Journal of Applied Informatics and Computing (JAIC)
Vol.10, No.3, June 2026, pp. 3098~3106
e-ISSN: 2548-6861

Analysis of ResNet50 Model Response to Skin Tone Variations in Medical Image-Based Skin Disease Classification

Made Ireina Dwiandra Divayanti [1]*, I Gusti Ngurah Lanang Wijayakusuma[2]*
* Matematika, Universitas Udayana
divayanti.2308541035@student.unud.ac.id [1], lanang_wijaya@unud.ac.id [2]

Article Info

Article history:

Received 2026-04-18
Revised 2026-06-19
Accepted 2026-06-22

Keyword:

Deep learning,
Fitzpatrick skin type,
ResNet50 skin disease classification,
skin tone bias.

ABSTRACT

Skin disease classification with deep learning has shown promising performance, however many models are primarily trained on datasets featuring light skin tones, which raises question about their effectiveness across a variety of akin types. This study analyses the response of a ResNet50 model based on transfer learning when faced with different skin tones in order to classifying skin disease using medical images. The model was trained on the HAM100000 which categorized into three classes: benign, malignant, and non-neoplastic. A bias analysis was then performed using the Fitzpatrick 17k dataset. The model demonstrated an overall accuracy of 70.85%, a precision rate of 74.03%, and a recall rate of 65.51%. Further analysis showed that the model had a consistent pattern of predicting malignant cases, which increased with darker skin tones, rising from 54% to 68.3%. To mitigate this issue, a threshold tuning approach was applied. After mitigation, the model achieved an accuracy of 74%, a weighted F1-score of 76%, dan a macro F1-score of 55%. Fairness evaluation after mitigation showed tha the proportion of malignant predictions increased from 56,3% in FST I to 69,9% in FST VI. These findings suggest that threshold tuning can improve classification performance and partially reduce bias intensity.

I. PENDAHULUAN

Kulit merupakan organ terluar dan terbesar dari tubuh yang memiliki kontak langsung dengan lingkungan sehingga kulit sangat kmudah untuk terserang penyakit. Menurut WHO, penyakit kulit termasuk penyakit yang umum menyerang manusia dan hamper 900 juta orang di dunia mengidap penyakit kulit dengan 80% diantaranya adalah dermatitis/eksim [1]. Selain itu, pada tahun 2021, terdapat lebih dari 1,73 miliar kasus infeksi jamur pada kulit di seluruh dunia. Wilayah Asia memiliki tingkat insedensi tertinggi, sedangkan Oseania memiliki tingkat terendah [2]. Angka kejadian penyakit kulit yang tinggi menunjukkan bhwa permasalahan ini menjadi tantangan serius dalam bidang kesehatan. Umumnya, diagnosis penyakit kulit dilakukan secara manual melalui obeservasi langsung oleh tenaga medis. Metode ini sangat bergantung pada keahlian dokter, sehingga potensi kesalahan diagnosisinya tergolong tinggi dan membutuhkan waktu yang sama. Dengan berkembangnya teknologi, berbagai metode telah diusulkan untuk mendukung proses diagnose penyakit. Salah satnya adalah *deep learning* yang memiliki kemampuan memahami pola rumit dari data yang jumlahnya besar. *Convolutional Neural Network* (CNN) telah menunjukkan hasil yang sangat memuaskan dalam mengklasifikasikan citra, termasuk citra medis.

Berbagai model CNN telah digunakan dalam berbagai penelitian untuk mendeteksi penyakit kulit dan memperoleh akurasi yang cukup tinggi. Klasifikasi dan segmentasi penyakit kulit dengan model EfficientNetV2Small memperoleh akurasi 94,04% [3]. Penelitian tersebut menguji akurasi beberapa model seperti, VGG16, Xception, EfficientNetB3, dan MobileNetV2 memiliki performa yang paling bagus dengan akurasi pelatihan 90,352% dan pengujian 98,374% kemudian diikuti model Xception yang memperoleh akurasi pelatihan 87,789% dan akurasi pengujian 87,985%, sementara EfficientNetB3 meraih 88,857% akurasi pelatihan dan 97,968% akurasi pengujian. Meskipun terdapat beberapa fluktuasi, VGG16 memiliki akurasi pengujian yang tinggi yakni, 95,858% namun dengan konvergensi yang lebih lambat [4]. Model MobileNetV2 yang

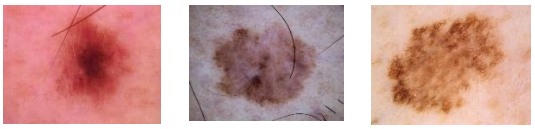
diintegrasikan dengan LSTM terbukti efisien untuk klasifikasi dengan mendeteksi penyakit kulit dengan akurasi 85,34% [5]. Penelitian-penelitain tersebut menggunakan dataset HAM100000 yang memiliki 10.015 gambar dermaskopi. Dataset HAM10000 merupakan data yang dikumpulkan dari institusi di Austria dan Australia yang didominasi oleh populasi berkulit putih (*Caucasian*) [6]. Hal ini menyebabkan model CNN yang dilatih oleh dataset tersebut berpotensi memiliki keterbatasan dalam mengenali penyakit kulit pada tipe kulit lain yang dapat mempengaruhi tampilan dari suatu penyakit meskipun performa model sudah mencapai akurasi lebih dari 90%. Dengan demikian, performa akurasi yang tinggi belum tentu memiliki kemampuan yang optimal terhadap seluruh variasi kulit. Penelitian yang membandingkan ratusan dermatolog dan dokter layanan primer dalam mendiagnosis penyakit kulit pada kulit gelap menunjukkan bahwa para akurasi diagnosis pada kulit gelap lebih rendah hingga 4 poin presentase dibandingkan kulit terang, meskipun deep learning memang dapat meningkatkan akurasi diagnosa dokter secara keseluruhan, namun kesenjangan akurasi dari yang awalnya sebesar 4 poin presentase naik menjadi 5 poin presentase [7]. Penelitian tersebut menunjukkan bahwa adanya potensi bias dalam sistem deep learning yang disebabkan oleh kurangnya representasi data pada tipe kulit tertentu. Bias terhadap warna kulit juga terjadi pada arsitektur SkinGPT-4 yang memggunakan Vision Transformer (ViT) dan modul Q-formeer untuk menghasilkan *image* embeddings. Berdasarkan tingkat *demographic parity* yang diperoleh, terlihat perbedaan ckup mencolok antara warna kulit paling terang dengan paling gelap yakni, sebesar 0,10 dan 0,15 [8]. Peneliti menyebutkan bahwa hal tersebut terjadi karena model dilatih dengan dataset yang didominasi oleh warna kulit yang terang, sehingga akurasi pada warna kulit yang lebih gelap berkurang.

