

BAB 1

PENDAHULUAN

1.1 Latar Belakang Masalah

Kecerdasan buatan (*Artificial Intelligence*, AI) berbasis *deep learning* telah menunjukkan potensi besar dalam memperluas akses layanan dermatologi, khususnya di wilayah dengan rasio dokter spesialis kulit yang rendah terhadap jumlah populasi [2]. Sejumlah studi rujukan melaporkan bahwa model *deep learning* mampu mencapai performa diagnostik yang setara, bahkan pada beberapa kasus melampaui, penilaian dokter spesialis dalam mengklasifikasikan lesi kulit dari citra klinis maupun dermoskopi [3, 4]. Namun demikian, potensi tersebut hanya dapat terwujud secara nyata apabila model yang dibangun terbukti *robust* dan *equitable*, yaitu mampu mengenali seluruh jenis penyakit secara proporsional, sekaligus memberikan performa yang setara pada seluruh kelompok demografis pasien [2]. Kesenjangan antara performa model yang dilaporkan pada kondisi eksperimen terkontrol dan performa aktualnya pada populasi klinis yang beragam inilah yang menjadi titik tolak permasalahan dalam penelitian ini.

Kesenjangan tersebut, pada praktiknya, bersumber dari dua kondisi yang melekat pada data dermatologi di dunia nyata. Prevalensi penyakit kulit di lapangan secara alami tidak seimbang, di mana studi beban penyakit global mencatat bahwa penyakit infeksi kulit dan kondisi dermatologis umum menyumbang proporsi kasus jauh lebih besar dibandingkan kondisi langka seperti kelainan autoimun atau neoplasma kulit [5]. Ketika model *deep learning* dilatih pada distribusi semacam ini tanpa penanganan khusus, model akan cenderung bias terhadap kelas mayoritas dan gagal mengenali kelas minoritas yang justru sering merupakan kondisi yang membutuhkan diagnosis cepat dan akurat [6]. Pada dataset DermaCon-IN yang digunakan dalam penelitian ini, tingkat ketidakseimbangan kelas tercatat sangat tinggi dengan *Imbalance Ratio* (IR) mencapai 58,32:1 antara kelas dengan jumlah sampel terbanyak dan tersedikit [1]. Kondisi ini menjadikan metrik evaluasi konvensional seperti akurasi keseluruhan tidak lagi representatif, sebab model dapat mencapai akurasi tinggi hanya dengan mengenali kelas mayoritas sembari mengabaikan kelas minoritas sepenuhnya, sebuah kelemahan yang secara langsung mengikis klaim *robustness* yang diharapkan dari sistem CAD dermatologi.

Di sisi lain, dan justru lebih sering diabaikan dalam pengembangan model

CAD dermatologi, adalah bias performa terhadap warna kulit pasien. Groh dkk. menunjukkan bahwa model klasifikasi lesi kulit yang dievaluasi pada dataset Fitzpatrick17k secara konsisten memiliki akurasi lebih rendah pada pasien dengan warna kulit gelap dibandingkan warna kulit terang, fenomena yang dikenal sebagai *skin tone bias* [7]. Temuan serupa dikonfirmasi oleh Daneshjou dkk. melalui *Diverse Dermatology Images* (DDI), yang menunjukkan bahwa model *state-of-the-art* secara konsisten mengalami penurunan performa substansial pada warna kulit gelap dan penyakit yang jarang terdokumentasi [8]. Akar masalah ini sebagian besar disebabkan oleh representasi warna kulit yang tidak merata pada dataset pelatihan publik, sehingga model yang tampak akurat secara agregat dapat menyembunyikan disparitas performa yang signifikan antar kelompok *Fitzpatrick Skin Type* (FST) [9] maupun *Monk Skin Tone* (MST) [10], risiko yang bertentangan langsung dengan prinsip pemerataan layanan kesehatan (*health equity*) yang menjadi dasar klaim *equitable* pada paragraf sebelumnya [2].

Yang membuat kedua kondisi ini sulit diselesaikan secara terpisah adalah keterkaitan struktural di antara keduanya. Penanganan *imbalance* yang hanya berorientasi pada peningkatan metrik agregat seperti akurasi atau *macro F1-score*, tanpa mempertimbangkan dampaknya terhadap kelompok demografis, berisiko justru memperbesar kesenjangan performa antar warna kulit [7]. Sebaliknya, upaya menjamin *fairness* tanpa memperhitungkan *imbalance* kelas akan sulit tercapai, karena kelas penyakit minoritas dalam dataset dermatologi kerap berkorelasi kuat dengan kelompok demografis minoritas pula, di mana manifestasi klinis suatu lesi (warna eritema, kontras pigmentasi, tekstur permukaan) dapat tampak sangat berbeda pada spektrum warna kulit yang berbeda, sehingga kondisi tertentu yang jarang terdokumentasi pada satu kelompok warna kulit cenderung lebih sering salah diklasifikasikan [2, 7]. Kondisi inilah yang mendasari penelitian ini untuk tidak memperlakukan penanganan *class imbalance* dan analisis *fairness* sebagai dua agenda yang independen, melainkan mengevaluasi *trade-off* antara performa klasifikasi dan keadilan lintas kelompok warna kulit secara simultan, sebagai satu kesatuan tujuan penelitian.

