

Klasifikasi Multi-Kelas Kanker Payudara Berdasarkan Citra Mammogram Menggunakan Arsitektur EfficientNetV2

Helmi Efendi Lubis
Fakultas Informatika
Universitas Telkom
Bandung, Indonesia
helmiel@student.telkomuniversity.ac.id

Untari Novia Wisesty
Fakultas Informatika
Universitas Telkom
Bandung, Indonesia
untarinw@telkomuniversity.ac.id

Abstrak — Kanker payudara merupakan penyebab utama kematian wanita, di mana deteksi dini menggunakan citra *mammogram* bersifat krusial namun rentan kesalahan interpretasi manual akibat kelelahan visual dan kualitas citra. Pendekatan *deep learning* saat ini memiliki keterbatasan berupa kompleksitas komputasi yang tinggi dan umumnya terbatas pada klasifikasi biner. Penelitian ini mengembangkan sistem klasifikasi multikelas menggunakan arsitektur EfficientNetV2 yang efisien untuk mendeteksi lima kondisi, yaitu normal, *benign mass*, *malignant mass*, *benign calcification*, dan *malignant calcification* pada dataset DDSM dan CBIS-DDSM. Pengembangan model dilakukan melalui pembersihan data dan serangkaian skenario percobaan yang mencakup *hyperparameter tuning*, penanganan *class imbalance* dengan *undersampling* dan *augmentasi*, serta peningkatan kontras menggunakan CLAHE. Model dievaluasi menggunakan metrik akurasi, *macro F1-score*, dan *weighted F1-score*. Hasil penelitian menunjukkan bahwa konfigurasi *optimizer AdamW*, *learning rate* $1e-4$, dan resolusi asli citra tanpa manipulasi tambahan menghasilkan performa terbaik dengan akurasi 94,71%, *macro F1-score* 73,87%, dan *weighted F1-score* 94,69%. Temuan ini mengindikasikan bahwa EfficientNetV2 bekerja dengan efektif pada data asli dan teknik penanganan *class imbalance* serta peningkatan kontras justru menurunkan kemampuan generalisasi model.

Kata kunci— kanker payudara, *mammogram*, EfficientNetV2, klasifikasi, *class imbalance*, CLAHE

I. PENDAHULUAN

Kanker payudara merupakan salah satu penyebab utama kematian pada wanita di seluruh dunia akibat pertumbuhan sel abnormal yang tidak terkendali dan berpotensi menyebar ke jaringan lain jika tidak ditangani secara dini [1]. *World Health Organization* (WHO) melaporkan bahwa pada tahun 2022 terdapat 2,3 juta kasus kanker payudara dengan 670.000 kematian secara global [1]. Tingginya angka kematian tersebut menegaskan pentingnya deteksi dini, khususnya pada tahap awal ketika tumor masih berukuran kecil dan lebih mudah ditangani [2].

Mammogram merupakan metode skrining utama yang digunakan dalam deteksi dini kanker payudara [3]. Namun, interpretasi citra *mammogram* secara manual masih

menghadapi berbagai tantangan, seperti kelelahan visual radiolog, kualitas citra yang rendah, serta kompleksitas variasi jaringan payudara [2], [3]. Kondisi ini dapat menyebabkan kesalahan diagnosis, terutama dalam membedakan kanker *benign* dan *malignant*.

Seiring dengan perkembangan teknologi kecerdasan buatan, pendekatan berbasis *deep learning*, khususnya *Convolutional Neural Networks* (CNN) banyak diterapkan dalam sistem *Computer-Aided Diagnosis* (CAD). Sistem ini berperan penting dalam melakukan klasifikasi, segmentasi, dan deteksi pada citra *mammogram* sehingga dapat membantu mengurangi angka kematian akibat kanker payudara [2]. Salah satu pendekatan penting dalam CNN adalah *transfer learning*, yaitu teknik pemanfaatan kembali pengetahuan yang telah dipelajari oleh model dari kasus sebelumnya sebagai titik awal dalam pelatihan model untuk kasus yang berbeda.

Penerapan *transfer learning* menggunakan arsitektur seperti EfficientNetB7, InceptionResNetV2, NASNetLarge, dan VGG16 dalam klasifikasi multikelas kanker payudara telah terbukti mampu menghasilkan akurasi klasifikasi yang tinggi pada citra *mammogram* [2], [3], [4]. Namun, arsitektur tersebut memiliki kompleksitas yang tinggi sehingga membutuhkan sumber daya komputasi yang besar. Kompleksitas ini ditinjau dari jumlah parameter yang dimiliki, di mana EfficientNetB7 memiliki sekitar 66 juta parameter [5], InceptionResNetV2 sekitar 56 juta parameter [5], NASNetLarge sekitar 88,9 juta parameter, dan VGG16 mencapai 138 juta parameter [6], [7]. Selain itu, beberapa penelitian sebelumnya masih berfokus pada klasifikasi biner kanker payudara [8], [9], [10].

Berdasarkan permasalahan tersebut, penelitian ini mengembangkan sistem klasifikasi multikelas kanker payudara menggunakan arsitektur EfficientNetV2 yang lebih efisien dengan jumlah parameter sekitar 24 juta. EfficientNetV2 mengombinasikan blok MBConv dan Fused-MBConv serta menerapkan strategi *progressive learning with adaptive regularization* untuk meningkatkan efisiensi pelatihan dan performa model [11]. Beberapa penelitian terdahulu menunjukkan potensi EfficientNetV2 dalam

klasifikasi kanker payudara, baik pada citra *mammogram* maupun histopatologi [12], [13].

Penelitian ini menangani permasalahan *class imbalance* pada dataset melalui penerapan teknik *undersampling* dan augmentasi data, serta mengevaluasi pengaruh peningkatan kontras citra menggunakan *Contrast Limited Adaptive Histogram Equalization* (CLAHE). Klasifikasi dilakukan terhadap lima kelas, yaitu normal, *benign mass*, *malignant mass*, *benign calcification*, dan *malignant calcification*.