Meskipun berbagai arsitektur CNN telah menunjukkan performa tinggi dalam klasifikasi penyakit kulit dengan dataset HAM1000, Sebagian besar penelitian berfokus pada peningkatan akurasi tanpa mempertimbangkan keadilan model terhadap variasi tipe kulit. Oleh karena itu, penelitian ini bertujuan untuk mengevaluasi performa model yang dikembangkan serta menganalisis keterbatasannya, khususnya dalam mendeteksi penyakit kulit pada tipe kulit yang lebih gelap. Oleh karena itu, penelitian ini tidak hanya berfokus pada tingkat akurasi model, tetapi juga mempertimbangkan potensi bias yang mempengaruhi hasil klasifikasi. Penelitian ini diharapkan memberikan wawasan terkait pentingnya variasi data dalam pengembangan model serta menjadi referensi dalam pengembangan sistem pendektesi penyakit kulit yang dapat diandalkan untuk berbagai tipe kulit.

II. METODE

1) *Dataset*

Penelitian ini menggunakan dua dataset yakni, HAM100000 dan Fitzpatrick 17k yang diperoleh melalui platform Kaggle. Dataset HAM100000 terdiri dari 10.015 gambar dermaskopi yang terbagi menjadi 7 kelas penyakit kulit yakni, *melanocytic nevi* (nv) dengan 6705 gambar, *melanoma* (mel) 1113 gambar, *benign keratosis-like lesions* (bkl) 1099 gambar ,*basal cell carcinoma* (bcc) dengan 514 gambar, *actinic keratoses* (akiec) dengan 327 gambar, *vascular lesions* (vasc) dengan 142 gambar, dan *dermatifibroma* (df) dengan 115 gambar. Dataset ini banyak digunakan dalam penelitian klasifikasi karena jumlah data yang memadai.



Gambar 1. Sampel Dataset HAM100000

Dataset Fitzpatrick 17k terdiri dari 15.504 fotografi klinis dengan 114 kondisi kulit yang terbagi menjadi 6 tipe kulit yakni dari tipe I yang merupakan kulit paling terang hingga tipe VI yang merupakan kulit paling gelap. Masing-masing tipe kulit memiliki jumlah data yang berbeda yakni sebagai berikut, tipe I (5561), II (3180), III (2934), IV (1974), V (1293), dan VI (562) [9],[10]. Dataset ini dipilih karena menyediakan variasi jenis kulit, sehingga memadai digunakan untuk menganalisis respons model.

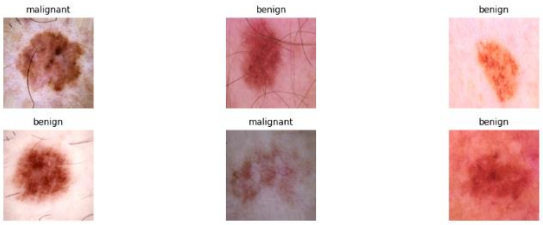
Gambar 2. Sampel Dataset Fitzpatrick 17k

Dataset HAM100000 sebagai dataset utama digunakan dalam seluruh tahap utama seperti, pelatihan, validasi, hingga pengujian model. Model akan dilatih dengan dataset ini untuk mengenali pola citra medis serta mengevaluasi model seperti, akurasi, presisi, dan recall. Selain itu, dataset Fitzpatrick 17k sebagai dataset pendukung tidak terlibat dalam proses pelatihan maupun evaluasi, namun digunakan sebagai data uji tambahan untuk menganalisis respons model terhadap variasi tipe warna kulit, khususnya pada tipe kulit yang lebih gelap. Penggunaan dataset ini bertujuan untuk mengidentifikasi potensi data bias pada model dalam mengklasifikasikan penyakit kulit pada tipe kulit yang lebih gelap yang kurang terwakili dalam dataset utama.

2) *Pre-processing Data*

Dataset HAM100000 terdiri dari gambar berukuran 300 × 300 piksel dan 600 × 600 piksel. Rata-rata gambar berukuran 400 × 650 piksel dimana ukuran tersebut tidak sesuai dengan masukan yang diperlukan model ResNet50 yakni, 224 × 224 piksel [11]. Oleh karena itu, *pre-processing* data diperlukan untuk menyesuaikan nilai input yang diterima model. Selanjutnya, dilakukan proses normalisasi dengan mengubah rentang nilai dari 0-255 menjadi 0-1 supaya proses pelatihan lebih cepat dan stabilitas model meningkat. Demi mengurangi risiko *overfitting*, Teknik augmentasi data dilakukan pada

data pelatihan. Teknik augmentasi yang dilakukan adalah *rotation* hingga ± 20 derajat, *horizontal flip, zoom in/zoom out* hingga $\pm 10\%$, dan *brightness adjustment* yang divariasikan antara 80% hingga 120%. Proses ini bertujuan untuk meningkatkan generalisasi model terhadap variasi data [12].



Gambar 3. Sampel Gambar Setelah *Pre-processing*

Dataset utama akan dibagi menjadi 2 yakni, 80% untuk data pelatihan dan 20% untuk data pengujian, kemudian data pelatihan kembali dibagi menjadi 2 bagian yakni, 72% untuk pelatihan dan 8% untuk validasi. Selain itu proses mapping dilakukan terhadap label dataset utama yang memiliki tujuh kelas penyakit kulit menjadi tiga kategori utama, yaitu non-neoplastic, benign, dan malignant yang bertujuan untuk menyederhanakan klasifikasi serta meningkatkan keseimbangan distribusi data antar kelas. Kategori non-neoplastic terdiri gabungan kelas dermatofibroma (df) dan vascular lesion (vasc) yang termasuk penyakit kulit dengan kondisi ringan. Kemudian kategori benign terdiri dari gabungan kelas melanocytic nevi (nv) dan benign keratosis (bkl) yang sudah termasuk tumor jinak namun tidak menyebar dan tidak ganas. Sementara, kategori malignant terdiri dari gabungan kelas melanoma (mel), basal cell carcinoma (bcc), dan actinic keratosis (akiec) yang sudah termasuk tumor ganas atau yang sering disebut kanker kulit.