Untuk dapat mengevaluasi keterkaitan ini secara presisi, diperlukan dataset yang menyediakan anotasi lokasi lesi secara eksplisit, bukan sekadar label diagnosis pada tingkat citra secara keseluruhan. Pada dataset DermaCon-IN, setiap citra klinis dianotasi tidak hanya dengan label diagnosis, tetapi juga dengan skor FST (6 tingkat) dan MST (10 tingkat) pada tingkat pasien, serta *concept-level annotation* berupa *bounding box* yang menandai lokasi lesi secara spesifik pada citra [1].

Anotasi *bounding box* inilah yang menjadi *marker* penting dalam penelitian ini untuk memastikan bahwa evaluasi *fairness* dilakukan pada area lesi yang relevan secara klinis, bukan pada keseluruhan citra yang dapat mengandung latar belakang atau bagian tubuh yang tidak terkait diagnosis. Dengan demikian, hubungan antara kelas penyakit dan kelompok demografis dapat dianalisis secara lebih presisi, dan potensi *confounding factor* akibat variasi latar citra dapat diminimalkan.

Kebutuhan akan anotasi demografis dan *concept-level* semacam ini justru menjadi kendala pada mayoritas dataset dermatologi publik yang digunakan secara luas dalam riset AI, seperti ISIC, HAM10000, atau Fitzpatrick17k, yang didominasi oleh citra pasien berkulit terang dari populasi Barat, sehingga model yang dilatih di atasnya secara inheren kurang representatif terhadap populasi dengan spektrum warna kulit lebih gelap [7]. Kondisi ini menciptakan kesenjangan riset yang signifikan, mengingat populasi Asia Selatan, termasuk India, mencakup salah satu spektrum variasi warna kulit terluas secara global sekaligus menanggung beban morbiditas penyakit kulit yang besar menurut studi beban penyakit global [5].

Pemilihan India, dan lebih spesifik India Selatan, sebagai konteks penelitian bukan tanpa mempertimbangkan alternatif populasi lain yang juga relevan secara demografis. Indonesia sebagai negara tropis dengan beban penyakit kulit yang juga tinggi sesungguhnya merupakan konteks yang relevan secara epidemiologis [11], namun hingga saat ini belum tersedia dataset dermatologi Indonesia yang dikurasi secara prospektif dalam skala besar dan dilengkapi anotasi demografis warna kulit setara DermaCon-IN. Studi klasifikasi penyakit kulit di Indonesia umumnya masih mengandalkan dataset sekunder berskala kecil seperti DermNet yang tidak menyertakan label FST maupun MST secara sistematis [11]. Populasi Afrika Sub-Sahara, yang memiliki warna kulit lebih gelap secara dominan, juga menjadi kandidat yang secara teoritis menarik untuk penelitian *fairness*, namun dataset dermatologi berkualitas dari wilayah ini masih sangat terbatas dan baru mulai dikembangkan dalam skala kecil, seperti dataset citra penyakit kulit tropis terabaikan di Afrika Barat yang jumlah citra dan variasi diagnosisnya jauh lebih sedikit dibandingkan DermaCon-IN [12]. Ketimbang memaksakan penggabungan beberapa dataset kecil dari populasi berbeda, yang berisiko menimbulkan *confounding factor* akibat perbedaan protokol pencitraan dan peralatan antar-institusi [13], penelitian ini memilih satu dataset tunggal yang secara *intrinsic* sudah memiliki variasi warna kulit yang cukup luas untuk dianalisis. Populasi India Selatan secara klinis didominasi oleh Fitzpatrick tipe IV hingga VI, sebuah rentang yang lebih gelap dan lebih beragam dibandingkan populasi Kaukasoid yang

mendominasi dataset dermatologi Barat, karakteristik yang telah dibuktikan secara empiris melalui modifikasi skala Fitzpatrick khusus untuk populasi India [14]. Karakteristik ini menjadikan satu populasi India Selatan sudah cukup representatif untuk mengungkap disparitas performa model lintas warna kulit tanpa memerlukan penggabungan data dari berbagai benua yang secara metodologis lebih rawan bias pengukuran.