Penelitian ini bertujuan untuk mengembangkan sistem klasifikasi multikelas kanker payudara yang efisien menggunakan arsitektur EfficientNetV2 pada dataset *Digital Database for Screening Mammography* (DDSM) dan *Curated Breast Imaging Subset of DDSM* (CBIS-DDSM) [14], serta mengevaluasi pengaruh *hyperparameter tuning*, penanganan *class imbalance*, dan peningkatan kontras citra terhadap performa model. *Hyperparameter tuning* dibatasi pada pemilihan *optimizer*, *learning rate*, dan *image size*, sementara parameter lainnya menggunakan nilai *default* yang telah dirancang agar bekerja secara *robust* dalam berbagai masalah [15], [16]. Dengan demikian, penelitian ini diharapkan dapat memberikan kontribusi dalam mendukung proses diagnosis dini kanker payudara secara otomatis dan efisien.

II. KAJIAN TEORI

A. Literatur Review

Klasifikasi kanker payudara pada citra *mammogram* menggunakan pendekatan *transfer learning* telah banyak dikembangkan dalam penelitian sebelumnya. Rani dkk. menerapkan arsitektur VGG16 pada dataset DDSM dan UPMC dengan tahapan *preprocessing* berupa *wiener filter*, CLAHE, serta segmentasi menggunakan *binary thresholding*. Model berhasil memperoleh akurasi sebesar 92%–95% untuk klasifikasi empat kelas. Meskipun menghasilkan performa yang baik, arsitektur VGG16 memiliki jumlah parameter yang besar sehingga membutuhkan sumber daya komputasi yang tinggi [2].

Boudouh dan Bouakkaz mengusulkan pendekatan klasifikasi dua tahap menggunakan *dual branch CNN* dengan kombinasi EfficientNetB7 dan InceptionResNetV2, serta memperoleh akurasi sebesar 95,86% pada dataset DDSM dan CBIS-DDSM. Namun, pendekatan tersebut terlalu kompleks dan memerlukan komputasi yang besar [3]. Tiryaki mengevaluasi berbagai arsitektur *deep learning* seperti NASNetLarge, EfficientNetB7, ResNet50, Xception, dan NASNetMobile dengan skema *fine-tuning* bertahap. NASNetLarge memberikan performa terbaik dengan nilai akurasi 0,7318 dan AUC 0,9404 pada klasifikasi lima kelas, tetapi memiliki kebutuhan komputasi yang tinggi [4].

Beberapa penelitian lainnya masih berfokus pada klasifikasi biner. Alruwaili dan Gouda mengembangkan arsitektur MOD-RES berbasis ResNet50 untuk klasifikasi *benign* dan *malignant* pada dataset MIAS dengan akurasi sebesar 89,5% [8]. Oyebo dkk. membandingkan EfficientNetB0, ResNet50, DenseNet121, dan VGG16 pada dataset DDSM dan memperoleh akurasi terbaik sebesar 95,23% menggunakan EfficientNetB0, namun masih terbatas pada dua kelas [9]. Ansar dkk. menggunakan MobileNet untuk membedakan *benign mass* dan *malignant mass* dengan menerapkan *preprocessing* berupa *median filter* dan CLAHE.

Model memiliki akurasi 86,8% pada dataset DDSM dan 74,5% pada CBIS-DDSM [10].

Berdasarkan literatur *review*, sebagian besar penelitian sebelumnya masih menghadapi keterbatasan berupa kompleksitas komputasi yang tinggi serta fokus pada klasifikasi biner. Oleh karena itu, penelitian ini mengusulkan penggunaan arsitektur EfficientNetV2 yang lebih efisien untuk melakukan klasifikasi multikelas kanker payudara pada citra *mammogram*.

B. Kanker Payudara

Kanker payudara merupakan jenis kanker yang paling umum terjadi pada wanita dan disebabkan oleh pertumbuhan sel abnormal pada jaringan payudara. Penyakit ini bersifat multifaktorial dan dipengaruhi oleh faktor usia, genetik, serta gaya hidup. WHO menekankan bahwa deteksi dini dan diagnosis yang akurat berperan penting dalam menurunkan angka kematian akibat kanker payudara [1].

C. Optimizer

Optimizer merupakan metode yang digunakan untuk memperbarui parameter bobot pada *neural network* dengan tujuan meminimalkan *loss function* [15]. Penelitian ini mengevaluasi beberapa jenis *optimizer*, yaitu *Stochastic Gradient Descent* (SGD), RMSprop, Adam, dan AdamW. SGD memperbarui parameter berdasarkan gradien terkini dan sangat bergantung pada pemilihan *learning rate*, sehingga sensitif terhadap fluktuasi gradien [17], [18]. RMSprop mengatasi permasalahan gradien tidak stabil dengan membagi gradien menggunakan akar rata-rata kuadrat gradien sebelumnya sehingga menghasilkan pembaruan parameter yang lebih stabil [16].

Adam menggabungkan keunggulan AdaGrad dan RMSprop dengan menghitung estimasi momen tingkat pertama dan kedua dari gradien secara adaptif [15]. AdamW merupakan pengembangan dari Adam yang memisahkan mekanisme *weight decay* dari pembaruan gradien sehingga mampu meningkatkan kemampuan generalisasi model dan memberikan fleksibilitas dalam pengaturan parameter [18].

D. Class Imbalance

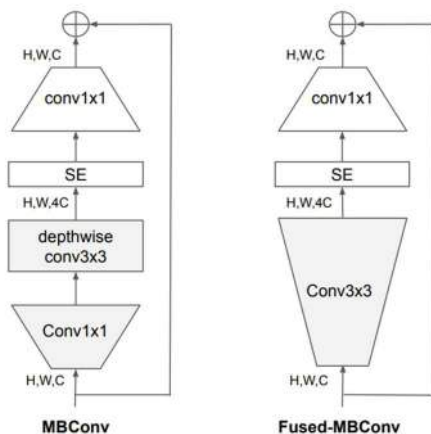
Class imbalance merupakan kondisi ketika distribusi kelas dalam dataset tidak seimbang, di mana kelas mayoritas mendominasi jumlah sampel dibandingkan kelas minoritas [19]. Kondisi ini dapat menyebabkan model bias terhadap kelas mayoritas dan mengabaikan kelas minoritas [20]. Untuk mengatasi permasalahan tersebut, penelitian ini menerapkan teknik *undersampling* pada kelas mayoritas dan augmentasi data untuk meningkatkan variasi data pada kelas minoritas sehingga kemampuan generalisasi model dapat ditingkatkan [19], [20].

E. Contrast Limited Adaptive Histogram Equalization

CLAHE merupakan teknik peningkatan kontras citra yang bertujuan untuk meningkatkan kontras lokal citra sehingga struktur jaringan payudara dapat terlihat lebih jelas tanpa memperkuat *noise* secara berlebihan [21]. Beberapa penelitian menerapkan CLAHE dalam *preprocessing* citra *mammogram* untuk meningkatkan visibilitas detail jaringan payudara [2], [10].