TABEL I
DISTRIBUSI KELAS SETELAH PROSES MAPPING

Kelas	Jumlah Distribusi
Benign	7804
Malignant	1954
Non-neoplastic	257

3) *Implementasi ResNet50*

ResNet50 adalah arsitektur CNN yang diperkenalkan oleh He *et al.* pada tahun 2016 melalui paper *Deep Residual Learning for Image Recognition* [13]. Pendekatan *residual learning* yang digunakan pada arsitektur ResNet50 memungkinkan jaringan untuk mempelajari fungsi residual terhadap input melalui *shortcut connection* demi mengatasi permasalahan degradasi yakni, performa model yang menurun saat penambahan lapisan [14]. Sesuai dengan namanya, ResNet50 terdiri dari 50 lapisan dengan desain *bottleneck block* yang tersusun tiga lapisan konvolusi berukuran 1×1, 3×3, dan 1×1 di setiap blok residual [14]. Lapisan konvolusi 1×1 berguna untuk mereduksi dan merestorasi dimensi, sehingga lapisan 3×3 dapat beroprasi pada dimensi yang lebih kecil. Arsitektur ResNet50 terbagi menjadi 5 tahap konvolusi yang diikuti lapisan *Global Average Pooling* dan lapisan klasifikasi penuh. Model ResNet50 terdiri dari lima blok dengan setiap bloknya memiliki jumlah filter yang meningkat fitur yang semakin kompleks [15]. Untuk menjaga informasi penting selama propogasi, *shortcut connection* di setiap blok residual menambahkan input langsung ke output [15]. Setelahnya, *average pooling* digunakan untuk mengurangi dimensi data tanpa kehilangan fitur penting [15]. Selanjutnya, lapisan yang *fully connected* akan menggunakan fungsi aktivasi *softmax* untuk klasifikasi multi kelas [15]. Fungsi aktivasi *softmax* dirumuskan sebagai berikut [16]:

$$\sigma(z_i) = \frac{e^{z_i}}{\sum_{j=1}^{K} e^{z_j}} \quad (1)$$

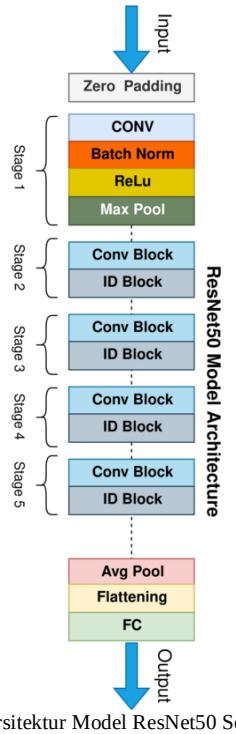
Dengan:
$\sigma(z_i)$: Probabilitas prediksi untuk kelas ke-i
z_i : Output dari neuron ke-i
K : Jumlah kelas

Selanjutnya, *loss function categorical crossentropy* digunakan untuk mengukur kesalahan prediksi selama pelatihan yang dirumuskan sebagai berikut [16]:

$$L = -\sum_{i=1}^{K} y_i \cdot \log(\hat{y}_i) \quad (2)$$

Dimana:
L : Nilai *loss*
y_i : Label aktual ke-i (bernilai 1 jika benar, 0 jika tidak)
$\hat{y}_i$: Probabilitas prediksi kelas ke-i dari *softmax*
K : Jumlah kelas

Untuk menganalisa respons model terhadap variasi warna kulit dengan dataset Fitzpatrick 17k, evaluasi dilakukan secara terpisah untuk setiap kelompok tipe kulit guna mengidentifikasi potemsi penurunan performa model pada tipe kulit yang lebih gelap. Pendekatan ini sejalan dengan dengan harapan penelitian sebelumnya yang menekankan evaluasi terhadap variasi tipe kulit untuk mengetahui kesenjangan performa yang tidak terlihat hanya dengan mengukur akurasi keseluruhan.



Gambar 4. Arsitektur Model ResNet50 Secara Umum

Jaringan pada ResNet50 hanya mempelajari fungsi residual $F(x)$ terhadap input x melalui *shortcut connection* supaya lebih mudah dioptimasi karena hanya mempelajari penyimpangan kecil, bukan pemetaan penuh [13].

$$H(x) = F(x) + x \quad (3)$$

Keterangan:
$H(x)$: Output yang diinginkan dari blok residual
$F(x)$: Fungsi residual
x : Input yang diteruskan melalui *shortcut connection*

Fungsi $F(x)$ diterapkan sebagai tiga lapisan kovolusi berurutan dengan fungsi aktivasi ReLU (*Rectified Linear Unit*) pada arsitektur *bottleneck* ResNet50. ReLU didefinisikan dengan persamaan (2) [17]:

$$ReLU(x) = \max(0, x) \quad (4)$$

Penelitian ini mengimplementasikan ResNet50 menggunakan pendekatan *transfer learning.* Pendekatan ini dipilih karena model memungkinkan untuk mengadopsi representasi fitur visual tingkat rendah seperti tepi dan tekstur, hingga fitur tingkat tinggi seperti bentuk dan pola untuk klasifikasi citra medis [11]. ResNet50 dirancang khusus untuk mengatasi masalah *vanishing gradient* dengan memanfaatkan residual atau *skip connections*. Model ini efektif untuk digunakan pada dataset yang memiliki ukuran besar dan kompleks karena kemampuannya dalam mempertahankan kestabilan selama proses pelatihan.

4) *Mitigasi Bias*

Bias dataset dapat terjadi karena model menempatkan kelompok-kelompok tertentu pada posisi yang kurang menguntungkan [18]. Oleh karena itu, mitigasi bias diperlukan demi membangun model yang dapat dipercaya. Secara umum, metode mitigasi bias dibagi menjadi tiga, yaitu *pre-processing, in-processing,* dan *post-processing* [19]. Penelitian ini menerapkan ketiga strategi bias tersebut dengan *pre-processing* berupa *rebalancing dataset, in-processing* berupa *focal loss* sebagai upaya *fairness aware training*, serta *post-processing* dengan *threshold tuning*.

