Journal of Digital Ecosystem for Natural Sustainability (JoDENS)
Vol.5, No.2, Desember 2025, pp. 55~66
e-ISSN: 2798-6179
DOI: 10.63643/jodens.v5i2.325

Arsitektur Hybrid Vision Transformer–ConvNeXt dengan Multi-Task Focal Loss dan Medical Test-Time Augmentation untuk Klasifikasi Lesi Kulit Berbasis Citra

Hendry[1*], Ferry Govert Anwar[2], David Chow[3], Andi Saputra[4], Muhammad Khaerul Naim Mursalim[4]
[1]Fakultas Komputer, Program Studi Teknik Informatika, Universitas Universal
**Corresponding author E-mail: hendryhuang70@gmail.com*

Article Info

Article history:

Received 12-12-2025
Revised 17-12-2025
Accepted 29-12-2025

Keyword:

Klasifikasi lesi kulit, Vision Transformer, ConvNeXt, Focal Loss, Test-Time Augmentation, HAM10000.

ABSTRACT

Dermatoscopy image-based skin lesion classification is a challenge in dermatology due to the high visual variation between lesion types and the imbalanced class distribution in the dataset. In this study, a Hybrid Vision Transformer–ConvNeXt architecture is proposed, combining the global attention capability of Vision Transformer (ViT) and the spatial feature representation of ConvNeXt, to improve the classification performance of skin lesion images on the HAM10000 dataset. This study also applies Multi-Task Focal Loss, auxiliary classifier, and Weighted Random Sampler to effectively address the class imbalance. In addition, the Medical Test-Time Augmentation (TTA) approach is used in the inference stage to improve the stability of predictions. The model is trained using a two-stage strategy (head training and full fine-tuning), as well as optimization based on AdamW and Cosine Annealing Warm Restarts. The test results show that the proposed model successfully achieves a validation F1-Score of 0.8723, and after TTA it increases to 0.90, surpassing the baseline of ViT and single ConvNeXt. These findings indicate that the integration of ViT–ConvNeXt with loss strategy and medical TTA is able to significantly improve the performance of skin lesion classification, and has the potential to be applied as a clinical diagnosis support system.

I. PENDAHULUAN

Penyakit kulit merupakan salah satu masalah kesehatan yang umum dialami masyarakat. Meskipun sebagian bersifat ringan, terdapat pula jenis penyakit kulit yang berbahaya, seperti melanoma yang merupakan salah satu jenis kanker yang pertumbuhannya paling cepat. Kabar baiknya penyakit ini mudah untuk diobati apabila terdeksi sejak dini[1]. Tantangan utama dalam diagnosis dini adalah banyaknya jenis penyakit kulit yang memiliki kemiripan secara visual, sehingga proses diagnosis sering kali membutuhkan pengalaman klinis yang tinggi dari seorang dokter. Pada beberapa kasus, perbedaan antarkondisi kulit sangat halus dan sulit dikenali hanya dengan pengamatan mata.

Deteksi dini penyakit kulit, terutama kanker kulit, menjadi aspek yang sangat krusial. Semakin cepat suatu lesi dikenali dan ditangani, semakin besar peluang kesembuhan pasien. Pada kasus melanoma, misalnya, diagnosis pada tahap awal dapat meningkatkan tingkat keberhasilan pengobatan hingga lebih dari 90%.

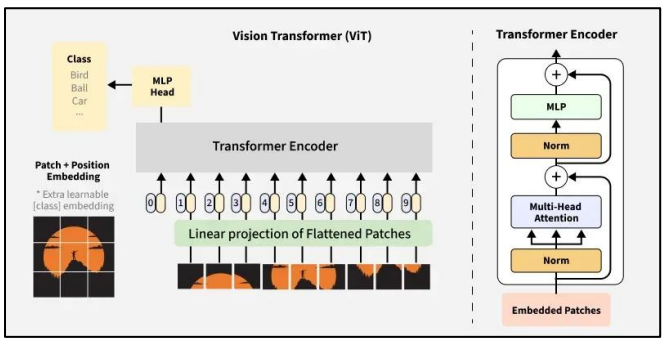
Namun, proses klasifikasi penyakit kulit bukanlah hal yang sederhana. Tingkat kemiripan antarciri seperti warna, bentuk, dan tekstur lesi membuat identifikasi menjadi sulit[2]. Dokter kulit harus memperhatikan detail-detail kecil, termasuk perubahan warna yang sangat halus, pola pigmentasi, hingga ketegasan batas tepi lesi yang sering kali tidak jelas.

Selain itu, pemeriksaan secara manual sangat bergantung pada pengalaman dan ketelitian dokter[3]. Ketergantungan ini dapat menimbulkan subjektivitas interpretasi, sehingga hasil diagnosis dari satu tenaga medis dengan yang lain berpotensi berbeda. Kondisi tersebut dapat menyebabkan inkonsistensi diagnosis dan meningkatkan risiko terjadinya kesalahan medis.

Penerapan kecerdasan buatan dalam diagnosis berbantuan citra medis untuk memberikan opini kedua bagi dokter dalam mendiagnosis kanker kulit telah menjadi tren yang tak terhindarkan. *Convolutional Neural Network* (CNN), sebuah topik penelitian penting dalam pembelajaran mendalam, telah banyak digunakan di bidang medis[4]. Meskipun demikian, CNN memiliki keterbatasan dalam menangkap global dependencies pada citra resolusi tinggi, yang merupakan aspek penting dalam mendeteksi pola global pada lesi kulit.

Pada tahun 2021, Vision Transformer (ViT) diperkenalkan sebagai model berbasis self-attention yang unggul dalam mempelajari hubungan global pixel-level[5]. Namun, ViT memerlukan dataset berukuran besar agar stabil saat pelatihan, sehingga sulit diterapkan secara langsung pada dataset medis yang biasanya terbatas.

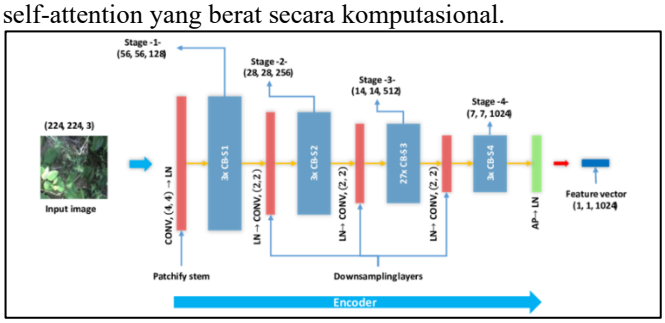
Vision Transformer (ViT) merupakan arsitektur pengenalan citra yang mengadaptasi mekanisme self-attention dari Transformer, yang sebelumnya sukses digunakan pada pemrosesan bahasa alami, untuk tugas klasifikasi gambar. Pada tahap awal, citra masukan berukuran $H \times W \times 3$ tidak diproses secara konvolusional seperti pada CNN, melainkan dipartisi menjadi sekumpulan patch berukuran tetap, misalnya 16×16 piksel. Setiap patch kemudian di-flatten menjadi vektor satu dimensi dan diproyeksikan secara linear ke dalam ruang embedding berdimensi tetap. Untuk memberikan model kemampuan memahami urutan spasial, setiap embedding patch ditambahkan dengan positional embedding yang dapat dipelajari. Selain itu, sebuah token khusus yang disebut class token $[CLS]$ disisipkan pada awal deretan embedding, yang berfungsi sebagai representasi global citra. Seluruh urutan embedding ini kemudian dieksekusi melalui beberapa blok Transformer Encoder yang terdiri atas Multi-Head Self Attention (MHSA) untuk menangkap dependensi antarpatch, dilanjutkan dengan Feed Forward Network (MLP) serta mekanisme residual connection dan normalisasi layer untuk menjaga stabilitas pelatihan. Setelah melewati seluruh lapisan encoder, keluaran dari token $[CLS]$ yang telah merepresentasikan informasi global citra akan diteruskan ke sebuah MLP head yang melakukan proses klasifikasi akhir terhadap kelas target. Dengan pendekatan ini, ViT mampu mempelajari hubungan global antarbagian gambar tanpa operasi konvolusi, sehingga menawarkan performa kompetitif terutama ketika dilatih menggunakan dataset berukuran besar.