F. EfficientNetV2

EfficientNetV2 merupakan pengembangan dari arsitektur EfficientNet yang dirancang untuk meningkatkan efisiensi pelatihan dan penggunaan parameter. Arsitektur ini ditemukan melalui pendekatan *training-aware Neural Architecture Search* (NAS) dan memperkenalkan blok Fused-MBConv untuk mempercepat proses pelatihan pada lapisan awal jaringan seperti yang terlihat pada Gambar 1. Selain itu, EfficientNetV2 menerapkan strategi *progressive learning with adaptive regularization* yang memungkinkan model mempelajari representasi fitur secara bertahap. Pendekatan ini memungkinkan model mencapai kinerja yang tinggi dengan waktu pelatihan yang lebih singkat dan jumlah parameter yang lebih sedikit dibandingkan arsitektur sebelumnya [11].



GAMBAR 1
STRUKTUR MBCONV DAN FUSED-MBCONV [11]

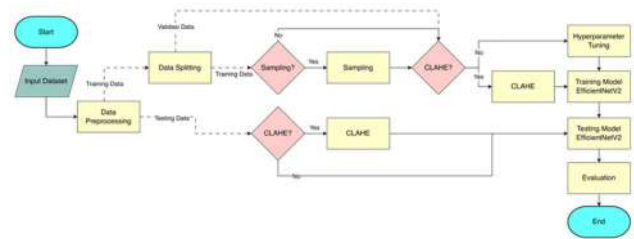
Desain spesifik dari model EfficientNetV2, termasuk susunan blok Fused-MBConv dan MBConv di setiap layer dapat dilihat secara rinci pada Tabel 1.

TABEL 1
ARSITEKTUR EFFICIENTNETV2S [11]

Stage	Operator	Stride	Channels	Layers
0	Conv 3×3	2	24	1
1	Fused-MBConv1, k3×3	1	24	2
2	Fused-MBConv4, k3×3	2	48	4
3	Fused-MBConv4, k3×3	2	64	4
4	MBConv4, k3×3, SE0.25	2	128	6
5	MBConv6, k3×3, SE0.25	1	160	9
6	MBConv6, k3×3, SE0.25	2	256	15
7	Conv 1×1 & Pooling & FC	-	1280	1

III. METODE

Penelitian ini merancang sistem klasifikasi multikelas kanker payudara pada citra *mammogram* menggunakan arsitektur EfficientNetV2. Alur perancangan model dimulai dari pengumpulan dataset, *preprocessing*, penanganan *class imbalance*, peningkatan kontras dengan CLAHE, pelatihan model, *hyperparameter tuning*, hingga evaluasi performa model. Diagram alur model berupa *flowchart* ditunjukkan pada Gambar 2.



GAMBAR 2
FLOWCHART SISTEM KLASIFIKASI KANKER PAYUDARA

A. Dataset

Dataset yang digunakan merupakan gabungan data dari DDSM dan CBIS-DDSM [14]. DDSM adalah dataset *mammogram* yang mencakup 2.620 kasus yang terdiri dari kasus normal, *benign without callback*, *benign*, dan *malignant* [22]. CBIS-DDSM merupakan versi perbaruan yang berisi 753 kasus *calcification* dan 891 kasus *mass* [23]. Data yang digunakan pada penelitian ini disimpan dalam format TFRecords dan dibagi menjadi data *training*, *validation*, dan *testing*. Namun, file *validation* hanya berisi kasus *calcification* dan *testing* hanya berisi kasus *mass* [14] sehingga penelitian ini menggabungkan keduanya menjadi data *testing* lengkap dan membagi data *training* dan *validation* dengan rasio 80:20. Data dikategorikan menjadi lima kelas, yaitu normal, *Benign Calcification* (BC), *Benign Mass* (BM), *Malignant Calcification* (MC), *Malignant Mass* (MM). Sampel data citra tiap kelas dapat dilihat pada Gambar 3.



GAMBAR 3
SAMPEL DATA TIAP KELAS

Tabel 2 menunjukkan jumlah citra tiap kelas kanker payudara sebelum *preprocessing*.

TABEL 2
DISTRIBUSI DATA SEBELUM *PREPROCESSING*

Class	Training	Testing
Normal	48596	13360
BC	2103	558
BM	1911	642
MC	1463	369
MM	1812	435
Total	55885	15364

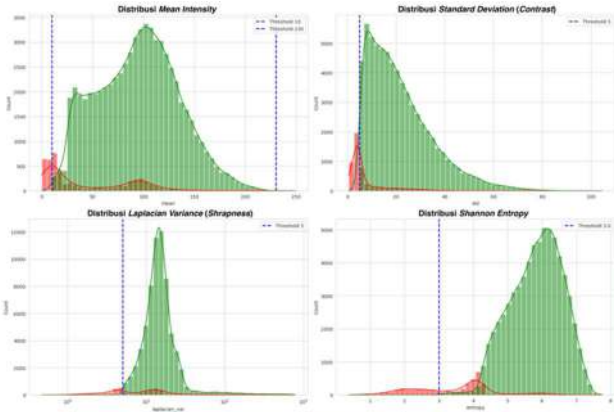
B. Data Preprocessing

Kualitas data memiliki pengaruh signifikan terhadap kinerja model *deep learning*. Berdasarkan hasil pemeriksaan kualitas citra pada dataset, ditemukan citra sampah berupa citra yang dominan berwarna hitam dan citra tulisan seperti yang terlihat pada Gambar 4.



GAMBAR 4
CITRA SAMPAH PADA DATASET

Citra tersebut dapat menimbulkan kesalahan dalam proses ekstraksi fitur yang dihasilkan oleh model. Pembersihan data dilakukan menggunakan beberapa metrik untuk memisahkan data *outlier* berdasarkan karakteristik intensitas [24], ketajaman [25], dan kandungan informasi citra [26] dengan menetapkan *threshold* yang ditentukan melalui distribusi histogram yang terlihat pada Gambar 5.