Tabel II menunjukkan distribusi data pelatihan dari dataset HAM100000 sebelum proses *oversampling* yang terlihat memiliki ketidakseimbangan yang ekstrem.

TABEL II
DISTRIBUSI KELAS SEBELUM OVERSAMPLING

Kelas	Sebelum Oversampling
Benign	5632
Malignant	1376
Non-neoplastic	168

Untuk mengatasi hasl tersebut, dilakukan *random oversampling with replacement* pada data pelatihan dari dataset HAM100000. Kelas malignant dan non-neoplastic diduplikasi hingga 1,5 kali jumlah kelas mayoritas demi mendapatkan dataset yang seimbang tanpa menambahkan data baru.

TABEL III
DISTRIBUSI KELAS SETELAH OVERSAMPLING

Kelas	Sebelum Oversampling
Benign	5632
Malignant	5632
Non-neoplastic	5632

Focal loss diperkenalkan oleh Lin *et al.* dengan menambahkan faktor $(1 - p_t)^{\gamma}$ ke fungsi *cross-entropy* standar. Ditetapkan $\gamma > 0$ untuk mengurangi nilai *relative loss* pada *well-classified examples* ($p_t > 5$), sehingga model lebih berfokus pada sampel minoritas dan salah diklasifikasikan. *Focal loss* memiliki formula sebagai berikut [20]:

$$\mathrm{FL}(p_t) = -(1 - p_t)^{\gamma} \log(p_t) \quad (5)$$

Focusing parameter (γ) berperan untuk mengatur seberapa besar kontribusi pada sampel yang mudah diklasifikasi akan dikurangi. Semakin besar nilai γ, akan semakin kuat juga pengaruh faktor modulasinya. Penelitian tersebut menemukan bahwa nilai $\gamma = 2$ memberikan hasil yang terbaik [20]. Oleh karena itu, penelitian ini menggunakan nilai $\gamma = 2$ dengan bobot $\alpha = [0{,}2; 0{,}4; 0{,}4]$ untuk kelas benign, malignant, dan non-neoplastic.

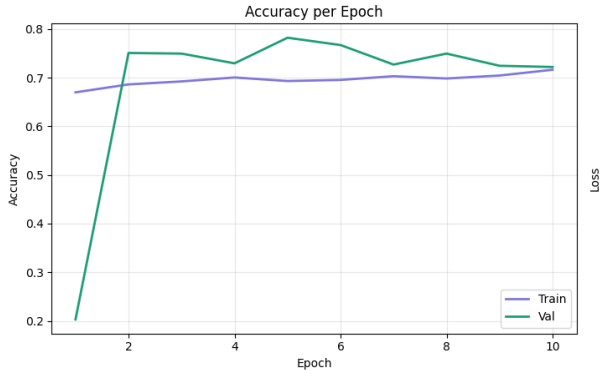
Penerapan *focal loss* dan *oversampling* menghasilkan model dengan distribusi probabilitas yang *overconfident* terhadap kelas minoritas. Hal ini terjadi karena dua strategi mitigasi bias yang mendorong model untuk berfokus pada

kelas minoritas saat pelatihan. Oleh karena itu, diterapkan *threshold tuning* untuk mengatasi ketidakseimbangan tersebut. *Threshold tuning* adalah pendekatan umum dalam metode *post-processing* untuk mengurangi bias dengan keputusan klasifikasi diatur kembali berdasarkan metrik tertentu setelah model selesai dilatih [21]. Dalam penelitian ini, nilai *threshold* optimal per kelas ditentukan melalui pencarian grid yang bertujuan meningkatkan *macro F1-score* pada data validasi. Diperoleh *threshold* optimal untuk masing-masing kelas adalah 0,22 untuk kelas benign, 0,32 untuk kelas malignant, dan 0,46 untuk kelas non-neoplastic.

III. HASIL DAN PEMBAHASAN

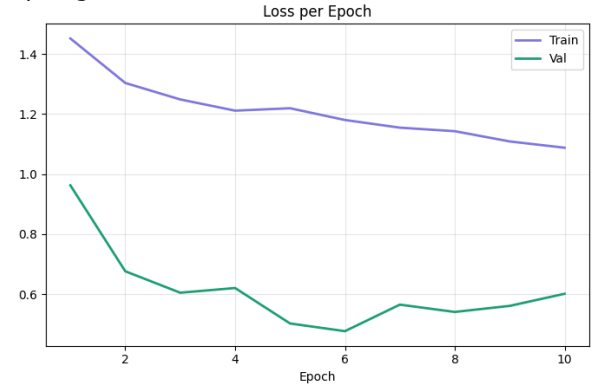
Model ResNet50 berbasis *transfer learning* dievaluasi dengan 2.041 gambar dermaskopi yang belum pernah dilihat oleh model selama proses pelatihan. Sebelum dievaluasi, performa model dipantau melalui kurva *training history* yang menunjukkan nilai akurasi dan *loss* setiap epoch.

Terlihat pada Gambar 5, *validation accuracy* meningkat secara konsisten dari *epoch* pertama hingga mencapai nilai terbaik pada *epoch* 5 yakni sebesar 78,20% yang kemudian distabilkan dengan *early stopping*.



Gambar 5. Kurva *Training* dan *Validation Accuracy* per *Epoch*

Gambar 6 menunjukkan penurunan dari 0,96 hingga 0,5 tanpa fluktuasi tajam yang signifikan di *validation loss*. Pola konvergensi yang stabil ini mengindikasikan proses pelatihan yang berlangsung dengan baik dan model tidak *overfitting*.



Gambar 6. Kurva *Training* dan *Validation Loss* per *Epoch*

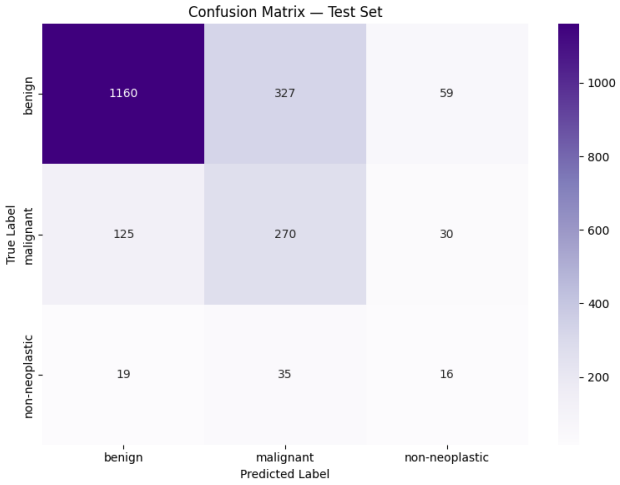
Hasil evaluasi data uji menunjukkan bahwa model memperoleh akurasi sebesar 70,85%, presisi 74,03%, dan *recall* 65,51% dengan nilai *loss* sebesar 0,6228. Dengan pendekatan *feature extraction* yang membekukan seluruh lapisan *backbone* ResNet50 dan hanya melatih lapisan yang baru ditambahkan yakni, *Global Average Pooling*, lapisan *dense* dengan aktivasi ReLU, dan lapisan *softmax* untuk tiga kelas, performa tersebut dapat tercapai.

