

ANALISIS PERBANDINGAN KINERJA ARSITEKTUR CNN, RESNET50, DAN EFFICIENTNET-B0 UNTUK PENGENALAN EKSPRESI WAJAH (MACRO EXPRESSION)

Fitto Ardiansyah, Anggay Luri Pramana

Teknik Informatika, Universitas NU Sidoarjo

Rangkah Kidul, Kec. Sidoarjo, Kabupaten Sidoarjo, Jawa Timur 61234, Indonesia.

fittoardiansyah12@gmail.com

ABSTRAK

Penelitian ini membahas analisis perbandingan kinerja tiga arsitektur deep learning, yaitu Convolutional Neural Network (CNN), ResNet50, dan EfficientNet-B0 dalam tugas pengenalan ekspresi wajah (Facial Expression Recognition/Fer). Permasalahan utama dalam penelitian ini adalah bagaimana menentukan arsitektur yang paling efektif untuk dataset ekspresi wajah beresolusi rendah dengan distribusi kelas yang telah diseimbangkan. Penelitian bertujuan untuk mengevaluasi dan membandingkan performa ketiga model berdasarkan metrik *accuracy*, *precision*, *recall*, dan *F1-score*. Dataset yang digunakan adalah CK+ yang telah melalui tahap preprocessing meliputi *resize*, *normalisasi*, dan *augmentasi data*. Setiap model dilatih menggunakan dua *skenario epoch*, yaitu 50 dan 70 epoch. Hasil penelitian menunjukkan bahwa CNN memberikan performa terbaik dengan akurasi mencapai 0,83, serta menunjukkan stabilitas yang lebih baik dibandingkan ResNet50 mencapai akurasi 0,67 dan EfficientNet-B0 mencapai akurasi 0,50. Sementara itu, ResNet50 menunjukkan peningkatan moderat, dan EfficientNet-B0 memiliki performa terendah akibat keterbatasan dalam generalisasi pada dataset beresolusi rendah. Temuan ini menunjukkan bahwa kompleksitas model tidak selalu berbanding lurus dengan peningkatan performa.

Kata kunci : *cnn, deep learning, efficientnet-b0, pengenalan ekspresi wajah, ResNet50*

1. PENDAHULUAN

Facial Expression Recognition (FER) merupakan salah satu bidang penting dalam computer vision yang berfokus pada identifikasi kondisi emosional manusia melalui citra wajah. Ekspresi wajah menjadi media komunikasi non-verbal yang mengandung informasi signifikan terkait kondisi psikologis individu dan banyak dimanfaatkan dalam berbagai aplikasi seperti human-computer interaction, sistem keamanan, serta analisis perilaku berbasis visual [1]. Namun, kompleksitas variasi ekspresi, perbedaan individu, serta kondisi pencahayaan menjadikan proses klasifikasi ekspresi wajah sebagai permasalahan yang menantang.

Pendekatan tradisional dalam FER umumnya bergantung pada ekstraksi fitur manual seperti Local Binary Pattern (LBP) dan Histogram of Oriented Gradients (HOG). Meskipun metode tersebut cukup efektif pada kondisi tertentu, kemampuannya dalam menangkap representasi fitur yang kompleks masih terbatas. Perkembangan deep learning, khususnya Convolutional Neural Network (CNN), telah membawa peningkatan signifikan dengan kemampuan pembelajaran fitur secara otomatis dan hierarkis langsung dari data citra [2].

Seiring dengan perkembangan tersebut, berbagai arsitektur yang lebih kompleks seperti ResNet50 dan EfficientNet-B0 diperkenalkan untuk meningkatkan performa klasifikasi. ResNet50 mengatasi permasalahan vanishing gradient melalui residual learning [3]. Sedangkan EfficientNet-B0 mengoptimalkan efisiensi model melalui pendekatan

compound scaling [4]. Meskipun demikian, performa relatif dari ketiga arsitektur ini pada dataset ekspresi wajah dengan resolusi rendah masih belum sepenuhnya teridentifikasi secara komprehensif.