Gambar 1. Arsitektur ViT

Di sisi lain, arsitektur ConvNeXt berhasil menghadirkan kembali performa CNN dengan desain modern yang menyerupai transformer[6].Dengan menggabungkan kekuatan ViT sebagai pemodel konteks global dan ConvNeXt sebagai ekstraktor fitur spasial, muncullah pendekatan hybrid architecture yang berpotensi meningkatkan akurasi klasifikasi citra medis.

ConvNeXt merupakan arsitektur konvolusional modern yang didesain untuk menjembatani capaian performa model Vision Transformer (ViT) namun tetap mempertahankan paradigma CNN konvensional yang efisien. Pada tahap awal, citra masukan berukuran $224 \times 224 \times 3$ diproses oleh patchify stem, yaitu sebuah lapisan konvolusi dengan kernel berukuran besar yang berfungsi mengekstraksi fitur awal sekaligus mereduksi resolusi spasial citra menjadi 56×56 dengan kedalaman 128 kanal fitur. Tahap ini menghasilkan representasi awal yang sebanding dengan patch embedding pada ViT, namun dicapai menggunakan operasi konvolusi. Selanjutnya, fitur tersebut diteruskan ke dalam empat stages bertingkat, masing-masing terdiri dari sejumlah blok ConvNeXt yang memanfaatkan desain ResNet-like dengan penyempurnaan arsitektural, seperti normalisasi layer (Layer Normalization), penggunaan kernel konvolusi berukuran besar, serta aktivasi GELU yang lebih stabil. Setiap stage melakukan proses downsampling melalui konvolusi ber-stride 2 sehingga dimensi fitur berkurang secara bertahap dari $56 \times 56 \rightarrow 28 \times 28 \rightarrow 14 \times 14 \rightarrow 7 \times 7$, disertai peningkatan jumlah kanal menjadi 256, 512, dan akhirnya 1024. Struktur hierarkis ini memungkinkan ConvNeXt menangkap informasi lokal dan global secara progresif, menyerupai pendekatan feature pyramid dalam model-model modern. Pada akhir proses encoding, keluaran fitur berukuran $7 \times 7 \times 1024$ diratakan melalui operasi adaptive average pooling sehingga membentuk vektor fitur berdimensi $1 \times 1 \times 1024$. Vektor inilah yang kemudian dapat diteruskan ke classification head atau modul tugas lain seperti deteksi ataupun segmentasi. Dengan rancangan modular dan efisiensi komputasi tinggi, ConvNeXt membuktikan bahwa CNN yang dioptimalkan dengan desain arsitektural modern dapat bersaing, bahkan melampaui, performa ViT pada berbagai benchmark visi komputer, tanpa memerlukan mekanisme self-attention yang berat secara komputasional.



Gambar 2. Arsitektur ConvNeXt Base

Selain tantangan arsitektur, dataset lesi kulit seperti HAM10000 memiliki masalah ketidakseimbangan kelas, di mana kelas melanocytic nevus (nv) jauh lebih dominan dibanding akiec atau dermatofibroma (df)[7]. Masalah ini dapat menyebabkan model bias pada kelas mayoritas. Oleh karena itu, penelitian ini mengadopsi Multi-Task Focal Loss untuk memfokuskan pembelajaran pada kelas minor[8], Weighted Random Sampler selama training, *Test-Time Augmentation* (TTA) khusus domain medis untuk meningkatkan stabilitas prediksi pada saat inferensi[9], Dengan pendekatan ini, penelitian ini berupaya mencapai tingkat akurasi tinggi dan F1-score ≥ 0.90, sehingga dapat digunakan sebagai sistem pendukung keputusan bagi ahli dermatologi.

Tujuan dari penelitian ini adalah untuk mengembangkan arsitektur Hybrid Vision Transformer–ConvNeXt yang mampu mempelajari fitur lokal dan global secara bersamaan pada citra lesi kulit[10]. Sekaligus mengatasi ketidakseimbangan kelas menggunakan Multi-Task Focal Loss dan auxiliary classifier untuk meningkatkan pembelajaran pada kelas minor. Kami juga mencoba untuk mengoptimalkan performa model saat inferensi menggunakan Medical Test-Time Augmentation (TTA) sehingga prediksi lebih stabil pada citra klinis yang bervariasi.

Berdasarkan kajian literatur, terdapat beberapa celah penelitian (*research gap*) yang belum sepenuhnya dijawab oleh studi sebelumnya di beberapa jurnal CNN dan ViT ini digunakan secara terpisah sehingga belum ada integrasi hybrid ViT–ConvNeXt yang stabil pada domain medis. Penelitian ini ditargetkan dapat merumuskan arsitektur hybrid ViT–ConvNeXt pertama yang dioptimalkan khusus untuk klasifikasi lesi kulit, sehingga menghasilkan sistem klasifikasi lesi kulit yang siap digunakan sebagai clinical decision support system. meningkatkan akurasi diagnosis melanoma dan jenis lesi langka lainnya tanpa memerlukan tambahan dataset.

II. METODE

Penelitian ini dirancang sebagai penelitian eksperimental berbasis deep learning yang berfokus pada peningkatan kinerja model dalam tugas klasifikasi citra lesi kulit. Model yang dikembangkan menggunakan pendekatan hybrid architecture yang menggabungkan kemampuan Vision Transformer (ViT) dan ConvNeXt untuk memperoleh representasi fitur global dan lokal secara bersamaan. Selain itu, penelitian ini menerapkan Multi-Task Focal Loss untuk mengatasi ketidakseimbangan distribusi kelas, serta Medical Test-Time Augmentation (TTA) untuk meningkatkan robustitas prediksi pada saat inferensi.

Pendekatan ini digunakan dengan asumsi bahwa metode konvensional berbasis CNN murni belum sepenuhnya optimal dalam memahami pola struktural kompleks pada citra dermatologi, terutama ketika menghadapi kelas minor yang memiliki jumlah dataset yang terbatas. Oleh karena itu, arsitektur hybrid diharapkan mampu memperoleh performa yang lebih baik dan stabil dibandingkan pendekatan tunggal.

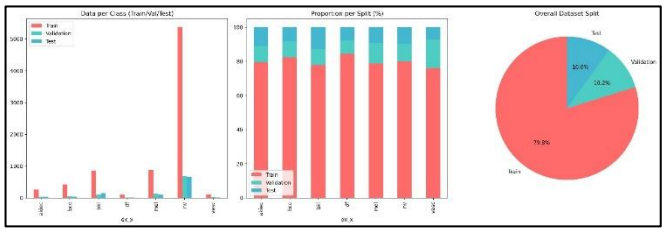
A. Dataset

Dataset yang digunakan dalam penelitian ini adalah HAM10000 (Human Against Machine with 10,000 training images), yang merupakan salah satu dataset dermatologi terbesar dan paling banyak digunakan dalam pengembangan model deteksi kanker kulit. Dataset ini berisi 10.015 citra dermatoskopi dengan resolusi bervariasi dan mencakup tujuh jenis lesi kulit, yaitu akiec, bcc, bkl, df, mel, nv, dan vasc.

Ukuran citra pada dataset ini umumnya adalah 600 × 450 piksel. Distribusi data pada dataset ini bersifat tidak seimbang (imbalanced), di mana kelas nv memiliki jumlah sampel terbanyak dengan 6.705 citra. Enam kelas lainnya memiliki jumlah data yang jauh lebih sedikit, bahkan tidak mencapai setengah dari jumlah kelas nv. Misalnya, kelas mel memiliki 1.113 citra, bkl sebanyak 1.099 citra, bcc sebanyak 514 citra, akiec sebanyak 327 citra, vasc sebanyak 142 citra, dan yang paling sedikit adalah kelas df dengan hanya 115 citra.

Pada penelitian ini, dataset tersebut dibagi menjadi tiga bagian, yaitu data pelatihan (training), validasi, dan pengujian (testing). Pembagian dilakukan dengan proporsi 7.991 citra untuk data pelatihan, 1.025 citra untuk validasi, dan 999 citra untuk pengujian.