GAMBAR 5
DISTRIBUSI HISTOGRAM KUALITAS CITRA MAMMOGRAM

Selanjutnya dilakukan pembagian data *training* dan *validation* dengan rasio 80:20. Tabel 3 menunjukkan jumlah citra tiap kelas setelah dilakukan pembersihan data dan sebelum proses *sampling*. Pada *training* model, data dimuat dalam format *grayscale* dengan ukuran 299×299 piksel sesuai dengan kondisi asli dataset. Setiap citra kemudian dipastikan dalam skala [0, 255] agar dapat dilakukan normalisasi secara otomatis oleh model. Kemudian citra diubah dari *grayscale* (1 *channel*) menjadi RGB (3 *channel*) agar menyesuaikan dengan kebutuhan *transfer learning* arsitektur EfficientNetV2.

TABEL 3
DISTRIBUSI DATA SEBELUM PROSES *SAMPLING*

Class	Training	Validation	Testing
Normal	31053	7764	10400
BC	1674	419	558
BM	1526	382	639
MC	1163	290	368
MM	1445	361	435
Total	36861	9216	12400

C. Undersampling dan Augmentasi

Teknik *undersampling* dan augmentasi diterapkan untuk mengatasi permasalahan *class imbalance* yang ditunjukkan pada Tabel 3.

TABEL 4
DISTRIBUSI DATA *TRAINING* SETELAH PROSES *SAMPLING*

Class	Sebelum	Setelah	Rasio Perubahan
Normal	31053	10000	-0,32x
BC	1674	8000	+4,78x
BM	1526	8000	+5,24x
MC	1163	8000	+6,88x
MM	1445	8000	+5,54x

Kelas mayoritas (normal) secara acak dikurangi hingga 10.000 sampel dan augmentasi dilakukan untuk tiap kelas minoritas hingga 8.000 sampel tiap kelas. Augmentasi dilakukan dengan teknik transformasi geometrik seperti *flip*, *rotate*, *zoom*, dan *translation*. Tabel 4 menunjukkan perubahan jumlah data *training* sebelum dan sesudah proses *undersampling* dan augmentasi.

D. Hyperparameter Tuning

Hyperparameter tuning dilakukan untuk menemukan parameter terbaik untuk model EfficientNetV2. Tahapan ini dilakukan pada data *training* tanpa proses *sampling* dan tanpa penerapan CLAHE. Pada penelitian ini, *hyperparameter* yang diuji meliputi:

1. Optimizer

Optimizer adalah algoritma yang berfungsi untuk mengubah bobot model selama proses pelatihan. *Optimizer* bertujuan untuk meminimalkan fungsi *loss*. *Optimizer* yang diuji dalam penelitian ini adalah Adam, AdamW, RMSprop, dan SGD.

2. Learning Rate

Learning rate merupakan nilai yang mempengaruhi seberapa besar perubahan yang dilakukan pada setiap *optimizer* melakukan perubahan bobot. Eksperimen dilakukan dengan nilai *learning rate* 1e-3, 1e-4, dan 1e-5 untuk setiap *optimizer* serta melihat pengaruhnya terhadap konvergensi model.

3. Image Size

Image size merupakan ukuran citra yang digunakan untuk pelatihan model. Penentuan ukuran citra penting dilakukan untuk melihat pengaruh berbagai ukuran citra terhadap performa model. Eksperimen dilakukan dengan mencoba ukuran citra 224×224, 256×256, 299×299, dan 384×384.

E. Contrast Limited Adaptive Histogram Equalization

CLAHE adalah teknik yang digunakan untuk meningkatkan kontras lokal secara adaptif pada citra dengan memanfaatkan proses *equalization* dan pemotongan histogram.

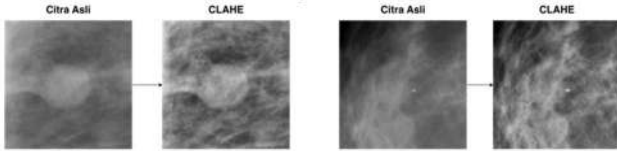
```

Algoritma 1. Peningkatan Kontras Citra
Menggunakan CLAHE
Input: Citra mammogram
Output: Citra mammogram CLAHE
1: IMG_SIZE ← 299
2: clahe ← cv2.createCLAHE(clipLimit=2.0,
tileGridSize=(8,8))
3: function apply_clahe(image):
4:   image ← image.reshape(IMG_SIZE, IMG_SIZE)
5:   return clahe.apply(image)
6: end function

```

Algoritma 1 di atas menunjukkan penerapan CLAHE yang diimplementasikan dengan memanfaatkan *library* OpenCV. Parameter *tileGridSize* berfungsi untuk menentukan berapa banyak *tile* yang dipakai untuk proses *equalization* lokal. Nilai *tileGridSize*=(8,8) berarti citra dipecah menjadi 8×8 petak dan histogram tiap petak diseimbangkan secara terpisah, lalu hasilnya digabung melalui proses *bilinear interpolation* untuk menghindari batas *tile* yang terlihat. Parameter *clipLimit* digunakan untuk mengatur batas pemotongan histogram (*clipping*) sebelum *equalization* sehingga puncak histogram tidak terlalu tinggi. Nilai *clipLimit*=2.0 berarti nilai maksimum puncak histogram lokal dibatasi hingga 2 kali dari tinggi rata-rata histogram. Nilai pada kedua parameter tersebut juga diterapkan pada penelitian sebelumnya yang menggunakan teknik CLAHE

pada *preprocessing* citra [2]. Perbandingan hasil peningkatan kontras citra *mammogram* dapat dilihat pada Gambar 6.



GAMBAR 6
HASIL PENERAPAN CLAHE

F. Pelatihan Model

Model klasifikasi yang digunakan dalam penelitian ini dirancang untuk mengklasifikasikan beberapa jenis kanker payudara berdasarkan citra *mammogram*. Gambar 7 menampilkan rancangan model untuk klasifikasi kanker payudara. Citra melewati arsitektur EfficientNetV2 dengan bobot *pre-trained* dari ImageNet yang digunakan dalam tahap awal pelatihan. Keluaran *feature map* dari EfficientNetV2 kemudian diringkas melalui *global average pooling layer*, diikuti *dropout* untuk meningkatkan generalisasi, lalu *dense* dan *dropout* tambahan. Terakhir, *dense* dengan fungsi aktivasi *softmax* digunakan untuk menghasilkan distribusi probabilitas lima kelas. Proses pelatihan dilakukan dengan melakukan *fine-tuning* keseluruhan layer karena bobot ImageNet dilatih dengan jenis citra yang berbeda.