Seperti yang dijabarkan pada Tabel III, terlihat adanya kesenjangan performa yang cukup signifikan. Kelas benign memperoleh nilai presisi tertinggi yakni sebesar 89% dan *F1-score* sebesar 81% yang berarti model dapat mengenali kelas mayoritas dengan baik. Sementara itu, kelas malignant memperoleh nilai presisi sebesar 43% dengan *F1-score* 51% menandakan tingginya jumlah *false positive* yang berpotensi berbahaya secara klinis. Kelas non-neoplastic memiliki performa paling rendah dengan nilai presisi sebesar 15% dan nilai *recall* sebesar 23% yang disebabkan oleh sampel kelas yang sangat sedikit dalam data uji yakni hanya sebanyak 70 sampel akibat ketidaksetimbangan distribusi kelas pada dataset HAM100000.

TABEL IV
CLASSIFICATION REPORT

	Precision	***Recall***	***F1-score***
Benign	0,89	0,75	0,81
Malignant	0,43	0,64	0,51
Non-neoplastic	0,15	0,23	0,18

Melalui *confusion matrix*, dapat terlihat seberapa banyak sampel yang berhasil diklasifikasikan oleh model. Pada kelas benign, model berhasil mengklasifikasi 1.160 sampel dengan benar meski terdapat 327 sampel yang salah diklasifikasikan sebagai malignant dan 59 sampel diklasifikasikan sebagai non-neoplastic. Sementara itu, 270 sampel berhasil diklasifikasikan dengan benar pada kelas malignant oleh model dengan 125 sampel salah diklasifikasikan sebagai benign dan 30 sampel salah diklasifikasikan sebagai non-neoplastic. Tingginya jumlah *false negative* pada kelas ini membahayakan secara klinis karena lesi yang bersifat ganas tidak terdeteksi. Adapun kelas non-neoplastic dengan performa paling rendah yang hanya dapat mengklasifikasikan 16 sampel dengan benar dan sebanyak 19 sampel dan 35 sampel salah diklasifikasikan sebagai benign dan malignant. Hal ini dapat terjadi karena adanya tumpang tindih fitur visual antar kelas non-neoplastic dengan kelas lainnya yang lebih diperburuk dengan terbatasnya jumlah sampel pada kelas ini. Penambahan data pada kelas minoritas atau menerapkan teknik *oversampling* dapat meningkatkan representasi kelas minoritas.



Gambar 7. Confusion Matrix

Model yang telah dilatih kemudian diuji dengan dataset Fitzpatrick 17k yang memiliki 15.504 fotografis klinis yang diklasifikasi menjadi 6 tipe kulit untuk menganalisis respons model terhadap variasi warna kulit. Secara keseluruhan, model memprediksi 9612 gambar dari dataset Fitzpatrick 17k sebagai kelas malignant dimana proporsi ini lebih tinggi dari kelas malignant pada data pelatihan. Perbedaan ini mengindikasikan adanya *distributional shiuft* dimana model yang dilatih pada citra dermaskopi belum tentu dapat digeneralisasi dengan baik pada jenis citra klinis yang berbeda.

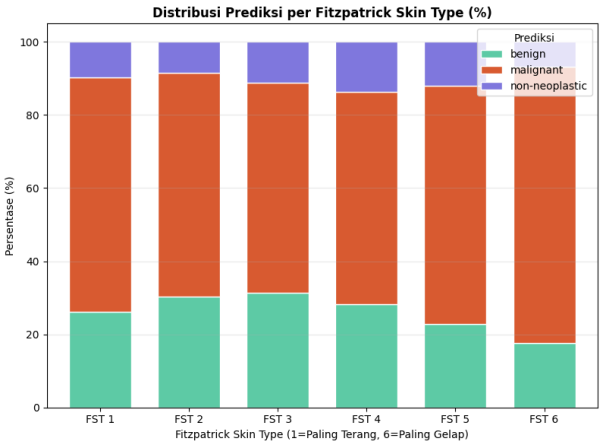
TABEL V
DISTRIBUSI PREDIKSI DATASET FITZPATRICK 17K

Kelas	Distribusi Prediksi
Benign	4290
Malignant	9612
Non-neoplastic	1602

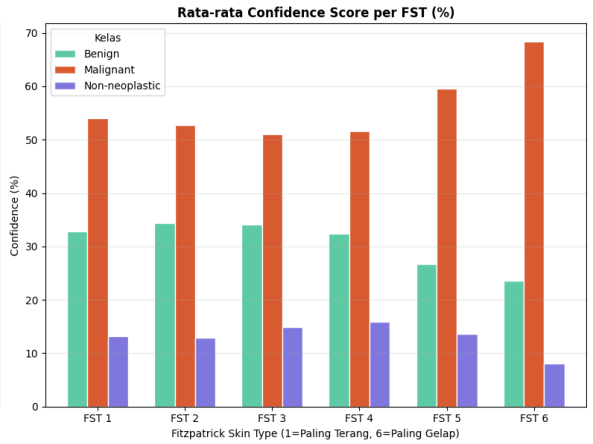
Berdasarkan Tabel V, Gambar 8 dan Gambar 9, terlihat pola bias yang konsisten terhadap warna kulit gelap. Prediksi maglinant pada FST 1 yakni, kulit yang paling terang adalah 64% dengan rata-rata *confidence score* sebesar 54%, sedangkan prediksi maglinant pada FST 6 yakni, kulit paling gelap meningkat hingga 75,6% dengan *confidence score* sebesar 68,3%. Peningkatan ini menunjukkan kecenderungan prediksi malignant seiring dengan meningkatnya tingkat kegelapan warna kulit. Dikarenakan dataset HAM100000 yang digunakan sebagai data pelatihan didominasi oleh populasi berkulit terang sehingga model menghasilkan representasi fitur yang kurang optimal untuk tipe kulit yang lebih gelap. Hal ini terlihat dari menurunnya *confidence score* kelas benign pada FST 1 sebesar 32,9% menjadi 23,5% pada FST 6 yang berarti model semakin tidak yakin dalam mengidentifikasi lesi jinak pada kulit gelap.