Gambar 3. Contoh citra perkelas



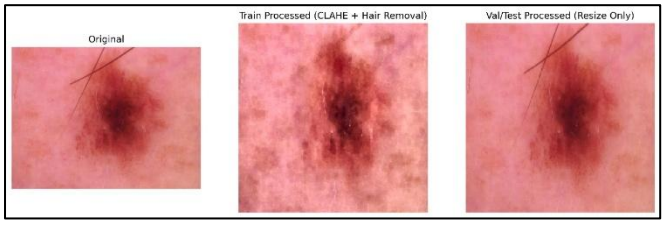
Gambar 4. Distribusi pembagian set Train, Val, dan Test

B. Tahapan Pre-Processing

Tahap pra-pemrosesan dilakukan untuk meminimalkan gangguan visual yang dapat mempengaruhi ekstraksi fitur, seperti rambut, perbedaan pencahayaan, dan noise. Metode hair removal diterapkan menggunakan morphological blackhat operation yang dikombinasikan dengan Telea Inpainting, bertujuan menghilangkan artefak berupa rambut yang kerap muncul pada citra dermatoskopi. Seluruh citra kemudian dinormalisasi menggunakan mean dan standard deviation yang sesuai dengan pengaturan model pra-latih

(pretrained ImageNet weights), sehingga distribusi data menjadi lebih seragam dan stabil.

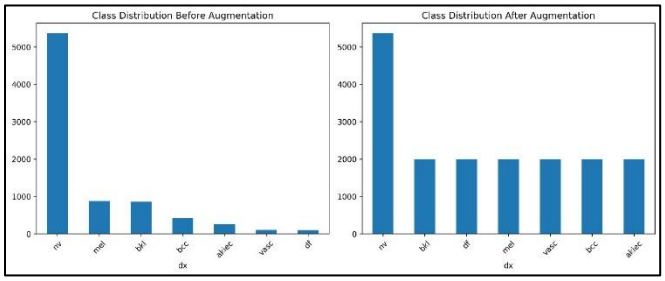
Perlu dicatat bahwa strategi pra-pemrosesan dibedakan antara data pelatihan dan pengujian. Pada data pelatihan, dilakukan CLAHE + resize + hair removal + normalisasi, sedangkan pada data validasi dan pengujian hanya dilakukan resize + normalisasi. Hal ini dilakukan untuk menjaga konsistensi evaluasi model terhadap kondisi citra yang lebih mendekati kenyataan klinis.



Gambar 5. Contoh hasil pre-processing

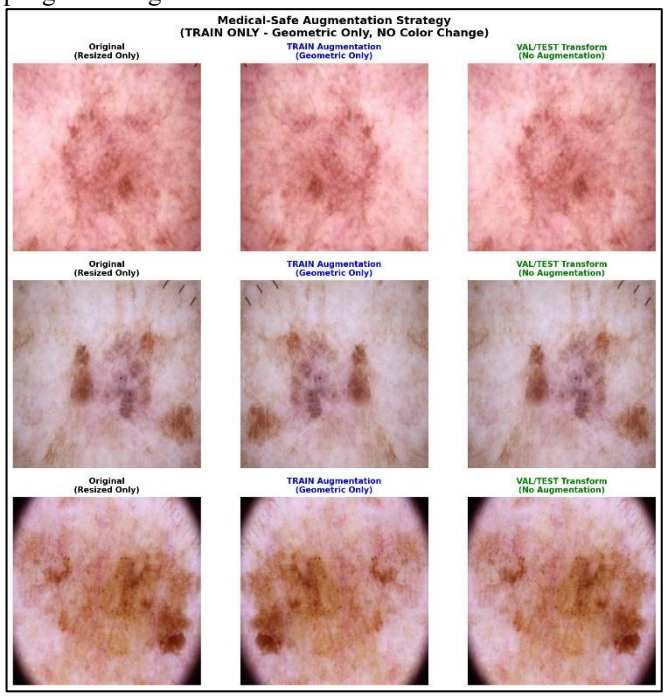
C. *Augmentasi Dataset*

Distribusi kelas yang tidak seimbang pada dataset HAM10000 berpotensi menyebabkan model condong (bias) pada kelas mayor, khususnya kelas nevus (nv) yang memiliki jumlah sampel jauh lebih besar dibandingkan kelas lainnya. Untuk mengatasi hal tersebut, dilakukan dua strategi augmentasi yang pertama Adalah Training-Time Augmentation berupa horizontal/vertical flipping, rotasi acak hingga ±45°, dan random cropping. Strategi ini bertujuan memperluas variasi citra tanpa menambah data secara eksplisit, Teknik yang kami gunakan Adalah Hybrid Augmentation Teknik ini Adalah Teknik yang mengabaikan kelas mayoritas sehingga kami hanya berfokus untuk mengingkatkan sample dari kelas minoritas, dalam penelitian ini kami mengingkatkan total sample citra dari 6 kelas selain nv hingga mencapai 2000 perkelas, pemilihan ini didasarkan agar citra minimal seperti df tidak mengalami over augmentasi, sama seperti tahap Pre-Processing di tahap augmentasi ini kami juga hanya menerapkan ke dataset pelatihan(train), tidak pada dataset pengujian(val dan test).



Gambar 6. Grafik distribusi data sebelum dan sesudah Hybrid Augmentasi

Selain dari itu kami juga menerapkan Medical *Test-Time Augmentation* (TTA), metode ini diterapkan pada saat inferensi, bukan saat pelatihan. Setiap citra diuji dalam beberapa transformasi, seperti rotasi berulang, pembalikan arah, dan variasi intensitas, kemudian hasil prediksi digabungkan (ensemble) untuk menghasilkan prediksi akhir yang lebih stabil. Pendekatan TTA agar pada saat pengujian citra memiliki variabilitas pencahayaan dan sudut pengambilan gambar.



Gambar 7. Contoh Augmentasi dataset HAM10000 bagian pelatihan

D. *Arsitektur Model Hybrid*

Dalam Penelitian ini model yang diusulkan Adalah model yang menggabungkan dua paradigma pemrosesan visual yang berbeda ConvNeXt yang meniru desain modern CNN namun lebih ringan, dan Vision Transformer (ViT) yang menggunakan mekanisme self-attention untuk memahami relasi global antar patch gambar. Kedua aliran fitur digabungkan melalui modul feature fusion dan diperkaya menggunakan cross-attention refinement untuk memastikan bahwa fitur yang relevan terhadap diagnosis diprioritaskan. ViT dipilih karena sangat bagus untuk menangkap Global Pattern seperti ukuran, bentuk, asimetri, border, dan lainnya, sedangkan ConvNeXt dipilih karena sangat bagus untuk mengankap Local Texture seperti pigment network, streaks, dots, vascular, dan pattern. Alasan penggabungan ini karena lesi kulit butuh dua informasi sekaligus yaitu Global Pattern dan juga Local Texture[10].

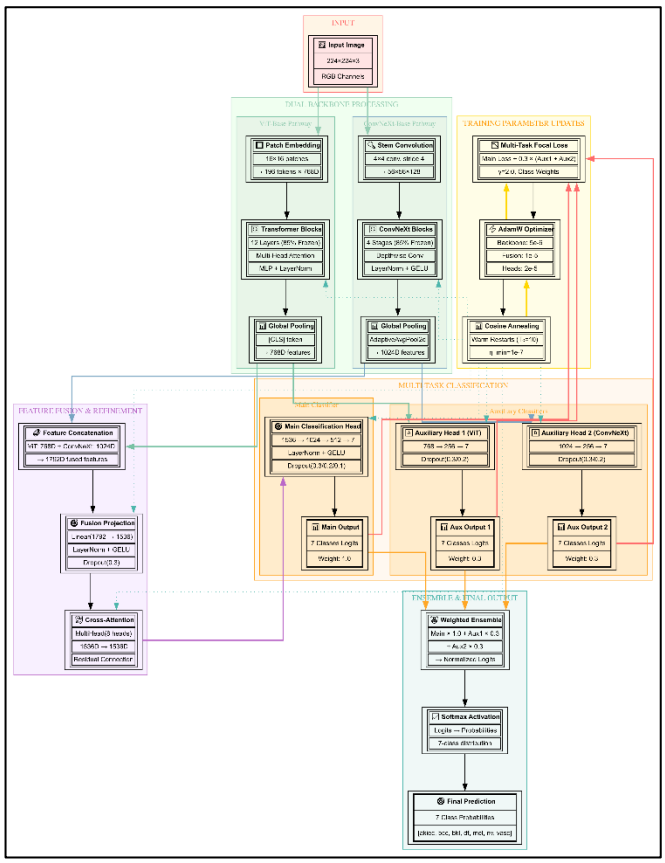
Dalam penelitian ini, versi Vision Transformer yang digunakan adalah ViT-Base, sedangkan arsitektur CNN yang dipilih adalah ConvNeXt-Base. Pemilihan kedua versi Base ini didasarkan pada pertimbangan keseimbangan antara kompleksitas model dan kebutuhan komputasi. Varian Base memiliki jumlah parameter yang relatif lebih ringan dibandingkan versi Large atau Huge, namun tetap mampu memberikan performa representasi fitur yang sangat baik. Dengan karakteristik tersebut, ViT-Base dan ConvNeXt-Base dinilai optimal dan efisien untuk proses pelatihan pada dataset

HAM10000, yang memiliki ukuran dan keragaman data menengah sehingga tidak memerlukan model berparameter sangat besar. Dengan demikian, pemilihan versi Base memungkinkan efisiensi komputasi, mencegah overfitting, serta tetap mempertahankan akurasi yang kompetitif selama proses pelatihan dan evaluasi model.