GAMBAR 7
RANCANGAN MODEL PREDIKSI KANKER PAYUDARA

G. Skenario Penelitian

Penelitian ini melakukan eksperimen dengan tiga skenario untuk mengevaluasi performa model EfficientNetV2 dalam mendeteksi kanker payudara. Skenario dilakukan untuk memahami pengaruh dari setiap percobaan yang dilakukan sehingga ditemukan kombinasi dengan performansi terbaik untuk mengklasifikasikan kanker payudara. Skenario penelitian yang dilakukan meliputi:

1. Hyperparameter Tuning

Skenario ini dilakukan pada data *training* asli dan tanpa penerapan CLAHE untuk menemukan kombinasi *optimizer*, *learning rate*, dan *image size* yang optimal. Percobaan selanjutnya akan menggunakan kombinasi *optimizer*, *learning rate*, dan *image size* yang dihasilkan dari data tersebut.

2. Undersampling dan Augmentasi

Melakukan eksperimen dengan menerapkan proses *undersampling* dan augmentasi data untuk mengatasi permasalahan *class imbalance* serta mengevaluasi pengaruhnya terhadap performa model. Hal itu dilakukan agar model tidak bias pada kelas mayoritas dan mengabaikan kelas minoritas.

3. Contrast Limited Adaptive Histogram Equalization

Menerapkan CLAHE untuk meningkatkan kontras citra dan mengevaluasi pengaruhnya terhadap performa model. Hal itu dilakukan agar model lebih mudah belajar dengan kualitas citra yang telah disesuaikan.

H. Evaluasi Model

Pada tahap evaluasi, model akan diuji dengan data *testing* yang belum pernah dipelajari oleh model sebelumnya. *Confusion matrix* digunakan untuk mengevaluasi kinerja model berdasarkan nilai *True Positive* (TP), *False Negative* (FN), *False Positive* (FP), dan *True Negative* (TN). Bentuk *confusion matrix* dapat dilihat pada Gambar 8.

		Nilai Prediksi	
		Positif	Negatif
Nilai Aktual	Positif	TP	FN
	Negatif	FP	TN

GAMBAR 8
CONFUSION MATRIX

Performa model dievaluasi dengan metrik akurasi, *macro F1-score*, dan *weighted F1-score*.

1. Akurasi mengukur proporsi prediksi yang benar dari seluruh total prediksi [27]. Akurasi dihitung menggunakan Persamaan (1).

$$\text{Akurasi} = \frac{TP + TN}{TP + TN + FP + FN} \quad (1)$$

2. *Precision* mengukur proporsi prediksi positif yang benar dari total prediksi positif yang dibuat oleh model [27]. *Precision* dihitung menggunakan Persamaan (2).

$$\text{Precision} = \frac{TP}{TP + FP} \quad (2)$$

3. *Recall* mengukur kemampuan model dalam mengidentifikasi semua kasus positif yang sebenarnya [27]. *Recall* dihitung menggunakan Persamaan (3).

$$\text{Recall} = \frac{TP}{TP + FN} \quad (3)$$

4. *F1-score* merupakan rata-rata harmonik antara *precision* dan *recall* [27]. *F1-score* dihitung menggunakan Persamaan (4).

$$\text{F1-score} = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \quad (4)$$

5. *Macro F1-score* merupakan rata-rata *F1-score* dari setiap kelas dengan memberikan bobot yang sama untuk tiap kelas [28]. *Macro F1-score* dihitung menggunakan Persamaan (5).

$$\text{Macro F1-score} = \frac{1}{N} \sum_{i=1}^N \frac{2 \times \text{Precision}_i \times \text{Recall}_i}{\text{Precision}_i + \text{Recall}_i} \quad (5)$$

6. *Weighted F1-score* merupakan rata-rata *F1-score* setiap kelas dengan bobot sesuai jumlah sampel tiap kelas [28]. *Weighted F1-score* dihitung menggunakan Persamaan (6).

$$\text{Weighted F1-score} = \sum_{i=1}^N w_i \frac{2 \times \text{Precision}_i \times \text{Recall}_i}{\text{Precision}_i + \text{Recall}_i} \quad (6)$$

$$w_i = \frac{TP_i + FN_i}{\sum_{j=1}^N TP_j + FN_j}$$

IV. HASIL DAN PEMBAHASAN

Penelitian ini melakukan tiga skenario percobaan untuk mengevaluasi kinerja model EfficientNetV2 dalam

mengklasifikasikan lima kelas kanker payudara pada citra *mammogram*. Ketiga skenario meliputi *hyperparameter tuning*, *undersampling* dan augmentasi, serta peningkatan kontras dengan CLAHE. Setiap skenario diimplementasikan selama 50 *epoch* diikuti penerapan *early stopping* dengan batas *patience* 10 *epoch* dan penurunan *learning rate* sebesar 0,2 dengan batas *patience* 4 *epoch*. Evaluasi dilakukan menggunakan metrik akurasi, *macro F1-score*, dan *weighted F1-score*.

Skenario pertama dirancang untuk menemukan kombinasi parameter terbaik model dengan pengujian *hyperparameter tuning* pada performansi data *validation*. Percobaan dilakukan pada data *training* asli dan tanpa CLAHE. *Hyperparameter* yang diuji adalah *optimizer*, *learning rate*, dan *image size*. Evaluasi kinerja model diawali dengan pengujian kombinasi jenis *optimizer* dan *learning rate* menggunakan ukuran asli citra *mammogram*. Hasil pengujian *optimizer* dan *learning rate* ditunjukkan pada Tabel 5.

TABEL 5
HASIL PENGUJIAN OPTIMIZER DAN LEARNING RATE PADA PERFORMANSI VALIDATION

No	Optimizer	Learning Rate	Akurasi	Macro F1-Score	Weighted F1-Score
1	Adam	1e-3	93,96%	70,56%	93,71%
2	AdamW	1e-3	94,40%	72,84%	94,35%
3	RMSprop	1e-3	94,52%	73,90%	94,47%
4	SGD	1e-3	94,50%	73,51%	94,44%
5	Adam	1e-4	96,51%	83,14%	96,52%
6	AdamW	1e-4	96,66%	83,93%	96,66%
7	RMSprop	1e-4	96,33%	81,95%	96,31%
8	SGD	1e-4	87,90%	38,99%	85,40%
9	Adam	1e-5	96,31%	82,42%	96,33%
10	AdamW	1e-5	96,30%	82,18%	96,30%
11	RMSprop	1e-5	95,38%	77,84%	95,38%
12	SGD	1e-5	84,24%	18,29%	77,04%

Berdasarkan hasil pada Tabel 5, model yang dilatih menggunakan kombinasi *optimizer* AdamW dan *learning rate* 1e-4 menghasilkan performa terbaik dengan akurasi 96,66%, *macro F1-score* 83,93%, dan *weighted F1-score* 96,66%. Penurunan *learning rate* menjadi 1e-5 pada *optimizer* adaptif (Adam, AdamW, RMSprop) menurunkan performa karena pembaruan bobot menjadi terlalu lambat. Sebaliknya, *optimizer* SGD menunjukkan penurunan performa signifikan seiring mengecilnya *learning rate* yang mengindikasikan karakteristik sensitif SGD terhadap parameter tersebut [17]. Percobaan selanjutnya adalah optimasi *image size*.

