid tersebut membutuhkan sumber daya komputasi yang signifikan lebih besar. Penelitian ini menunjukkan bahwa MobileNetV3-Small arsitektur lightweight dengan parameter minimal dapat mencapai AUC yang kompetitif (0,9318 dengan SMOTE) sambil secara signifikan meningkatkan recall melanoma, yang merupakan metrik prioritas dalam konteks skrining klinis.

4.6. Diskusi Static vs Dynamic Cost-Sensitive Learning

Berdasarkan seluruh hasil analisis, penelitian ini memberikan jawaban terhadap research question utama. Secara metrik agregat, tidak ada satu metode yang secara absolut superior dalam seluruh dimensi evaluasi. Masing-masing metode menunjukkan keunggulan pada aspek yang berbeda: Class Weight unggul pada recall, SMOTE unggul pada AUC dan F1, Kombinasi unggul pada recall namun dengan biaya FP yang sangat tinggi.

Namun, ketika analisis diperluas ke dimensi kualitatif melalui Grad-CAM, Focal Loss metode dinamis satu-satunya menunjukkan keunggulan yang konsisten pada dua aspek yang paling relevan secara klinis: (1) distribusi probabilitas tertinggi pada sampel melanoma yang sulit dideteksi (hard negatives), dan (2) distribusi probabilitas terendah pada kasus false positive (overconfidence paling rendah). Karakteristik ini merupakan konsekuensi langsung dari mekanisme dynamic focusing yang membedakan Focal Loss dari metode statis secara fundamental.

Dari perspektif metodologi penelitian, temuan ini memperkuat argumen bahwa evaluasi model AI medis tidak cukup dilakukan hanya berdasarkan metrik agregat. Metrik seperti recall dan AUC mencerminkan performa rata-rata pada keseluruhan test set, namun tidak menangkap bagaimana model berperilaku pada kasus-kasus yang paling kritis kasus yang secara klinis paling sulit dan paling berisiko jika salah diklasifikasi. Analisis Grad-CAM memberikan dimensi evaluasi tambahan yang bersifat komplementer dan sangat relevan untuk validasi model dalam konteks medis (Hicks et al., 2022).

Keterbatasan penelitian ini meliputi: (1) eksperimen dijalankan dengan satu random seed sehingga tidak tersedia estimasi variabilitas statistik; (2) analisis Grad-CAM hanya dilakukan pada enam sampel representatif sehingga tidak dapat digeneralisasi secara penuh; dan (3) perbandingan dibatasi pada arsitektur MobileNetV3-Small sehingga temuan perlu divalidasi pada arsitektur yang berbeda.

5. Simpulan

Penelitian ini secara sistematis membandingkan strategi cost-sensitive learning statis (Class Weight, SMOTE, WRS, Kombinasi Class Weight + SMOTE) dan cost-sensitive learning dinamis (Focal Loss) untuk deteksi melanoma menggunakan MobileNetV3-Small pada dataset HAM10000. Seluruh metode berhasil meningkatkan recall melanoma secara signifikan dari baseline 0,5676, dengan peningkatan tertinggi dicapai oleh Kombinasi (0,9099) dan Class Weight (0,8649). SMOTE menghasilkan AUC tertinggi (0,9318) dengan keseimbangan terbaik antara recall dan precision. Analisis Grad-CAM mengungkap bahwa Focal Loss metode dinamis secara konsisten menghasilkan probabilitas prediksi tertinggi pada kasus melanoma yang sulit dideteksi (rata-rata $P(\text{mel}) = 0,295$ pada sampel FN vs 0,006 pada Baseline) dan probabilitas terendah pada kasus false positive (rata-rata $P(\text{mel}) = 0,714$ vs 0,958 pada Kombinasi), menunjukkan distribusi probabilitas yang lebih terkalibrasi. Temuan ini menunjukkan bahwa evaluasi model AI medis perlu melampaui metrik agregat dan menyertakan analisis

perilaku model pada kasus-kasus kritis yang paling relevan secara klinis.