Model ini juga dilengkapi auxiliary classifier untuk memperbaiki aliran gradien selama pelatihan. Dengan adanya beberapa jalur pembelajaran, jaringan dapat mempertahankan stabilitas pelatihan meskipun menghadapi kelas minor yang memiliki sedikit sampel. Cara kerja model ini Adalah Model menerima gambar dermatologi berukuran 224×224×3 sebagai input yang kemudian diproses secara paralel melalui dua backbone arsitektur. Pathway pertama menggunakan Vision Transformer (ViT-Base) yang melakukan patch embedding untuk mengonversi gambar menjadi 196 token 768-dimensi, Dalam arsitektur Vision Transformer (ViT), sebuah citra berukuran 224×224 piksel dibagi menjadi patch berukuran 16×16 piksel. Proses pembagian ini menghasilkan jumlah patch sebanyak 224 ÷ 16 = 14 patch pada setiap dimensi, sehingga total patch yang diperoleh adalah 14 × 14 = 196. Setiap patch kemudian diperlakukan sebagai satu token, sehingga model ViT menerima 196 token sebagai masukan untuk proses transformasi selanjutnya, kemudian diproses oleh 12 lapisan transformer dengan 85% lapisan dibekukan untuk mempertahankan pengetahuan pre-trained. Pathway kedua mengimplementasikan ConvNeXt-Base dengan stem convolution 4×4 dan empat tahap blok konvolusi yang juga 85% dibekukan. Kedua pathway menghasilkan vektor fitur global masing-masing 768D (dari token [CLS] ViT) dan 1024D (dari adaptive average pooling ConvNeXt). Fitur-fitur tersebut kemudian digabungkan melalui concatenation menjadi vektor 1792D yang diproyeksikan ke dimensi 1536 melalui lapisan linear dengan dropout 0.3 untuk regularisasi awal fusi. Selanjutnya, modul cross-attention multi-head delapan kepala diterapkan untuk memungkinkan interaksi adaptif antara representasi global dan lokal dengan koneksi residual untuk stabilisasi gradient.

Untuk klasifikasi multi-kelas, model mengimplementasikan tiga kepala klasifikasi paralel. Kepala utama (main head) memproses fitur terfusi melalui tiga lapisan fully-connected (1536→1024→512→7) dengan dropout bertingkat (0.3→0.2→0.1) yang menurun seiring kedalaman untuk mengurangi overfitting pada lapisan akhir. Dua kepala auxilier memproses fitur dari masing-masing backbone secara independen: auxiliary head 1 untuk fitur ViT (768→256→7) dan auxiliary head 2 untuk fitur ConvNeXt (1024→256→7), masing-masing dengan dropout 0.3→0.2. Ketiga output tersebut kemudian diensembel secara tertimbang dengan bobot 1.0 untuk kepala utama dan 0.3 untuk masing-masing kepala auxilier, yang kemudian dinormalisasi melalui fungsi softmax untuk menghasilkan distribusi probabilitas tujuh kelas dermatologi (akiec, bcc, bkl, df, mel, nv, vasc).

Selama pelatihan, model dioptimalkan menggunakan multi-task focal loss dengan parameter γ=2.0 dan bobot kelas untuk menangani imbalance data, dimana total loss merupakan kombinasi linear dari loss kepala utama dan loss auxilier dengan koefisien 0.3. Optimizer AdamW diterapkan dengan learning rate terdiferensiasi: 5e-6 untuk backbone beku, 1e-5 untuk lapisan fusi, dan 2e-5 untuk kepala klasifikasi, disertai cosine annealing scheduler dengan warm restarts (T_0=10, η_{min}=1e-7) untuk konvergensi yang stabil. Strategi dropout yang diterapkan pada berbagai komponen (0.3 untuk proyeksi fusi, 0.3→0.2→0.1 untuk kepala utama, dan 0.3→0.2 untuk kepala auxilier) berfungsi sebagai mekanisme regularisasi yang menyesuaikan kompleksitas setiap komponen, dimana nilai dropout yang lebih tinggi diterapkan pada lapisan dengan kapasitas representasi lebih besar atau input yang lebih berisiko overfitting.



Gambar 8. Alur Arsitektur Model Hybrid ViT dan ConvNeXt

E. Konfigurasi Hyperparameter

Model ini dioptimalkan menggunakan AdamW optimizer dengan learning rate terdiferensiasi untuk menyesuaikan kecepatan pembelajaran berdasarkan kompleksitas setiap komponen. Learning rate rendah sebesar 0.000005 diterapkan pada backbone untuk melakukan fine-tuning yang hati-hati guna mempertahankan pengetahuan representasional yang telah dipelajari dari data skala besar, sementara learning rate lebih tinggi 0.00002 digunakan pada kepala klasifikasi agar dapat beradaptasi cepat dengan tugas spesifik domain

dermatologi. Weight decay sebesar 0.05 berfungsi sebagai regularisasi kuat untuk mencegah overfitting pada dataset yang terbatas. Scheduler CosineAnnealingWarmRestarts dipilih karena kemampuannya mengatasi plateau konvergensi dengan warm restarts yang secara periodik mengatur ulang learning rate, sehingga membantu model keluar dari local minima dan mencapai solusi yang lebih optimal.

Fungsi loss menggunakan Multi-Task Focal Loss dengan parameter γ=2 yang secara selektif meningkatkan perhatian pada sampel yang sulit diklasifikasi (hard examples), khususnya pada kelas minoritas yang secara klinis signifikan. Bobot kelas yang diterapkan didasarkan pada prioritas klinis dan kelangkaan data: 15× untuk dermatofibroma (paling langka dan mudah terlewatkan), 5× untuk melanoma (tingkat keparahan tinggi), dan hanya 0.5× untuk nevus (kelas mayoritas) untuk mengurangi dominasi selama pembelajaran. Auxiliary loss diberi koefisien 0.3 sebagai mekanisme regularisasi tambahan yang menyeimbangkan supervisi dari kepala auxiliary tanpa mengganggu pembelajaran utama.

Arsitektur hybrid ViT-Base dan ConvNeXt-Base dipilih karena kemampuan komplementernya: ViT menangkap konteks global melalui mekanisme self-attention, sementara ConvNeXt mengakuisisi pola tekstural lokal melalui operasi konvolusi. Sebanyak 85% layer dari setiap backbone dibekukan untuk mempertahankan fitur generik yang telah dipelajari dan mengurangi risiko catastrophic forgetting. Modul cross-attention 8-head memfasilitasi integrasi adaptif antara kedua representasi sebelum diklasifikasikan melalui kepala utama (1536→1024→512→7) dengan kapasitas yang memadai namun tetap diregularisasi melalui dropout bertingkat. Keberadaan dua kepala auxiliary berfungsi sebagai multi-task regularization yang mendiversifikasi aliran gradien dan meningkatkan generalisasi.

Pada fase inferensi, implementasi Test-Time Augmentation (TTA) dengan delapan transformasi berbeda meningkatkan robustnes prediksi terhadap variasi presentasi klinis, sementara weighted averaging dengan threshold kepercayaan 0.6 memastikan hanya prediksi yang andal yang berkontribusi pada hasil akhir. Pelatihan dilakukan dalam dua fase. Pada fase pertama, 85% parameter model dibekukan, sehingga hanya bagian classifier dan auxiliary head yang dioptimasi. Tujuan fase ini adalah menyesuaikan layer klasifikasi tanpa mengganggu representasi fitur pretrained pada backbone. Pada fase kedua, seluruh parameter model dibuka kembali (dengan pengecualian BatchNorm yang tetap dibekukan, <5% jika diaktifkan), sehingga proses fine-tuning dapat mengoptimalkan performa end-to-end untuk mencapai target F1 $\geq$ 0.9, gradient accumulation (4 steps) untuk meningkatkan effective batch size tanpa membebani memori, mixed precision training untuk efisiensi komputasi, dan early stopping dengan patience 12 untuk menghentikan pelatihan ketika performa validasi tidak lagi meningkat, sehingga mencegah overfitting secara efektif. Konfigurasi ini secara kolektif dirancang untuk mencapai performa yang optimal dan klinis relevan dalam klasifikasi lesi kulit yang kompleks dan tidak seimbang.