Berdasarkan permasalahan tersebut, penelitian ini bertujuan untuk melakukan analisis komparatif terhadap kinerja CNN, ResNet50, dan EfficientNet-B0 dalam klasifikasi ekspresi wajah makro. Pendekatan yang digunakan meliputi preprocessing data, augmentasi, serta evaluasi menggunakan metrik *accuracy*, *precision*, *recall*, dan *F1-score*. Hasil penelitian ini diharapkan dapat memberikan pemahaman yang lebih mendalam terkait efektivitas masing-masing arsitektur dalam konteks dataset FER.

2. TINJAUAN PUSTAKA

2.1. Facial Expression Recognition (FER)

Facial Expression Recognition (FER) merupakan bidang dalam computer vision yang bertujuan untuk mengidentifikasi ekspresi wajah manusia secara otomatis. Ekspresi wajah menjadi salah satu bentuk komunikasi non-verbal yang penting dalam menyampaikan emosi. Secara umum, terdapat tujuh ekspresi dasar yang bersifat universal, yaitu marah (anger), jijik (disgust), takut (fear), sedih (sadness), terkejut (surprise), senang (happiness), dan netral (neutral) [5]. Seiring perkembangan deep learning, metode berbasis Convolutional Neural Network (CNN) menjadi pendekatan utama dalam FER karena kemampuannya dalam mengekstraksi fitur citra secara otomatis dan hierarkis [2].

2.2. Dataset CK+

Dataset Extended Cohn-Kanade (CK+) merupakan salah satu dataset yang luas digunakan dalam penelitian FER. Dataset ini berisi urutan citra wajah dari kondisi netral hingga mencapai puncak ekspresi. Keunggulan CK+ terletak pada anotasi ekspresi yang jelas dan kualitas citra yang baik [6]. Namun, dataset ini memiliki tantangan berupa ketidakseimbangan distribusi kelas, di mana beberapa ekspresi seperti happiness memiliki jumlah data lebih banyak dibandingkan ekspresi lain seperti disgust atau fear. Kondisi ini dapat menyebabkan model cenderung bias terhadap kelas mayoritas [7].

2.3. Pre-processing Citra

Tahap pra-pemrosesan bertujuan untuk meningkatkan kualitas data dan menyesuaikan format input dengan kebutuhan model. Beberapa teknik yang umum digunakan meliputi resize citra, konversi warna, normalisasi, dan augmentasi data. Resize citra dilakukan untuk menyeragamkan dimensi input sesuai dengan arsitektur model [2]. Konversi citra ke grayscale dapat mengurangi kompleksitas komputasi tanpa menghilangkan informasi penting terkait tekstur wajah [8].

Normalisasi digunakan untuk menskalakan nilai piksel ke rentang tertentu, umumnya [0,1], guna meningkatkan stabilitas pelatihan model [9]. Selain itu, augmentasi data seperti rotasi, translasi, dan flipping digunakan untuk meningkatkan variasi data dan mengurangi overfitting [8].

2.4. Convolutional Neural Network (CNN)

Convolutional Neural Network (CNN) merupakan arsitektur deep learning yang dirancang untuk mengolah data berbentuk grid seperti citra. CNN terdiri dari beberapa lapisan utama, yaitu convolutional layer, pooling layer, dan fully connected layer. Lapisan konvolusi berfungsi mengekstraksi fitur lokal seperti tepi dan tekstur, sedangkan pooling layer digunakan untuk mengurangi dimensi fitur [2]. CNN memiliki keunggulan dalam menangkap pola spasial secara efisien melalui mekanisme shared weights dan local connectivity [10].