Berdasarkan Tabel 6, ukuran citra 299×299 menghasilkan performa terbaik dengan nilai akurasi, *macro F1-score*, dan *weighted F1-score* sebesar 96,66%, 83,93%, dan 96,66%. Penurunan performa pada ukuran citra 224×224 dan 256×256 terjadi karena citra kehilangan detail tekstur penting, sementara ukuran 384×384 tidak memberikan peningkatan performa dibandingkan ukuran 299×299 karena memperjelas fitur lain seperti *noise* yang mengganggu proses ekstraksi fitur. Dengan demikian, konfigurasi akhir yang digunakan dalam

percobaan selanjutnya adalah *optimizer* AdamW, *learning rate* 1e-4, dan *image size* 299×299.

TABEL 6
HASIL PENGUJIAN IMAGE SIZE PADA PERFORMANSI VALIDATION

No	Image Size	Akurasi	Macro F1-Score	Weighted F1-Score
1	224×224	96,00%	81,79%	95,92%
2	256×256	95,88%	81,40%	95,87%
3	299×299	96,66%	83,93%	96,66%
4	384×384	96,42%	82,82%	96,43%

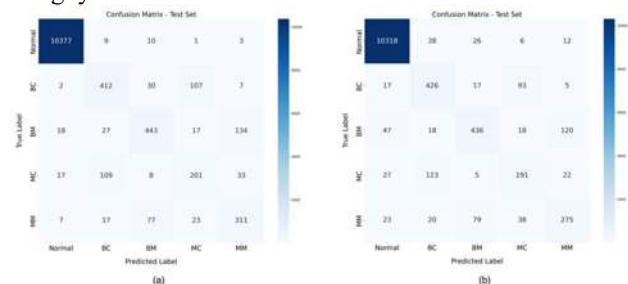
Skenario kedua bertujuan untuk mengevaluasi pengaruh *undersampling* dan augmentasi pada data *training* terhadap performa model dalam mendeteksi kanker payudara. Evaluasi dilakukan dengan membandingkan model yang dilatih menggunakan data *training* asli (tanpa *sampling*) dan data *training sampling*.

TABEL 7
HASIL PENGUJIAN SKENARIO UNDERSAMPLING DAN AUGMENTASI DATA TRAINING PADA PERFORMANSI VALIDATION DAN TESTING

Pengujian	Validation			Testing		
	Akurasi	Macro F1-Score	Weighted F1-Score	Akurasi	Macro F1-Score	Weighted F1-Score
Sampling	97,97%	91,93%	97,97%	93,92%	72,08%	93,87%
Tanpa Sampling	96,66%	83,93%	96,66%	94,71%	73,87%	94,69%

Berdasarkan hasil pada Tabel 7, model yang dilatih menggunakan data *training* asli memiliki performa yang lebih baik dibandingkan model yang dilatih menggunakan data *training sampling* pada performansi data *testing*. Meskipun model dengan skema *sampling* memiliki metrik *validation* yang sangat tinggi dengan *macro F1-score* 91,93%, performanya mengalami penurunan drastis saat diuji pada data *testing* menjadi 72,08%. Sebaliknya, model dengan data asli menunjukkan konsistensi yang lebih baik dengan *macro F1-score* 83,93% pada data *validation* dan 73,87% pada data *testing*. Hasil tersebut menunjukkan bahwa proses penyeimbangan data justru membuat model terlalu menghafal pola augmentasi (*overfitting*) sehingga menurunkan performa model.

Penyebab *overfitting* tersebut dapat ditinjau karena rasio perubahan seperti yang terlihat pada Tabel 4. Kelas minoritas mengalami peningkatan jumlah data dengan rasio di atas 4,5 kali lipat. Rasio ini mengindikasikan bahwa data yang dihasilkan dari proses augmentasi menyebabkan model terlalu menghafal pola hasil augmentasi yang memiliki kemiripan visual yang tinggi. Selain itu, proses *undersampling* pada kelas normal juga mengurangi performa model dalam membedakan kelas normal dan kanker akibat hilangnya variasi data normal.



GAMBAR 9

CONFUSION MATRIX (A) DATA TRAINING ASLI (B) DATA TRAINING SAMPLING

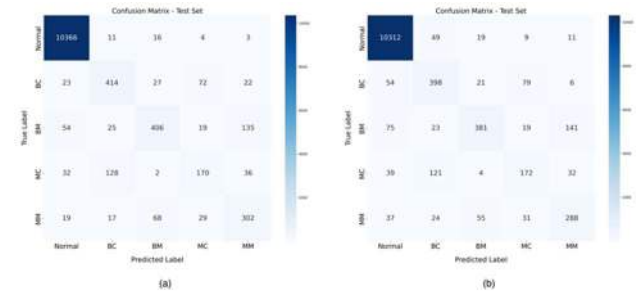
Berdasarkan *confusion matrix* pada Gambar 9, model yang dilatih dengan data *training sampling* menunjukkan penurunan performa pada kasus kanker dibandingkan model yang dilatih dengan data *training* asli. Pada kasus *malignant*, model yang dilatih menggunakan data *training sampling* salah memprediksi 123 kasus *malignant calcification* sebagai *benign calcification* dan 79 kasus *malignant mass* sebagai *benign mass*. Angka tersebut lebih tinggi dibandingkan model yang dilatih dengan data *training* asli, di mana model salah memprediksi 109 dan 77 pada kasus yang sama. Selain itu, penurunan performa terlihat pada *true positive* yang dihasilkan model yang dilatih pada data *training sampling* lebih rendah dibandingkan model yang dilatih pada data asli. Pada klasifikasi kasus normal, model yang dilatih menggunakan data *training sampling* salah memprediksi 82 kasus normal sebagai kanker (*false negative*) dan juga salah memprediksi 114 kasus kanker sebagai normal (*false positive*). Sedangkan, model yang dilatih menggunakan data *training* asli hanya menghasilkan 23 *false negative* dan 44 *false positive* untuk kasus normal. Dengan demikian, hasil percobaan ini membuktikan bahwa *undersampling* dan augmentasi justru menurunkan kemampuan generalisasi model karena model kehilangan variasi kelas normal dan teknik augmentasi mengakibatkan model terlalu belajar pola augmentasi.