TABEL VI
TABEL DISTRIBUSI PREDIKSI (%)

Label FST	Benign	Malignant	Non-neoplastic
1	26,2	64,0	9,8
2	30,3	61,3	8,5
3	31,3	57,6	11,1
4	28,2	58,1	13,7
5	22,9	65,2	11,9
6	17,6	75,6	6,8



Gambar 8. Distribusi Prediksi per Fitzpatrick Skin Type (%)



Gambar 9. Rata-rata Confidence Score per Fitzpatrick Skin Type (%)

Oleh karena itu, pengembangan sistem klasifikasi penyakit kulit dengan *deep learning* perlu mempertimbangkan keberagaman warna kulit entah dalam pengumpulan data maupun pelatihan model supaya model dapatt diterapkan pada seluruh populasi pasien.

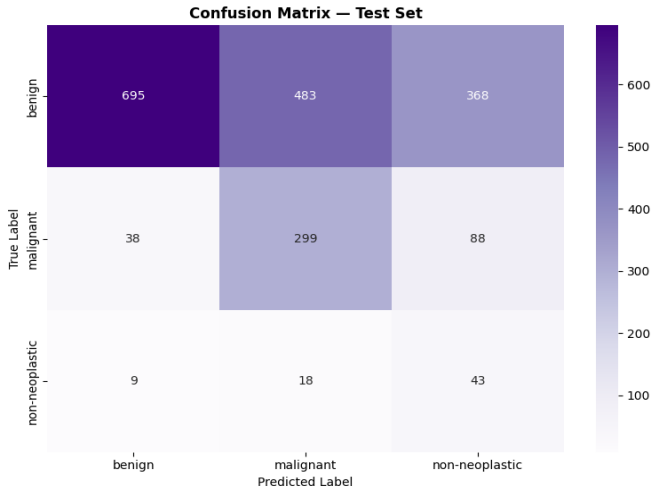
Model dikaji ulang dengan menerapkan mitigasi bias yang berhasil meningkatkan kemampuan model dalam mengenali kelas minoritas. *Pre-processing* berupa *rebalancing dataset* dan *in-processing* berupa *focal* loss menunjukkan performa yang belum seimbang antar kelasnya. Hasil evaluasi model yang telah dilakukan *rebalancing dataset* dan *focal loss* memperoleh akurasi sebesar 51% yang menandakan lebih dari sebagian data berhasil diklasifikasikan dengan benar

sesuai dengan kelasnya. Kelas benign memperoleh nilai *precision* paling besar yakni, sebesar 94% yang menunjukkan model dapat memprediksi kelas benign hampir selalu benar, sedangkan kelas non-neoplastic memperoleh nilai *precision* paling rendah yakni, hanya 9% yang menunjukkan model sering kali memprediksi kelas non-neoplastic dengan keliru.

TABEL VII
CLASSIFICATION REPORT SETELAH MITIGASI BIAS

	Precision	***Recall***	***F1-score***
Benign	0,94	0,45	0,61
Malignant	0,37	0,70	0,49
Non-neoplastic	0,09	0,61	0,15

Gambar 10 menunjukkan hanya 695 data yang berhasil diklasifikasi sebagai kelas benign dari 1.546 data, sebanyak 299 data berhasil dikenali sebagai malignant dari 425 data uji, dan untuk kelas non-neoplastic hanya 43 data yang berhasil diklasifikasi dari 70 data. Kondisi ini menunjukkan banyak sampel data kelas benign yang diprediksikan sebagai kelas malignant maupun kelas non-neoplastic. Terlihat bahwa model cenderung mengorbankan kemampuan memprediksi kelas benign demi meningkatkan sensitivas terhadap kelas malignant.



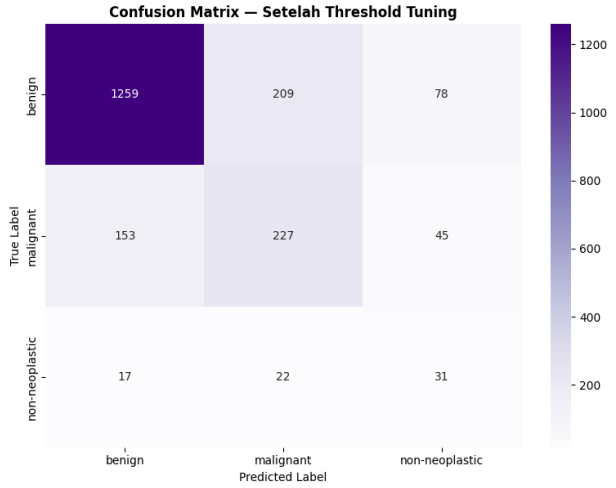
Gambar 10. Confusion Matrix Setelah Mitigasi Bias

Oleh karena kondisi kecenderungan model memperhatikan kelas minoritas, diterapkan *threshold tuning* yang menetapkan nilai batas yang berbeda-beda untuk setiap kelas demi memperoleh keseimbangan antara *precision* dan *recall* di setiap kelas. Perubahan signifikan terlihat pada nilai *precision* kelas non-neoplastic yang meningkat dari yang hanya 9% menjadi 20% yang berarti proporsi prediksi kelas non-neoplastic menjadi lebih baik. Peningkatan ini juga terjadi pada kelas lainnya, seperti meningkatnya nilai *precision* pada kelas malignant yang semula 37% menjadi 50% yang menjadikan peluang prediksi benar pada kelas malignant turut meningkat.