Daftar Pustaka

- Abohashish, S. M. M., Amin, H. H., & Elsedimy, E. I. (2025). Enhanced melanoma and non-melanoma skin cancer classification using a hybrid LSTM-CNN model. *Scientific Reports*, 15(1), 24994. <https://doi.org/10.1038/s41598-025-08954-8>
- Alotaibi, A., & AlSaeed, D. (2025). Skin Cancer Detection Using Transfer Learning and Deep Attention Mechanisms. *Diagnostics*, 15(1). <https://doi.org/10.3390/diagnostics15010099>
- Andrianto, M. R., Mandyartha, E. P., & Puspaningrum, E. Y. (2026). Effect of CBAM Integration on InceptionV3 for Improved Foot and Mouth Disease Detection Accuracy. *Bit-Tech*, 8(3), 3944–3954. <https://doi.org/10.32877/bt.v8i3.3756>
- Arnold, M., Singh, D., Laversanne, M., Vignat, J., Vaccarella, S., Meheus, F., Cust, A. E., de Vries, E., Whiteman, D. C., & Bray, F. (2022). Global Burden of Cutaneous Melanoma in 2020 and Projections to 2040. *JAMA Dermatology*, 158(5), 495–503. <https://doi.org/10.1001/jamadermatol.2022.0160>
- Chawla, N. V., Bowyer, K. W., Hall, L. O., & Kegelmeyer, W. P. (2002). SMOTE: Synthetic Minority Over-sampling Technique. In *Journal of Artificial Intelligence Research* (Vol. 16).
- 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>
- Hicks, S. A., Strümke, I., Thambawita, V., Hammou, M., Riegler, M. A., Halvorsen, P., & Parasa, S. (2022). On evaluation metrics for medical applications of artificial intelligence. *Scientific Reports*, 12(1), 5979. <https://doi.org/10.1038/s41598-022-09954-8>
- Howard, A., Sandler, M., Chu, G., Chen, L.-C., Chen, B., Tan, M., Wang, W., Zhu, Y., Pang, R., Vasudevan, V., Le, Q. V., & Adam, H. (2019). *Searching for MobileNetV3*. <http://arxiv.org/abs/1905.02244>
- Kingma, D. P., & Ba, J. (2017). *Adam: A Method for Stochastic Optimization*. <http://arxiv.org/abs/1412.6980>
- Kumar Lilhore, U., Simaiya, S., Sharma, Y. K., Kaswan, K. S., Rao, K. B. V. B., Rao, V. V. R. M., Baliyan, A., Bijalwan, A., & Alrooba, R. (2024). A precise model for skin cancer diagnosis using hybrid U-Net and improved MobileNet-V3 with hyperparameters optimization. *Scientific Reports*, 14(1). <https://doi.org/10.1038/s41598-024-54212-8>
- Lin, T.-Y., Goyal, P., Girshick, R. B., He, K., & Dollár, P. (2017). Focal Loss for Dense Object Detection. *CoRR*, abs/1708.02002. <http://arxiv.org/abs/1708.02002>
- Mateen, M., Hayat, S., Arshad, F., Gu, Y.-H., & Al-antari, M. A. (2024). Hybrid Deep Learning Framework for Melanoma Diagnosis Using Dermoscopic Medical Images. *Diagnostics*, 14(19). <https://doi.org/10.3390/diagnostics14192242>
- Nguyen, V. D., Bui, N. D., & Do, H. K. (2022). Skin Lesion Classification on Imbalanced Data Using Deep Learning

- with Soft Attention. *Sensors*, 22(19).
<https://doi.org/10.3390/s22197530>
- Pan, S. J., & Yang, Q. (2010). A Survey on Transfer Learning. *IEEE Transactions on Knowledge and Data Engineering*, 22(10), 1345–1359.
<https://doi.org/10.1109/TKDE.2009.191>
- Ravi, V. (2022). Attention Cost-Sensitive Deep Learning-Based Approach for Skin Cancer Detection and Classification. *Cancers*, 14(23).
<https://doi.org/10.3390/cancers14235872>
- 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. *2017 IEEE International Conference on Computer Vision (ICCV)*, 618–626.
<https://doi.org/10.1109/ICCV.2017.74>
- Tschandl, P., Rosendahl, C., & Kittler, H. (2018). Data descriptor: The HAM10000 dataset, a large collection of multi-source dermatoscopic images of common pigmented skin lesions. *Scientific Data*, 5.
<https://doi.org/10.1038/sdata.2018.161>