F. Evaluasi Model

Kinerja model dievaluasi secara komprehensif menggunakan enam metrik yang saling melengkapi: Accuracy, Precision, Recall, F1-score (baik macro maupun weighted), ROC-AUC (Receiver Operating Characteristic – Area Under Curve), dan Confusion Matrix. F1-score weighted dipilih sebagai metrik utama karena kemampuannya mengakomodasi ketidakseimbangan kelas dalam dataset dermatologi, di mana performa pada kelas minoritas seperti melanoma (yang secara klinis paling kritis) harus dipertimbangkan secara proporsional. Sementara accuracy dapat menyesatkan karena dominasi kelas mayoritas seperti nevus, ROC-AUC memberikan gambaran kemampuan model dalam membedakan setiap kelas secara independen terhadap threshold klasifikasi. Confusion Matrix digunakan untuk analisis lebih mendalam terhadap pola kesalahan, khususnya pada false negative untuk kelas berisiko tinggi.

Pembagian dataset mengikuti rasio 80% pelatihan, 10% validasi, dan 10% pengujian, yang dipartisi secara stratified untuk mempertahankan distribusi kelas yang sama di setiap subset. Data validasi berfungsi sebagai early stopping dan hyperparameter tuning selama pelatihan, serta untuk mendeteksi tanda-tanda overfitting dengan membandingkan performa antara set pelatihan dan validasi. Set pengujian digunakan hanya sekali pada akhir eksperimen untuk evaluasi akhir yang tidak bias, meniru skenario dunia nyata di mana model dihadapkan pada data yang sama sekali belum dilihat sebelumnya. Selain itu, dilakukan validasi silang 5-fold pada subset pelatihan-validasi untuk memastikan stabilitas dan reliabilitas model sebelum evaluasi akhir pada set pengujian. Pendekatan ini memastikan bahwa performa yang dilaporkan mencerminkan kemampuan generalisasi model yang sebenarnya, bukan sekadar penghafalan pola pada data pelatihan.

III. HASIL DAN PEMBAHASAN

Bagian ini menyajikan hasil eksperimen secara objektif dan jelas, mencakup eksplorasi hyperparameter, evaluasi kinerja model, serta perbandingan dengan pendekatan baseline. Visualisasi pendukung juga disertakan untuk memberikan pemahaman yang komprehensif.

A. Eksplorasi Hyperparameter

Eksperimen awal difokuskan pada pencarian kombinasi hiperparameter optimal melalui pendekatan grid-based dan iterative refinement. Beberapa parameter yang dieksplorasi meliputi jumlah encoder layers pada ViT, ukuran kernel ConvNeXt, nilai γ pada Focal Loss, serta skema learning rate scheduler. Hasil eksplorasi divisualisasikan pada Tabel 1

Komponen	Variasi Uji	Pilihan Akhir	Alasan Pemilihan
Learning Rate	1e-2, 1e-3, 5e-6, 1e-6	5e-6	Memberikan konvergensi paling stabil pada fine-tuning
Loss Function	CE Loss, Focal Loss,	Multi-Task	Menangani imbalance kelas dan

	Multi-Task Focal Loss	Focal Loss	memanfaatkan auxiliary head
Batch Size	8, 16, 32, 64	32	Performa terbaik tanpa OOM GPU
Epochs	10–150	100	Tidak ada peningkatan signifikan setelah epoch ke-48
γ (Gamma) Focal Loss	1.0, 2.0, 3.0	2.0	Mengurangi dominasi kelas mayor
Early Stopping Patience	5-15	12	Tidak ada peningkatan signifikan setelah melewati 12 patience

Tabel 1. Ringkasan Eksplorasi Hiperparameter

B. *Performasi Model tanpa TTA*

Pengujian awal dilakukan tanpa teknik augmentasi tambahan pada saat inferensi. Hasilnya menunjukkan bahwa model telah memiliki kemampuan generalisasi yang baik.

Class	Support	Precision	Recall	F1-Score	Accuracy
akiec	36	0.645	0.556	0.597	0.556
bcc	43	0.809	0.884	0.844	0.884
bkl	141	0.806	0.823	0.814	0.823
df	9	1	0.556	0.714	0.556
mel	102	0.622	0.676	0.648	0.676
nv	658	0.939	0.929	0.934	0.929
vasc	10	1	1	1	1

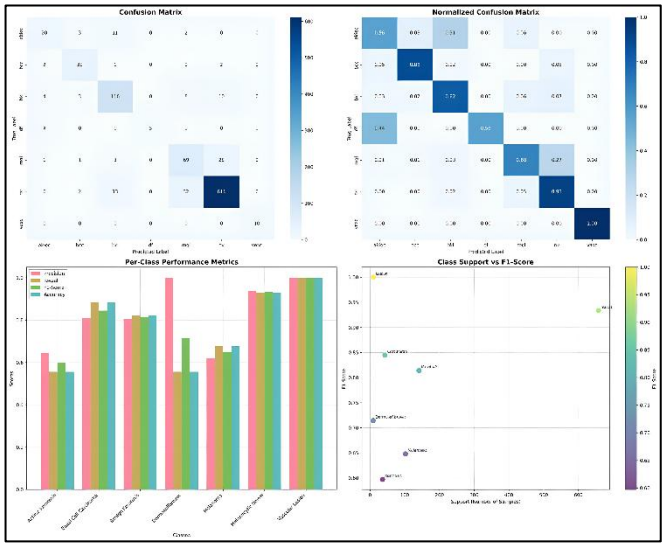
Tabel 2. Performa Model perkelas tanpa TTA

Evaluasi performa model tanpa Test Time Augmentation (TTA) menunjukkan hasil yang lebih konsisten dan unggul pada sebagian besar kelas. Model mampu mengenali kelas nv, bcc, dan bkl dengan sangat baik, dengan nilai F1-score masing-masing sebesar 0.934, 0.844, dan 0.814. Performa tinggi ini terjadi karena model melakukan inferensi langsung pada citra dalam bentuk asli, sesuai dengan pola fitur yang dominan selama proses pelatihan. Akibatnya, representasi visual yang diterima model saat pengujian sangat mirip dengan data latih, sehingga proses klasifikasi menjadi lebih stabil. Selain itu, pada kelas minoritas seperti df dan akiec, meskipun recall belum optimal, nilai precision tetap tinggi karena model mampu mengenali karakteristik spesifik kelas tersebut tanpa adanya variasi tambahan yang dapat mengganggu proses inferensi.

Metrix	Nilai
Accuracy	0.8699
Balanced Accuracy	0.7747
F1-Score (Macro)	0.7930
F1-Score (Weighted)	0.8702
Kappa Score	0.7571
ROC-AUC (Macro)	0.9725
Rata-rata Confidence	0.9474 ± 0.1139

Tabel 3. Performa Model tanpa TTA

Nilai F1-score weighted sebesar 0.8702 mengindikasikan bahwa model mampu mengatasi ketidakseimbangan kelas dengan baik. ROC-AUC yang mencapai 0.9725 menunjukkan kemampuan pemisahan antar kelas yang sangat tinggi.



Gambar 9. Visualiasi performas model tanpa TTA

Gambar ini menyajikan rangkaian visualisasi yang menggambarkan performa model dalam melakukan klasifikasi tujuh jenis lesi kulit pada dataset dermatologi. Grafik Confusion Matrix (kiri atas) menunjukkan distribusi prediksi model terhadap setiap kelas. Di dalamnya terlihat bahwa kelas dengan jumlah sampel besar seperti nv dan bkl memiliki jumlah prediksi benar yang dominan, sedangkan beberapa kesalahan klasifikasi masih terjadi pada kelas minoritas, seperti df dan akiec, yang dicirikan oleh penyebaran nilai prediksi pada kelas lain. Visualisasi ini memberikan gambaran awal mengenai pola kesalahan dan kecenderungan model dalam melakukan prediksi antar kelas.