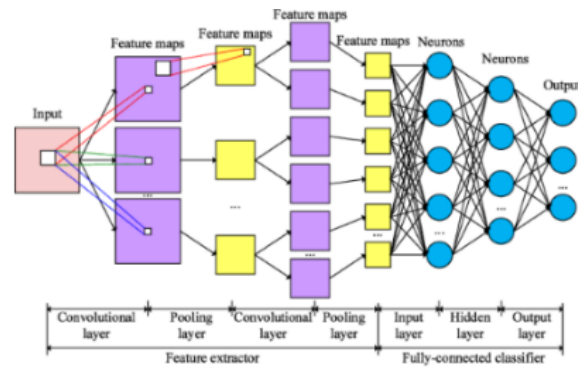
Operasi konvolusi pada CNN dapat dirumuskan sebagai:

$$S(i, j) = \sum_m \sum_n I(m, n) \cdot K(i - m, j - n) \quad (1)$$

di mana f adalah filter dan x adalah citra input. Fungsi aktivasi yang umum digunakan adalah ReLU:

$$f(x) = \max(0, x) \quad (2)$$

Gambar 1 menunjukkan arsitektur CNN yang terdiri dari proses konvolusi, pooling, dan klasifikasi melalui fully connected layer.



Gambar 1. Arsitektur Model ResNet50

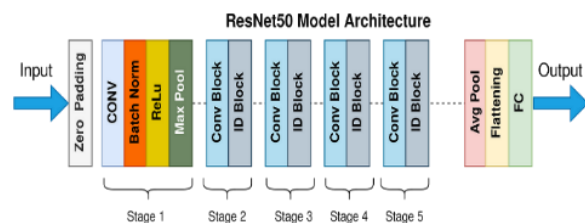
2.5. ResNet50

ResNet50 merupakan arsitektur deep neural network yang diperkenalkan untuk mengatasi permasalahan degradasi performa pada jaringan yang sangat dalam. Model ini menggunakan konsep residual learning dengan menambahkan skip connection yang memungkinkan aliran gradien tetap stabil selama proses pelatihan [11]. Dengan kedalaman 50 lapisan, ResNet50 mampu mengekstraksi fitur yang lebih kompleks dibandingkan CNN konvensional [12].

Secara matematis, *residual learning* dirumuskan sebagai:

$$y = F(x, W) + x \quad (3)$$

di mana x adalah input dan $F(x)$ adalah fungsi residual yang dipelajari oleh jaringan. Gambar 2 memperlihatkan arsitektur ResNet50 dengan mekanisme skip connection yang menghubungkan input langsung ke output layer tertentu.



Gambar 2. Arsitektur Model ResNet50

2.6. EfficientNet-B0

EfficientNet-B0 merupakan arsitektur yang dikembangkan dengan pendekatan compound scaling, yaitu menyeimbangkan skala kedalaman (depth), lebar (width), dan resolusi (resolution) jaringan secara sistematis. Model ini menggunakan blok MBConv yang dikombinasikan dengan mekanisme squeeze-and-excitation untuk meningkatkan representasi fitur [13]. EfficientNet dikenal memiliki performa yang baik dengan jumlah parameter yang relatif lebih sedikit [14].

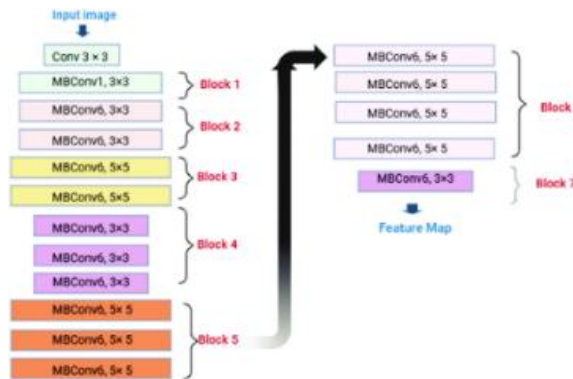
Pendekatan *compound scaling* dirumuskan sebagai:

$$\begin{aligned} \text{depth} &= \alpha^{\phi}(\text{depth}), \\ \text{width} &= \beta^{\phi}(\text{width}), \\ \text{resolution} &= \gamma^{\phi}(\text{resolution})[4] \end{aligned}$$

dengan constraint:

$$\alpha \cdot \beta^2 \cdot \gamma^2 \approx 2, \quad (4)$$

Gambar 3 menunjukkan arsitektur EfficientNet-B0 yang tersusun dari blok MBConv dengan mekanisme squeeze-and-excitation.