TABEL VIII
CLASSIFICATION REPORT SETELAH THRESHOLD TUNING

	Precision	***Recall***	***F1-score***
Benign	0,88	0,81	0,85
Malignant	0,50	0,53	0,51
Non-neoplastic	0,20	0,44	0,28

Berdasarkan gambar 11, terlihat bahwa performa model meningkat dibandingkan dengan sebelum *threshold tuning*. Model berhasil mengklasifikasikan sebanyak 1.259 data dengan benar pada kelas benign dimana nilai tersebut meningkat sebanyak 564 setelah *threshold tuning*. Untuk kelas malignant sebanyak 227 data berhasil dikenali dengan benar, sementara kelas non-neoplastic terdapat 70 data yang berhasil diklasifikasikan dengan benar. Penurunan nilai ini pada kelas malignant dan non-neoplastic dibandingkan dengan sebelum *threshold tuning* disertai dengan kenaikan nilai *precision* pada masing-masing kelas yang menunjukkan peluang prediksi benar menjadi lebih tinggi.



Gambar 11. Confusion Matrix Setelah Threshold Tuning

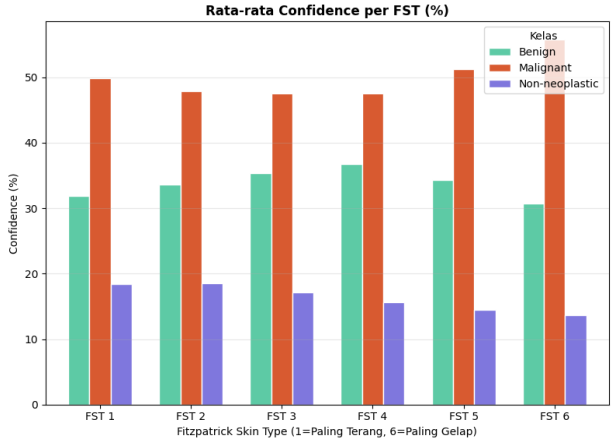
Secara keseluruhan, *threshold tuning* berhasil meningkatkan performa model. Akurasi model meningkat dari 51% menjadi 74%, *weighted F1-score* meningkat dari 0,57 menjadi 0,76, dan *macro F1-score* meningkat dari 0,42 menjadi 0,55. Meningkatnya nilai-nilai tersebut menunjukkan bahwa prediksi antar kelas menjadi lebih seimbang karena jumlah *false positive* berkurang dan akurasi meningkat menjadikan performa model secara keseluruhan menjadi lebih stabil.

Berdasarkan Tabel setelah diterapkan metode mitigasi bias, model masih menunjukkan kecenderungan memprediksi kelas malignant pada tipe kulit yang lebih gelap. Kondisi ini menunjukkan bahwa warna kulit mempengaruhi kecenderungan prediksi model, di mana tipe kulit yang lebih gelap cenderung memprediksi sebagai kelas malignant. Namun, pola bias pada model terlihat lebih seimbang dibandingkan sebelum dilakukan mitigasi bias.

TABEL IX
TABEL DISTRIBUSI PREDIKSI (%) SETELAH MITIGASI BIAS

Label FST	Benign	Malignant	Non-neoplastic
1	38,8	56,3	4,9
2	41,8	53,0	5,2
3	47,1	48,5	4,4
4	45,1	49,8	5,1
5	35,7	58,5	5,7
6	23,7	69,9	6,4

Terlihat pada Gambar 12, rata-rata *confidence* untuk kelas malignant di setiap tipe kulit terus meningkat yang mengindikasikan bahwa model tidak hanya memprediksi malignant lebih sering pada kulit gelap, tetapi juga memprediksi dengan tingkat keyakinan yang lebih tinggi.



Gambar 12. Rata-rata Confidence Score per Fitzpatrick Skin Type (%) Setelah Mitigasi Bias

Setelah mitigasi bias diterapkan, diperoleh nilai *confidence* kelas malignant yang meningkat dibandingkan sebelum mitigasi bias, diperoleh pada FST 1 sebesar 49,8% dan sebesar 55,7% pada FST 6. Dengan selisih *confidence* yang lebih kecil menunjukkan bahwa *threshold tuning* berhasil mengurangi intensitas bias meskipun tidak menghilangkan secara penuh.

IV. KESIMPULAN

Model ResNet50 berbasis *transfer learning* dapat mengklasifikasikan penyakit kulit ke dalam 3 kategori yakni, benign, malignant, dan non-neoplastic dengan dataset HAM100000 hingga mencapai akurasi sebesar 70,85%, presisi 74,03%, dan *recall* sebesar 65,51%. Meskipun performa pada kelas benign selaku kelas mayoritas sudah cukup baik dengan presisi 89%, kesenjangan performa terjadi pada kelas minoritas yakni, kelas non-neoplastic yang nilai presisinya hanya mencapai15%. Hal ini dipengaruhi oleh ketidakseimbangan distribusi kelas dalam dataset HAM100000 terhadap performa model. Dengan diujinya model dengan dataset Fitzpatrick 17k yang memiliki sampel tipe kulit yang lebih bervariatif menunjukkan bahwa adanya bias sistemik yang signifikan pada model. Model cenderung memprediksi sampel sebagai kelas malignant seiring dengan gelapnya warna kulit. Proporsi prediksi malignant meningkat dari 64% pada FST 1 menjadi 75,6% pada FST 6 yang disertai dengan meningkatnya *confidence score* dari 54% menjadi 68,3%. Bias ini disebabkan oleh dataset pelatihan HAM100000 yang didominasi oleh populasi berkulit putih sehingga model belum mampu merepresentasikan fitur visual penyakit kulit pada tipe kulit yang lebih gelap.

Metode mitigasi bias dengan *rebalancing dataset*, *focal loss*, dan *threshold tuning* diterapkan untuk mengurangi dampak bias tersebut. Mitigasi bias terbukti meningkatkan performa klasifikasi dengan akurasi sebesar 74%, *weighted F1-score* sebesar 76%, dan *macro F1-score* sebesar 55%. *Recall* pada masing-masing kelas juga meningkat yakni, 81% pada kelas benign, 53% pada kelas malignant, dan 44% pada kelas non-neoplastic. Nilai *precision* masing-masing kelas juga meningkat dibandingkan dengan sebelum dilakukan mitigasi bias yang menunjukkan bahwa model dapat menghasilkan prediksi yang lebih seimbang antar kelas. Nilai rata-rata *confidence* maglinant juga meningkat menjadi 68,3% hingga 55,7%. Dengan demikian, metode mitigasi bias berhasil memperbaiki keseimbangan klasifikasi model dan mengurangi dominasi kelas mayoritas.