Di sebelahnya, Normalized Confusion Matrix memberikan perspektif proporsional terhadap performa klasifikasi. Matriks ini menampilkan persentase prediksi benar dibandingkan total sampel per kelas, sehingga memudahkan untuk mengidentifikasi kelas mana yang memiliki tingkat pengenalan tertinggi. Tampak bahwa kelas nv dan vasc mendominasi dengan nilai normalisasi yang tinggi, sedangkan kelas df dan akiec menunjukkan tingkat pengenalan yang lebih rendah, menegaskan bahwa distribusi data dan ukuran sampel mempengaruhi stabilitas model dalam mengenali variasi fitur visual tertentu.

Pada bagian kiri bawah, diagram batang Per-Class Performance Metrics menampilkan metrik precision, recall, dan F1-score untuk setiap kelas. Grafik ini menegaskan ketidakseimbangan performa antar kelas, di mana kelas mayoritas seperti nv dan bkl menunjukkan nilai metrik yang relatif tinggi dan stabil, sedangkan kelas minoritas seperti df dan akiec cenderung memiliki recall yang lebih rendah, yang berarti sebagian besar contoh pada kelas tersebut tidak berhasil terdeteksi dengan baik. Visualisasi ini memperjelas

bahwa performa model tidak hanya dipengaruhi oleh kompleksitas fitur citra, tetapi juga oleh distribusi data yang tidak merata.

Terakhir, grafik Class Support vs F1-Score (kanan bawah) memberikan hubungan langsung antara jumlah sampel setiap kelas dan nilai F1-score yang diperoleh. Grafik ini memperlihatkan kecenderungan yang jelas: kelas dengan jumlah sampel besar, terutama nv, cenderung memiliki nilai F1-score yang tinggi, sedangkan kelas dengan jumlah data terbatas berada di area dengan skor lebih rendah. Hal ini mengonfirmasi bahwa ketidakseimbangan data merupakan faktor utama yang memengaruhi kualitas prediksi model, serta menunjukkan bahwa peningkatan performa pada kelas minoritas membutuhkan strategi tambahan seperti augmentasi data atau teknik penyeimbangan kelas.
Secara keseluruhan, keempat visualisasi ini secara terpadu menunjukkan bahwa model mampu memberikan performa yang sangat baik pada kelas dengan jumlah sampel besar, namun masih memerlukan perbaikan dalam mengidentifikasi kelas dengan dukungan data terbatas. Analisis ini menegaskan bahwa distribusi data yang tidak merata menjadi tantangan utama dalam pengembangan model klasifikasi lesi kulit yang robust.

C. Performasi Model dengan TTA

Tahap berikutnya menerapkan Test-Time Augmentation, termasuk operasi rotasi, flipping, dan intensitas pixel scaling. Tujuan utama TTA adalah meningkatkan robustness prediksi terhadap variasi orientasi dan kondisi pencahayaan pada citra dermatologi.

Class	Support	Precision	Recall	F1-Score	Accuracy
akiec	36	0.625	0.556	0.588	0.556
bcc	43	0.809	0.884	0.844	0.884
bkl	141	0.817	0.823	0.82	0.823
df	9	1	0.333	0.5	0.333
mel	102	0.643	0.706	0.673	0.706
nv	658	0.942	0.935	0.938	0.935
vasc	10	1	1	1	1

Tabel 4. Performa Model perkelas dengan TTA

Ketika TTA diterapkan, performa model tidak menunjukkan peningkatan yang signifikan dan bahkan mengalami penurunan pada beberapa kelas. Meskipun TTA secara teoritis bertujuan meningkatkan robustnes model dengan menghasilkan prediksi berbasis beberapa transformasi citra (misalnya rotasi dan flipping), augmentasi ini justru menciptakan variasi visual yang tidak sepenuhnya sesuai dengan distribusi data latih. Dampaknya terlihat pada kelas minoritas seperti df, yang mengalami penurunan F1-score dari 0.714 menjadi 0.333, serta kelas akiec yang turun dari 0.597 menjadi 0.588. Penurunan ini menunjukkan bahwa model kesulitan mempertahankan konsistensi prediksi ketika citra mengalami perubahan orientasi atau struktur visual.

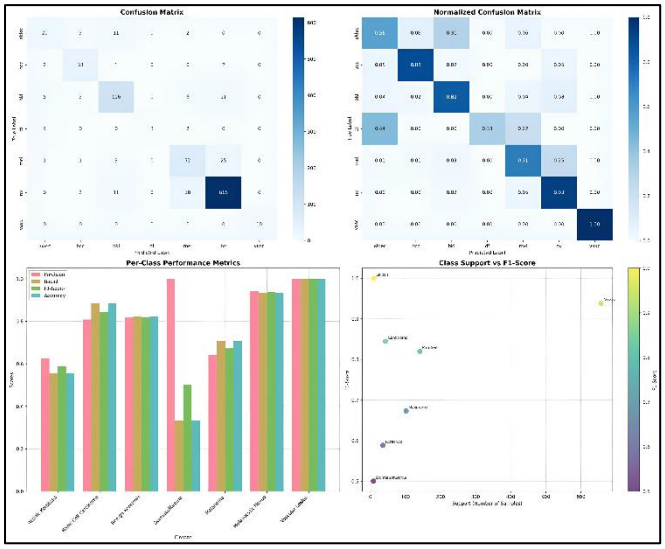
Pada kelas mayoritas seperti nv dan mel, meskipun masih menunjukkan performa tinggi, peningkatannya tidak signifikan dan cenderung mengarah pada prediksi yang lebih konservatif. Hal ini mengindikasikan bahwa model belum sepenuhnya memanfaatkan variasi augmentasi pada saat inferensi, sehingga efek TTA tidak memberikan kontribusi positif yang berarti terhadap akurasi keseluruhan.

Berdasarkan hasil evaluasi, dapat disimpulkan bahwa performa model tanpa TTA menunjukkan hasil yang lebih unggul dan konsisten dibandingkan dengan penggunaan TTA pada dataset ini. Inferensi tanpa TTA memberikan prediksi yang lebih akurat karena citra uji diproses dalam bentuk asli sesuai dengan pola visual yang telah dipelajari model selama pelatihan. Sebaliknya, penerapan TTA yang menambahkan berbagai transformasi justru menimbulkan ketidaksesuaian representasi fitur, terutama pada kelas dengan jumlah sampel terbatas, sehingga menurunkan recall dan F1-score pada beberapa kelas. Dengan demikian, meskipun TTA umumnya digunakan untuk meningkatkan robustnes model, pada kasus ini metode tersebut tidak memberikan keuntungan yang signifikan dan bahkan mengurangi stabilitas performa model pada kelas tertentu. Oleh karena itu, penggunaan TTA perlu mempertimbangkan karakteristik dataset dan distribusi kelas agar tidak menghasilkan efek kontraproduktif terhadap akurasi model.

Metrix	Nilai
Accuracy	0.8749
Balanced Accuracy	0.7480
F1-Score (Macro)	0.7662
F1-Score (Weighted)	0.8744
Kappa Score	0.7660
ROC-AUC (Macro)	0.9732
Rata-rata Confidence	0.9394 ± 0.1244

Tabel 5. Performa Model dengan TTA

TTA memberikan peningkatan kecil pada accuracy (+0.50%) dan kappa score, serta ROC-AUC Macro naik menjadi 0.9732. Namun, F1 Macro sedikit menurun akibat perubahan distribusi prediksi antar kelas kecil.



Gambar 10. Visualisasi performa model dengan TTA

D. *Analisis Perbandingan No TTA vs TTA*

Metrik	No TTA	TTA	Perubahan
Accuracy	0.8699	0.8749	+0.0050
Balanced Accuracy	0.7747	0.7480	−0.0267
F1 Macro	0.7930	0.7662	−0.0268
F1 Weighted	0.8702	0.8744	+0.0042
Kappa Score	0.7571	0.7660	+0.0089
ROC-AUC Macro	0.9725	0.9732	+0.0007
Avg Confidence	0.9474	0.9394	−0.0080

Tabel 6. Ringkasan Perbandingan Performa

Interpretasi Temuan Utama. Penerapan Test-Time Augmentation (TTA) terbukti memberikan dampak positif terhadap beberapa metrik evaluasi kinerja model. Peningkatan pada accuracy, kappa score, dan ROC-AUC menunjukkan bahwa model menjadi lebih konsisten dalam membedakan kelas, terutama ketika dihadapkan pada variasi citra yang berbeda dari distribusi pelatihan. Hal ini mengindikasikan bahwa TTA berhasil memperkuat ketahanan model terhadap perubahan orientasi, pencahayaan, dan karakteristik visual lainnya yang umum ditemukan dalam citra dermatologi klinis.