Skenario ketiga dilakukan untuk mengevaluasi pengaruh teknik peningkatan kontras menggunakan CLAHE terhadap performa model. CLAHE diterapkan pada seluruh data (*training*, *validation*, *testing*). Evaluasi dilakukan dengan membandingkan model yang dilatih menggunakan data *training* asli ditambah CLAHE dan model yang dilatih menggunakan data *training sampling* ditambah CLAHE.

TABEL 8
HASIL PENGUJIAN SKENARIO CLAHE PADA PERFORMANSI *VALIDATION*
DAN *TESTING*

Pengujian	Validation			Testing		
	Akurasi	Macro F1-Score	Weighted F1-Score	Akurasi	Macro F1-Score	Weighted F1-Score
Sampling + CLAHE	97,36%	89,62%	97,35%	93,15%	69,68%	92,98%
Tanpa Sampling + CLAHE	95,51%	79,22%	95,44%	94,02%	71,45%	93,86%

Berdasarkan hasil pada Tabel 8, peningkatan kontras citra justru menurunkan performa model dibandingkan hasil percobaan pada Tabel 7. Model yang dilatih menggunakan data *training* asli ditambah CLAHE menunjukkan penurunan akurasi 94,71% menjadi 94,02%, *macro* F1-score 73,87% menjadi 71,45%, dan *weighted* F1-score 94,69% menjadi 93,86%. Pola serupa juga terjadi pada model yang dilatih menggunakan data *training sampling* ditambah CLAHE, di mana akurasi 93,92% menjadi 93,15%, *macro* F1-score 72,08% menjadi 69,68%, dan *weighted* F1-score 93,87% menjadi 92,98%. Hal itu disebabkan karena CLAHE meningkatkan detail citra secara lokal dan menimbulkan *noise* yang dianggap sebagai fitur oleh model sehingga mengganggu proses ekstraksi fitur.



GAMBAR 10
CONFUSION MATRIX (A) DATA *TRAINING* ASLI DENGAN CLAHE (B) DATA
TRAINING SAMPLING DENGAN CLAHE

Berdasarkan *confusion matrix* pada Gambar 10, penerapan CLAHE meningkatkan ambiguitas visual antara kelas *malignant* dan *benign*. Pada model yang dilatih menggunakan data asli ditambah penerapan CLAHE, kesalahan prediksi kasus *malignant calcification* yang dianggap *benign calcification* meningkat menjadi 128 kasus. Jumlah ini lebih tinggi dibandingkan model yang dilatih pada data *training* asli tanpa penerapan CLAHE. Selain itu, pada kasus *malignant mass*, terdapat 68 kasus yang salah diklasifikasikan sebagai *benign mass* dan 19 kasus terdeteksi sebagai normal. Penurunan performa juga dapat dilihat dari *true positive* yang dihasilkan model.

Penurunan performa yang lebih parah terlihat pada model yang dilatih dengan data *training sampling* ditambah CLAHE. Pada kasus normal, terdapat 205 kasus *false positive* dan terdapat 88 kasus *false negative*. Dengan demikian, penerapan CLAHE justru menurunkan performansi model karena *noise* yang mengakibatkan model kesulitan dalam melakukan ekstraksi fitur.

V. KESIMPULAN

Penelitian ini telah berhasil mengembangkan sistem klasifikasi kanker payudara pada citra *mammogram* dengan memanfaatkan arsitektur EfficientNetV2. Berdasarkan evaluasi skenario percobaan yang dilakukan, model EfficientNetV2 mampu bekerja optimal pada data asli tanpa memerlukan teknik untuk mengatasi *class imbalance* ataupun peningkatan kontras dengan CLAHE. Konfigurasi terbaik diperoleh menggunakan *optimizer* AdamW, *learning rate* 1e-4, dan *image size* 299×299 piksel dengan akurasi 94,71%, *macro* F1-score 73,87%, dan *weighted* F1-score 94,69% pada performansi data *testing*.

Secara spesifik, analisis terhadap skenario penanganan data menunjukkan bahwa teknik *undersampling* dan augmentasi serta penerapan CLAHE tidak memberikan peningkatan performa. Teknik *sampling* pada data *training* menyebabkan model *overfitting* terhadap pola augmentasi dan kehilangan variasi pada kelas normal sehingga akurasi turun menjadi 93,92%, *macro* F1-score 72,08%, dan *weighted* F1-score 93,87%. Selain itu, penerapan CLAHE memunculkan *noise* yang meningkatkan ambiguitas antara kelas *malignant* dan *benign* sehingga mengakibatkan penurunan akurasi menjadi 94,02%, *macro* F1-score 71,45%, dan *weighted* F1-score 93,86%. Penurunan performa juga terjadi ketika dikombinasikan dengan teknik *undersampling* dan augmentasi, di mana akurasi menjadi 93,15%, *macro* F1-score 69,68%, dan *weighted* F1-score 92,98%. Dengan demikian, temuan ini mengindikasikan bahwa EfficientNetV2 bekerja secara efektif pada data asli dan teknik penanganan *class imbalance* serta peningkatan kontras justru menurunkan kemampuan generalisasi model.