DermaCon-IN menjawab kebutuhan tersebut sebagai dataset dermatologi prospektif pertama yang dikurasi langsung dari klinik rawat jalan di India Selatan, mencakup 5.450 citra klinis dari sekitar 3.000 pasien dengan ratusan diagnosis berbeda yang distrukturkan mengikuti taksonomi hierarkis dermatologi [1]. Karakteristik distribusi penyakit pada populasi ini bersifat spesifik-regional (*region-specific disease distribution*), dipengaruhi oleh iklim tropis, faktor sosioekonomi, dan pola rujukan rawat jalan setempat [5], sehingga tidak dapat diasumsikan identik dengan distribusi penyakit pada dataset dermatologi Barat. Dari sisi metodologis, dataset ini pula yang menyediakan anotasi warna kulit ganda (FST dan MST) pada tingkat pasien yang jarang tersedia secara simultan pada dataset dermatologi publik lain, menjadikannya *testbed* yang ideal untuk penelitian *fairness-aware learning* yang menjadi fokus utama skripsi ini.

Ketersediaan *testbed* yang representatif tersebut perlu dipasangkan dengan arsitektur klasifikasi yang mampu memanfaatkan karakteristik visual lesi kulit secara optimal. *Dual Attention Vision Transformer* (DaViT) dipilih sebagai *backbone* klasifikasi dalam penelitian ini karena mengintegrasikan dua mekanisme *self-attention* yang saling melengkapi dalam setiap *block*, yaitu *spatial attention* yang menangkap detail tekstur lokal lesi, dan *channel attention* yang memodelkan konteks global antar-*channel* fitur [15]. Kombinasi ini relevan untuk klasifikasi penyakit kulit karena diagnosis dermatologis seringkali memerlukan kombinasi informasi tekstur lokal (misalnya pola skuama atau vesikel) dan konteks morfologis global (misalnya bentuk dan distribusi lesi secara keseluruhan), kebutuhan yang sejalan dengan kemampuan arsitektur *Vision Transformer* dalam memodelkan dependensi jarak jauh pada citra [16]. Secara empiris, varian DaViT terbukti mengungguli baik model konvolusional (CNN) maupun *Vision Transformer* standar seperti Swin Transformer pada berbagai tugas klasifikasi citra [15, 17], sekaligus mempertahankan efisiensi komputasi yang relevan untuk sumber daya penelitian yang terbatas, di mana varian DaViT-Tiny yang digunakan dalam penelitian ini hanya memiliki sekitar 28,3 juta parameter dengan akurasi *top-1* 82,8% pada ImageNet-1K [15, 18]. Kombinasi *backbone* DaViT-Tiny dengan dataset

DermaCon-IN, yang menyediakan *concept-level annotation* (*bounding box* lesi), label diagnosis hierarkis, dan metadata demografis ganda (FST/MST) secara bersamaan, menjadikan keduanya saling melengkapi untuk mengevaluasi baik performa klasifikasi maupun *fairness* model secara menyeluruh, sebuah kombinasi yang belum banyak dieksplorasi pada studi *benchmark* arsitektur *transformer* terhadap dataset dermatologi non-Barat.

Berdasarkan uraian di atas, penelitian ini secara spesifik diarahkan untuk menjawab kebutuhan akan metode penanganan *class imbalance* yang tidak hanya unggul dari sisi performa klasifikasi, tetapi juga terbukti paling adil (*fairness-aware*) lintas kelompok warna kulit, dengan menggunakan *backbone* DaViT-Tiny pada dataset DermaCon-IN sebagai studi kasus representatif populasi dermatologi India Selatan.

1.2 Rumusan Masalah

Berdasarkan latar belakang yang telah diuraikan, rumusan masalah dalam penelitian ini adalah sebagai berikut.

1. Bagaimana performa klasifikasi model DaViT-Tiny pada dataset DermaCon-IN yang memiliki tingkat *class imbalance* ekstrem ($IR = 58,32:1$), dan metode penanganan *imbalance* apa yang memberikan peningkatan performa klasifikasi paling signifikan?
2. Bagaimana tingkat kesenjangan (*gap*) performa model antar kelompok demografis *Fitzpatrick Skin Type* (FST) dan *Monk Skin Tone* (MST), dan metode penanganan *imbalance* apa yang menghasilkan kesenjangan *fairness* paling rendah?
3. Metode penanganan *class imbalance* manakah yang memberikan keseimbangan (*trade-off*) paling optimal antara performa klasifikasi dan *fairness* lintas kelompok warna kulit, berdasarkan skor komposit yang mempertimbangkan kedua aspek tersebut secara simultan?

1.3 Tujuan Penelitian

Berdasarkan rumusan masalah di atas, tujuan penelitian ini adalah sebagai berikut.