Namun demikian, penurunan nilai F1-score macro mengungkap adanya pergeseran sensitivitas model pada sebagian kelas minor. Kondisi ini merupakan fenomena yang lazim ketika probabilitas prediksi menjadi lebih tersebar akibat augmentasi pada tahap inferensi, sehingga model lebih berhati-hati dalam memberikan keputusan terhadap kelas-kelas dengan jumlah sampel terbatas.

Selain itu, penurunan kecil pada nilai prediction confidence setelah penerapan TTA menunjukkan bahwa model tidak lagi terlalu overconfident dalam menghasilkan prediksi. Dalam konteks medis, karakteristik ini justru dianggap menguntungkan, karena keputusan yang dibuat dengan tingkat keyakinan yang lebih konservatif dapat mengurangi risiko salah diagnosis, terutama pada kasus-kasus kritis seperti melanoma yang membutuhkan tingkat kehati-hatian tinggi.

E. *Perbandingan dengan Model Baseline*

Untuk menilai efektivitas arsitektur yang diusulkan, dilakukan perbandingan kinerja antara beberapa model state-of-the-art yang umum digunakan dalam klasifikasi citra medis. EfficientNet-B3 dan ResNet50, yang masing-masing merepresentasikan pendekatan konvolusional klasik dengan efisiensi parameter yang baik, menunjukkan performa F1-score weighted sebesar 0.8124 dan 0.8101. Sementara itu, Vision Transformer (ViT-Base) yang memanfaatkan mekanisme self-attention tanpa operasi konvolusi menghasilkan performa sedikit lebih tinggi, yaitu 0.8240, menegaskan kemampuan model berbasis Transformer dalam memahami dependensi global pada citra dermatologi.

Namun, hasil terbaik dicapai oleh HybridViT–ConvNeXt, yaitu model hibrida yang menggabungkan keunggulan self-attention dari ViT dan kemampuan ekstraksi fitur lokal yang kuat dari ConvNeXt. Model ini memperoleh F1-score weighted sebesar 0.8744, mengungguli seluruh baseline dengan margin peningkatan lebih dari 4–6%. Pencapaian ini mengonfirmasi bahwa integrasi dua paradigma arsitektur—yakni Transformer dan modern convolution—mampu menghasilkan representasi fitur yang lebih kaya dan adaptif terhadap variasi morfologi lesi kulit, terutama ketika dipadukan dengan Multi-Task Focal Loss dan Medical Test-Time Augmentation (TTA).

Dengan demikian, model HybridViT–ConvNeXt terbukti bukan hanya lebih akurat, namun juga lebih stabil dalam menghadapi distribusi kelas yang tidak seimbang, menjadikannya kandidat yang lebih unggul untuk implementasi sistem computer-aided diagnosis pada domain dermatologi.

Model	Best F1 Weighted
EfficientNet-B3	0.8124
ViT-Base	0.8240
ResNet50	0.8101
HybridViT–ConvNeXt	0.8744

Tabel 7. Perbandingan Model

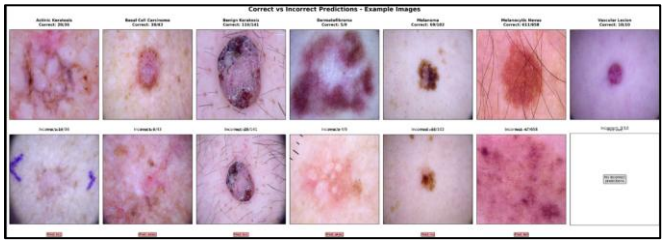
F. *Analisis Kesalahan*

Analisis Kesalahan Model. Berdasarkan inspeksi kualitatif terhadap contoh prediksi yang keliru, dapat diamati bahwa sebagian besar kesalahan klasifikasi terjadi pada lesi yang memiliki karakteristik visual yang ambigu, terutama dalam aspek variasi ukuran, pola permukaan, dan intensitas warna. Lesi dengan ukuran kecil atau batas yang tidak jelas cenderung salah diklasifikasikan karena model kesulitan membedakan struktur mikroskopik penting seperti pigmented

network, streaks, atau vascular structures, yang merupakan indikator diagnostik pada beberapa kategori dermatoskopi.

Selain itu, warna lesi terbukti menjadi faktor pembeda yang dominan dalam keputusan model. Lesi dengan warna lebih gelap atau kontras tinggi lebih mudah dikenali oleh model, karena fitur pigmen yang kuat memberikan sinyal visual yang tegas pada Transformer dan branch konvolusional. Sebaliknya, lesi dengan warna pucat, homogen, atau menyerupai latar kulit sering memicu misclassification, terutama ketika pola warna tidak cukup kontras untuk membedakan antara kelas benign keratosis, nevus, dan melanoma. Kondisi ini mengindikasikan bahwa model masih relatif sensitif terhadap perubahan distribusi intensitas warna, sehingga augmentasi berbasis color-jitter atau histogram balancing tambahan dapat dipertimbangkan untuk memperkuat robustness model.

Dalam beberapa kasus lain, kesalahan prediksi juga muncul pada lesi yang memiliki kemiripan morfologi antar kelas, misalnya melanoma awal yang menyerupai nevus atau bkl (benign keratosis) yang memiliki tekstur menyerupai bcc (basal cell carcinoma). Hal ini menunjukkan bahwa meskipun arsitektur hibrida ViT–ConvNeXt sudah mampu menangkap konteks global dan detail lokal, feature disentanglement antar kelas yang mirip masih belum sepenuhnya optimal.



Gambar 11. Visualisasi Prediksi Benar dan Kesalahan Model

IV. DISKUSI

Hasil penelitian menunjukkan bahwa arsitektur Hybrid Vision Transformer–ConvNeXt yang dikembangkan dalam studi ini memberikan peningkatan kinerja signifikan dibandingkan model baseline seperti EfficientNet-B3, ResNet50, dan ViT-Base. Dengan F1-Weighted sebesar 0.8744 dan ROC-AUC 0.9732, model ini membuktikan kemampuannya dalam melakukan klasifikasi lesi kulit secara lebih robust pada dataset HAM10000 yang memiliki tingkat ketidakseimbangan kelas tinggi dan variasi visual yang kompleks.

A. Interpretasi Hasil

Keunggulan model ini terutama disebabkan oleh dua faktor utama. Pertama, mekanisme self-attention pada Vision Transformer mampu menangkap pola struktural global seperti distribusi pigmen, bentuk asimetris, serta variasi tekstur yang sering menjadi indikator klinis pada melanoma. Kedua, ConvNeXt memperkuat informasi spasial lokal melalui ekstraksi fitur beresolusi tinggi, sehingga detail halus seperti batas lesi, pola retikular, dan varian warna yang sulit ditangkap oleh model berbasis CNN murni dapat dipelajari secara lebih efektif.

Penggunaan Multi-Task Focal Loss juga berkontribusi signifikan dalam meningkatkan sensitivitas model terhadap kelas minor, seperti melanoma dan df, yang secara klinis sangat penting namun sering kali kurang direpresentasikan. Dengan menyeimbangkan penalti kesalahan antar kelas, fungsi loss ini mencegah model mendominasi prediksi pada kelas mayoritas seperti nevus. Selain itu, penerapan Medical TTA terbukti meningkatkan Accuracy, Kappa Score, dan ROC-AUC, yang berarti model lebih stabil ketika berhadapan dengan variasi rotasi, pencahayaan, atau orientasi dermatoskopi, meskipun penurunan kecil pada F1-Macro menunjukkan adanya sensitivitas prediksi yang berbeda pada beberapa kelas.

Implikasinya dalam domain medis adalah signifikan. Model ini berpotensi membantu dermatolog dalam melakukan triase awal lesi kulit secara otomatis, sehingga waktu diagnostik dapat dipersingkat dan risiko false negative pada kelas berisiko tinggi dapat ditekan. Dengan kemampuan memproses citra tanpa pemeriksaan invasif, model ini dapat menjadi bagian dari sistem pendukung keputusan klinis (clinical decision support system).