REFERENSI

- [1] WHO, "Breast cancer," 13-Mar-2024. [Online]. Available: <https://www.who.int/news-room/factsheets/detail/breast-cancer>. [Accessed: 20-Apr-2025].
- [2] N. Rani, D. K. Gupta, and S. Singh, "Multi-class classification of breast cancer abnormality using transfer learning," *Multimedia Tools and Applications*, vol. 83, no. 30, pp. 75085–75100, Feb. 2024, doi: 10.1007/s11042-023-17832-2.
- [3] S. S. Boudouh and M. Bouakkaz, "Breast cancer: new mammography dual-view classification approach based on pre-processing and transfer learning techniques," *Multimedia Tools and Applications*, vol. 83, no. 8, pp. 24315–24337, Aug. 2023, doi: 10.1007/s11042-023-16431-5.
- [4] V. M. Tiriyaki, "Deep Transfer Learning to Classify Mass and Calcification Pathologies from Screen Film Mammograms," *Bitlis Eren Üniversitesi Fen Bilimleri Dergisi*, vol. 12, no. 1, pp. 57–65, Mar. 2023, doi: 10.17798/bitlisfen.1190134.
- [5] M. Tan and Q. Le V., "EfficientNet: Rethinking model scaling for convolutional neural networks," *arXiv.org*, May 28, 2019. <https://arxiv.org/abs/1905.11946>.
- [6] B. Zoph, V. Vasudevan, J. Shlens, and Q. Le V., "Learning transferable architectures for scalable image recognition," *arXiv.org*, Jul. 21, 2017. <https://arxiv.org/abs/1707.07012>.
- [7] K. Simonyan and A. Zisserman, "Very deep convolutional networks for Large-Scale image recognition," *arXiv.org*, Sep. 04, 2014. <https://arxiv.org/abs/1409.1556>.
- [8] M. Alruwaili and W. Gouda, "Automated Breast cancer detection models based on transfer learning," *Sensors*, vol. 22, no. 3, p. 876, Jan. 2022, doi: 10.3390/s22030876.
- [9] N. O. S. Oyeibanji et al., "Enhancing breast cancer detection accuracy through transfer learning: A case study using efficient net," *World Journal of Advanced Engineering Technology and Sciences*, vol. 13, no. 1, pp. 285–318, Sep. 2024, doi: 10.30574/wjaets.2024.13.1.0415.
- [10] W. Ansar, A. R. Shahid, B. Raza, and A. H. Dar, "Breast cancer detection and localization using MobileNet based transfer learning for mammograms," in *Communications in computer and information science*, 2020, pp. 11–21. doi: 10.1007/978-3-030-43364-2_2.
- [11] M. Tan and Q. V. Le, "EfficientNetV2: Smaller models and faster training," *arXiv (Cornell University)*, Apr. 2021, doi: 10.48550/arxiv.2104.00298.
- [12] M. Hasan, "Deep Learning for Breast Cancer Detection: Comparative Analysis of ConvNeXT and EfficientNet," pp. 1387–1391, Dec. 2024, doi: <https://doi.org/10.1109/iccit64611.2024.11021905>.
- [13] O. A. Abioye, A. E. Ewuekpae, and A. J. Awujo, "Performance Evaluation of EfficientNetV2 Models on the Classification of Histopathological Benign Breast Cancer Images," *Science Journal of University of Zakho*, vol. 12, no. 2, pp. 208–214, May 2024, doi: 10.25271/sjuoz.2024.12.2.1261.
- [14] E. A. Scuccimarra, "DDSM Mammography," 2018. [Online]. Available: <https://www.kaggle.com/datasets/skooch/ddsm-mammography/data>. [Accessed: 01-May-2025].
- [15] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," *arXiv (Cornell University)*, Dec. 2014, doi: 10.48550/arxiv.1412.6980.
- [16] M. C. Mukkamala and M. Hein, "Variants of RMSProp and Adagrad with Logarithmic Regret Bounds," *arXiv (Cornell University)*, pp. 2545–2553, Jun. 2017, doi: 10.48550/arxiv.1706.05507.
- [17] A. Brutzkus, A. Globerson, E. Malach, and S. Shalev-Shwartz, "SGD Learns Over-parameterized Networks that Provably Generalize on Linearly Separable Data," *arXiv (Cornell University)*, Oct. 2017, doi: 10.48550/arxiv.1710.10174.
- [18] I. Loshchilov and F. Hutter, "Decoupled weight decay regularization," *arXiv.org*, Nov. 14, 2017. <https://arxiv.org/abs/1711.05101>.
- [19] J. M. Johnson and T. M. Khoshgoftaar, "The Effects of Data Sampling with Deep Learning and Highly Imbalanced Big Data," *Information Systems Frontiers*, vol. 22, no. 5, pp. 1113–1131, Jun. 2020, doi: 10.1007/s10796-020-10022-7.
- [20] S. Yang, W. Xiao, M. Zhang, S. Guo, J. Zhao, and F. Shen, "Image Data augmentation for Deep Learning: A survey," *arXiv (Cornell University)*, Apr. 2022, doi: 10.48550/arxiv.2204.08610.
- [21] K. Zuiderveld, "Contrast limited adaptive histogram equalization," in *Graphics Gems IV*, 1994, pp. 474–485.
- [22] M. Heath, K. Bowyer, D. Kopans, R. Moore, and W. P. Kegelmeyer, "The Digital Database for Screening Mammography," in *Proc. 5th Int. Workshop on Digital Mammography*, M. J. YaZe, Ed. Madison, WI: Medical Physics Publishing, 2001, pp. 212–218, ISBN 1-930524-00-5.
- [23] R. S. Lee, F. Gimenez, A. Hoogi, K. K. Miyake, M. Gorovoy, and D. L. Rubin, "A curated mammography data set for use in computer-aided detection and diagnosis research," *Scientific Data*, vol. 4, no. 1, p. 170177, Dec. 2017, doi: 10.1038/sdata.2017.177.
- [24] R. C. Gonzalez and R. E. Woods, *Digital Image Processing*, 2018.
- [25] J. L. Pech-Pacheco, G. Cristóbal, J. Chamorro-Martínez, and J. Fernández-Valdivia, "Diatom autofocusing in brightfield microscopy: a comparative study," in *Proc. 15th Int. Conf. Pattern Recognit. (ICPR)*, vol. 3, pp. 314–317, Sept. 2000.
- [26] S. Vajapeyam, "Understanding Shannon's Entropy metric for Information," *arXiv.org*, Mar. 24, 2014. <https://arxiv.org/abs/1405.2061>.
- [27] H. M and S. MN, "A review on Evaluation Metrics for Data Classification Evaluations," *International Journal of Data Mining & Knowledge Management Process*, vol. 5, no. 2, pp. 01–11, Mar. 2015, doi: 10.5121/ijdkp.2015.5201.
- [28] M. C. H. Lee, J. Braet, and J. Springael, "Performance Metrics for multilabel emotion Classification: Comparing micro, macro, and weighted F1-Scores," *Applied Sciences*, vol. 14, no. 21, p. 9863, Oct. 2024, doi: 10.3390/app14219863.