1. Mengevaluasi dan membandingkan performa klasifikasi lima metode penanganan *class imbalance* (*Cross-Entropy Loss* sebagai *baseline*, *Weighted Random Sampler*, *Class-Balanced Loss*, *Balanced Softmax Loss*, dan *Logit Adjustment Loss*) pada model DaViT-Tiny menggunakan dataset DermaCon-IN.
2. Menganalisis tingkat *fairness* setiap metode berdasarkan kesenjangan akurasi (*accuracy gap*), *Equalized Odds Gap*, dan *Demographic Parity Gap* pada kelompok FST dan MST.
3. Menentukan metode penanganan *imbalance* dengan skor komposit (*performance-fairness composite score*) terbaik sebagai rekomendasi metode yang paling sesuai untuk implementasi sistem CAD dermatologi yang adil dan andal.

1.4 Batasan Masalah

Agar penelitian ini terarah dan fokus, ditetapkan batasan masalah sebagai berikut.

1. Penelitian menggunakan dataset DermaCon-IN yang telah dibagi menjadi 4.399 data latih dan 1.051 data uji, dengan 8 kelas utama penyakit kulit hasil pengelompokan taksonomi hierarkis dermatologi.
2. Arsitektur *backbone* yang digunakan dibatasi pada DaViT-Tiny dengan bobot *pretrained* ImageNet-1K, tanpa membandingkan arsitektur *backbone* lain secara ekstensif.
3. Evaluasi *fairness* dibatasi pada dua sumbu demografis yang tersedia pada metadata dataset, yaitu *Fitzpatrick Skin Type* (FST) dan *Monk Skin Tone* (MST), tanpa melibatkan variabel demografis lain seperti usia atau jenis kelamin secara mendalam.
4. Penelitian dilakukan pada lingkup komputasi lokal dengan GPU NVIDIA RTX 2060 Super (8 GB VRAM), sehingga jumlah *epoch* pelatihan dan konfigurasi *hyperparameter* disesuaikan dengan keterbatasan sumber daya tersebut.

1.5 Manfaat Penelitian

Manfaat Teoritis

Penelitian ini diharapkan dapat memperkaya literatur mengenai integrasi metode penanganan *class imbalance* dengan analisis *fairness* pada domain klasifikasi citra medis, khususnya pada konteks populasi non-Barat yang masih terbatas dieksplorasi dalam riset AI dermatologi.

Manfaat Praktis

Hasil penelitian ini diharapkan dapat menjadi rujukan bagi pengembang sistem CAD dermatologi dalam memilih metode penanganan *imbalance* yang tidak hanya meningkatkan performa klasifikasi, tetapi juga menjamin keadilan performa lintas kelompok warna kulit pasien, sehingga mendukung pemerataan akses layanan diagnosis dermatologi berbasis AI.

1.6 Sistematika Penulisan

Laporan penelitian ini disusun dalam lima bab dengan sistematika sebagai berikut:

- **Bab 1 PENDAHULUAN**
Menjelaskan latar belakang masalah yang melatarbelakangi penelitian, yaitu tantangan *class imbalance* dan bias demografis pada sistem AI dermatologi, rumusan masalah, batasan penelitian, tujuan, manfaat, dan sistematika penulisan.
- **Bab 2 LANDASAN TEORI**
Menjabarkan teori-teori yang mendasari penelitian, meliputi: dataset DermaCon-IN, arsitektur DaViT-Tiny, metode penanganan *class imbalance* (CB Loss, Balanced Softmax, Logit Adjustment), serta konsep dan metrik *fairness* dalam *machine learning*.
- **Bab 3 METODOLOGI PENELITIAN**
Mendeskripsikan metodologi penelitian secara rinci, meliputi alur penelitian, konfigurasi lingkungan eksperimen, desain ablasi lima eksperimen, protokol pelatihan, dan protokol evaluasi performa serta *fairness*.

- **Bab 4 HASIL DAN DISKUSI**

Menyajikan hasil eksperimen secara kuantitatif beserta analisis dan diskusi mendalam, meliputi dinamika pelatihan, perbandingan performa klasifikasi, analisis *fairness* demografis, dan perankingan akhir berdasarkan *composite score*.

- **Bab 5 SIMPULAN DAN SARAN**

Merangkum simpulan dari keseluruhan hasil penelitian, mencakup performa klasifikasi kelima metode penanganan *imbalance*, tingkat *fairness* demografis berdasarkan FST dan MST, serta metode terbaik secara holistik berdasarkan *composite score*, dan memberikan saran untuk peningkatan performa, peningkatan *fairness*, serta pengembangan dataset dan model pada penelitian lanjutan.