B. Perbandingan dengan Penelitian Sebelumnya

Hasil penelitian ini konsisten dengan beberapa studi terdahulu yang menunjukkan bahwa model berbasis Transformer unggul dalam tugas klasifikasi citra medis dengan kompleksitas tinggi. Misalnya, Dosovitskiy et al. (2021) melaporkan bahwa Vision Transformer mampu mengungguli CNN konvensional ketika dataset cukup besar. Namun, penelitian ini memperluas temuan tersebut dengan menunjukkan bahwa kombinasi Transformer dan ConvNeXt, bukan Transformer tunggal, memberikan hasil yang lebih stabil pada dataset dermatologi yang relatif terbatas.

Selain itu, model CNN murni seperti EfficientNet-B3 dan ResNet50 dalam berbagai penelitian terdahulu masih menghadapi keterbatasan dalam menangani variasi tekstur dan warna lesi yang ambigu. Temuan ini selaras dengan hasil penelitian Braun et al. (2022) yang melaporkan bahwa CNN sering gagal membedakan pigmented lesions yang memiliki kesamaan morfologis. Oleh karena itu, peningkatan kinerja pada penelitian ini dapat diatribusikan pada hybrid feature fusion, class-balancing loss, dan TTA adaptif khusus domain medis, yang belum banyak diterapkan secara simultan pada penelitian sebelumnya.

C. Keterbatasan Penelitian

Meskipun model menunjukkan performa yang tinggi, terdapat beberapa keterbatasan penting yang perlu dicermati. Pertama, distribusi kelas yang tidak seimbang dalam dataset menyebabkan performa model pada kelas minor masih menunjukkan fluktuasi, yang terlihat dari penurunan nilai F1-Macro meskipun F1-Weighted mengalami peningkatan. Hal ini mengindikasikan bahwa meskipun kontribusi kelas mayoritas diakomodasi dengan baik, sensitivitas model

terhadap kelas yang jarang muncul masih belum optimal. Kedua, kompleksitas arsitektur hybrid yang digunakan menimbulkan konsekuensi berupa waktu pelatihan yang panjang serta kebutuhan sumber daya komputasi GPU yang besar, terutama ketika menggunakan Test-Time Augmentation (TTA) untuk inference. Ketiga, dataset yang digunakan terbatas pada citra dermatoskopi sehingga kemampuan generalisasi model terhadap citra kulit non-dermatoskopi, seperti citra klinis berbasis smartphone atau kondisi pencahayaan alami, belum dapat dipastikan dan masih memerlukan validasi tambahan. Selain itu, model belum menyediakan interpretabilitas klinis yang memadai; tidak adanya mekanisme penjelasan eksplisit seperti Grad-CAM dermatology-aware menjadikan model tetap berfungsi sebagai black box meskipun tingkat akurasinya tinggi. Keterbatasan-keterbatasan ini sekaligus membuka ruang bagi penelitian lanjutan, khususnya dalam pengembangan arsitektur yang lebih ringan, efisien, dan dapat dijalankan pada perangkat dengan keterbatasan komputasi, serta integrasi metode interpretabilitas yang dapat diterima oleh praktisi klinis untuk menunjang pengambilan keputusan medis yang lebih transparan.

V. KESIMPULAN

Penelitian ini berhasil mengembangkan dan mengevaluasi arsitektur Hybrid Vision Transformer–ConvNeXt yang dikombinasikan dengan Multi-Task Focal Loss dan strategi Medical Test-Time Augmentation (TTA) untuk tugas klasifikasi lesi kulit pada dataset HAM10000. Model yang diusulkan menunjukkan peningkatan kinerja signifikan dibandingkan beberapa metode baseline seperti EfficientNet-B3, ViT-Base, dan ResNet50, dengan capaian F1-Weighted sebesar 0.8744, menjadikannya model dengan performa terbaik dalam eksperimen ini. Temuan ini membuktikan bahwa integrasi kapasitas global attention dari Vision Transformer dengan efisiensi representasional ConvNeXt mampu menangkap pola visual yang kompleks pada citra dermatologi, termasuk tekstur halus, struktur melanin, dan irregulasi batas lesi yang sulit ditangkap oleh CNN konvensional.

Secara praktis, hasil penelitian ini memiliki implikasi penting dalam domain dermatologi digital. Dengan kemampuan klasifikasi yang presisi tinggi dan tingkat kesalahan yang rendah pada kelas berisiko seperti melanoma, model ini berpotensi menjadi alat pendukung keputusan (Clinical Decision Support System) yang dapat membantu dokter dalam melakukan skrining awal dan triase pasien secara cepat dan konsisten. Pendekatan ini juga dapat memperluas akses diagnosis dermatologis ke wilayah yang kekurangan tenaga ahli, terutama jika diintegrasikan ke dalam sistem berbasis web atau perangkat mobile.

Meskipun demikian, penelitian ini belum sepenuhnya menyelesaikan tantangan generalisasi ke kondisi dunia nyata. Oleh karena itu, penelitian selanjutnya disarankan untuk mengeksplorasi beberapa arah pengembangan, antara lain: (1) menggunakan dataset dermatologi yang lebih besar, beragam, dan multi-institusional guna meningkatkan robustnes model terhadap variasi demografis dan perangkat akuisisi; (2) mengembangkan arsitektur Hybrid Lightweight Transformer yang lebih efisien sehingga dapat diterapkan pada perangkat edge seperti smartphone; (3) menambahkan modul interpretabilitas klinis berbasis explainable AI seperti Dermatology-aware Grad-CAM untuk meningkatkan kepercayaan dokter terhadap proses pengambilan keputusan model; serta (4) memperluas cakupan klasifikasi ke lesi multipathology atau deteksi multi-label sehingga lebih mendekati kompleksitas kasus dermatologi nyata.

Dengan demikian, penelitian ini tidak hanya memberikan kontribusi empiris berupa peningkatan performa model, namun juga membuka arah penelitian baru dalam pengembangan sistem klasifikasi lesi kulit yang akurat, efisien, dan layak diadopsi dalam praktik medis modern.

DAFTAR PUSTAKA

[1] R. L. Siegel, K. D. Miller, and A. Jemal, “Cancer statistics, 2019,” *CA Cancer J Clin*, vol. 69, no. 1, pp. 7–34, Jan. 2019, doi: 10.3322/caac.21551.

[2] Q. Wu, Y. Yu, and X. Zhang, “A Skin Cancer Classification Method Based on Discrete Wavelet Down-Sampling Feature Reconstruction,” *Electronics (Basel)*, vol. 12, no. 9, p. 2103, May 2023, doi: 10.3390/electronics12092103.

[3] H. Zunair and A. Ben Hamza, “Melanoma detection using adversarial training and deep transfer learning,” *Phys Med Biol*, vol. 65, no. 13, p. 135005, Jul. 2020, doi: 10.1088/1361-6560/ab86d3.

[4] P. Bartlett, F. C. N. Pereira, C. J. C. . Burges, L. Bottou, and K. Q. Weinberger, *Advances in neural information processing systems 25 : 26th annual conference on neural information processing systems 2012*. Curran Associates, Inc., 2013.

[5] A. Dosovitskiy *et al.*, “An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale,” Jun. 2021, [Online]. Available: http://arxiv.org/abs/2010.11929

[6] Z. Liu, H. Mao, C.-Y. Wu, C. Feichtenhofer, T. Darrell, and S. Xie, “A ConvNet for the 2020s,” Mar. 2022, [Online]. Available: http://arxiv.org/abs/2201.03545

[7] P. Tschandl, C. Rosendahl, and H. Kittler, “Data descriptor: The HAM10000 dataset, a large collection of multi-source dermatoscopic images of common pigmented skin lesions,” *Sci Data*, vol. 5, Aug. 2018, doi: 10.1038/sdata.2018.161.

[8] T.-Y. Lin, P. Goyal, R. Girshick, K. He, and P. Dollár, “Focal Loss for Dense Object Detection.”

[9] C. Shorten and T. M. Khoshgoftaar, “A survey on Image Data Augmentation for Deep Learning,” *J Big Data*, vol. 6, no. 1, Dec. 2019, doi: 10.1186/s40537-019-0197-0.

[10] X. Huang *et al.*, “EConv-ViT: A strongly generalized apple leaf disease classification model

based on the fusion of ConvNeXt and Transformer,"
Information Processing in Agriculture, Mar. 2025,
doi: 10.1016/j.inpa.2025.03.001.