Berdasarkan temuan tersebut, diharapkan penelitian selanjutnya dapat mempertimbangkan penggunaan dataset yang lebih beragam dengan representasi yang lebih merata dari seluruh tipe kulit demi mengurangi bias. Teknik augmentasi data yang mensimulasikan variasi warna kulit juga dapat menjadi alternatif untuk meningkatkan generalisasi model. Serta, evaluasi model berbasis *fairness metrics* dapat diterapkan dalam pengembangan sistem kecerdasan buatan di bidang dermatologi agar teknologi ini memberikan manfaat bagi seluruh kelompok populasi tanpa diskriminasi warna kulit. Selain itu, penerapan mitigasi bias dapat menjadi langkah lanjutan untuk mengurangi bias dan memastikan model dapat memberikan performa yang lebih adil pada berbagai kelompok populasi.

DAFTAR PUSTAKA

[1] A. Moloo, “Recognizing neglected skin diseases: WHO publishes pictorial training guide,” who.int.

[2] D. Li *et al.*, “Worldwide trends and future projections of fungal skin disease burden: a comprehensive analysis from the Global Burden of Diseases study 2021,” *Front. Public Health*, vol. 13, Jun. 2025, doi: 10.3389/fpubh.2025.1580221.

[3] N. Annalakshmi and S. Umarani, “SkinProNet: An attention-based deep learning system for skin disease classification and segmentation,” *Elsevier*, vol. 26, Oct. 2025, doi: 10.1016/j.simpa.2025.100798.

[4] S. Sharma, R. Mittal, N. Goyal, S. B. Goyal, and C. Verma, “Skin disease diagnostics through federated transfer learning on heterogeneous data,” *Sci. Rep.*, vol. 16, Jan. 2026, doi: 10.1038/s41598-025-31730-7.

[5] P. N. Srinivasu, J. G. Sivasai, M. F. Ijaz, A. K. Bhoi, W. Kim, and J. J. Kang, “Classification of Skin Disease Using Deep Learning Neural Networks with MobileNet V2 and LSTM,” *Sensors*, vol. 21, no. 8, Apr. 2021, doi: 10.3390/s21082852.

[6] 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.

[7] M. Groh *et al.*, "Deep learning-aided decision support for diagnosis of skin disease across skin tones," *Nat. Med.*, vol. 30, pp. 573–583, Feb. 2024, doi: 10.1038/s41591-023-02728-3.

[8] K. Nijjer *et al.*, "Adapting Large Language Models to Mitigate Skin Tone Biases in Clinical Dermatology Tasks: A Mixed-Methods Study," *Electrical Engineering and Systems Science*, Oct. 2025, [Online]. Available: http://arxiv.org/abs/2510.00055

[9] K. Mader, "Skin Cancer MNIST: HAM10000," Kaggle.

[10] M. Farabi, "fitzpatrick 17k tonewise splitted," Kaggle.

[11] A. Asriani, N. Lapatta, D. Nugraha, A. Amriana, and W. Wirdayanti, "Implementation of ResNet-50-Based Convolutional Neural Network For Mobile Skin Cancer Classification," *Journal of Applied Informatics and Computing (JAIC)*, vol. 9, no. 4, pp. 1569–1577, Jul. 2025, [Online]. Available: http://jurnal.polibatam.ac.id/index.php/JAIC

[12] F. Ritan and A. Chandra, "Analisis Perbandingan Kinerja Model CNN Resnet-50, VGG19 dan Mobilenet dalam Klasifikasi Penyakit pada Tanaman Mete," *JURNAL LOCUS: Penelitian & Pengabdian*, vol. 4, no. 8, pp. 7903–7918, Aug. 2025, [Online]. Available: https://locus.rivierapublishing.id/index.php/jl

[13] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, IEEE Computer Society, 2016, pp. 770–778. doi: 10.1109/CVPR.2016.90.

[14] W. Xu, Y. L. Fu, and D. Zhu, "ResNet and its application to medical image processing: Research progress and challenges," *Comput. Methods Programs Biomed.*, vol. 240, Jun. 2023, doi: 10.1016/j.cmpb.2023.107660.

[15] A. Sihabillah, A. Tholib, and I. I. Basit, "OPTIMASI Model ResNet50 untuk Klasifikasi Sampah," *INDEXIA: Informatic and Computational Intelligent Journal*, vol. 06, no. 02, pp. 102–111, Apr. 2025.

[16] A. Setiawan, A. ndruru, R. Rosnelly, and A. R. Zai, "Analisis Pengaruh Fine-Tuning pada Model ResNet-50 untuk Deteksi Multikategori Penyakit Mata Berdasarkan Citra Fundus Retina 1," *Remik: Riset dan E-Jurnal Manajemen Informatika Komputer*, vol. 10, no. 1, Mar. 2026, doi: 10.33395/remik.v10i1.15047.

[17] R. Zhang, Y. Zhu, Z. Ge, H. Mu, D. Qi, and H. Ni, "Transfer Learning for Leaf Small Dataset Using Improved ResNet50 Network with Mixed Activation Functions," *Forests*, vol. 13, no. 12, Dec. 2022, doi: 10.3390/f13122072.

[18] N. Kinyanjui *et al.*, "Estimating Skin Tone and Effects on Classification Performance in Dermatology Datasets," Oct. 2019, [Online]. Available: http://arxiv.org/abs/1910.13268

[19] M. Hort, Z. Chen, J. M. Zhang, M. Harman, and F. Sarro, "Bias Mitigation for Machine Learning Classifiers: A Comprehensive Survey," *ACM Journal on Responsible Computing*, vol. 1, no. 2, pp. 1–52, Jun. 2024, doi: 10.1145/3631326.

[20] T. Y. Lin, P. Goyal, R. Girshick, K. He, and P. Dollar, "Focal Loss for Dense Object Detection," in *Proceedings of the IEEE International Conference on Computer Vision*, Institute of Electrical and Electronics Engineers Inc., Dec. 2017, pp. 2999–3007. doi: 10.1109/ICCV.2017.324.

[21] D. Minatel, A. Parmezan, N. Santos, M. Curi, and A. Lopes, "A DIF-Driven Threshold Tuning Method for Improving Group Fairness," in *Proceedings of the ACM Symposium on Applied Computing*, Association for Computing Machinery, Apr. 2025. doi: 10.1145/3672608.3707875.