Gambar 3. Arsitektur Model *EfficientNet-B0*

2.7. Transfer Learning

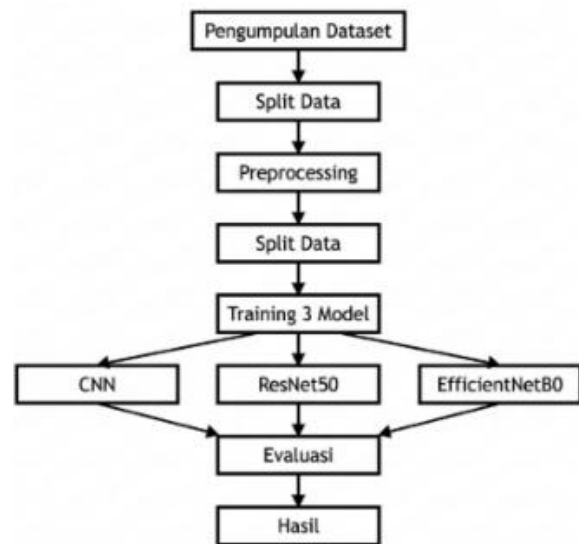
Transfer learning merupakan teknik yang memanfaatkan model yang telah dilatih sebelumnya pada dataset besar, seperti ImageNet, untuk menyelesaikan tugas baru. Pendekatan ini memungkinkan proses pelatihan menjadi lebih cepat dan efisien, terutama ketika dataset yang digunakan terbatas [3]. Dalam praktiknya, transfer learning dapat dilakukan dengan dua cara, yaitu sebagai feature extractor atau dengan melakukan fine-tuning pada beberapa lapisan model [15].

2.8. Evaluasi Model

Evaluasi kinerja model dilakukan menggunakan beberapa metrik, antara lain accuracy, precision, recall, dan F1-score. Accuracy mengukur tingkat ketepatan prediksi secara keseluruhan, sedangkan precision dan recall memberikan informasi terkait performa pada masing-masing kelas [16]. F1-score merupakan rata-rata harmonis antara precision dan recall yang digunakan untuk menangani ketidakseimbangan data [17]. Selain itu, confusion matrix digunakan untuk memvisualisasikan distribusi prediksi model terhadap kelas sebenarnya [4].

3. METODE PENELITIAN

Penelitian ini menerapkan pendekatan eksperimental kuantitatif untuk mengevaluasi efektivitas arsitektur *Deep learning* pada tugas klasifikasi emosi wajah. Tahapan penelitian dimulai dari pengumpulan dataset, pra-pemrosesan data untuk mengatasi ketidakseimbangan kelas (*imbalanced class*), perancangan model (CNN, *ResNet50*, *EfficientNet-B0*), pelatihan, hingga evaluasi kinerja. Pendekatan sistematis ini mengacu pada standar *pipeline* evaluasi model FER *modern* [4]. Alur kerja sistem secara keseluruhan direpresentasikan pada Gambar 4.



Gambar 4. Diagram Alur Penelitian

3.1. Pengumpulan Dataset

Dataset yang digunakan dalam penelitian ini adalah CK+ yang difokuskan pada pengenalan tujuh ekspresi wajah makro universal, yaitu anger, disgust, fear, sadness, surprise, happiness, dan neutral. Gambar 5 menampilkan contoh visual dari masing-masing ekspresi wajah tersebut sebagai representasi variasi emosi yang digunakan dalam proses pelatihan model.



Gambar 5. Contoh 7 Ekspresi Wajah

Salah satu tantangan utama dari dataset ini adalah ketidakseimbangan distribusi kelas (*class imbalance*), di mana beberapa emosi seperti happiness dan neutral memiliki jumlah data lebih banyak dibandingkan kelas lain seperti disgust dan fear. Kondisi ini dapat menyebabkan bias model terhadap kelas mayoritas [10].

Untuk mengatasi hal tersebut, dilakukan augmentasi data secara acak guna meningkatkan jumlah sampel pada kelas minoritas hingga distribusi menjadi seimbang dengan total 19.600 citra. Selanjutnya, dataset dibagi menjadi training set (70%), validation set (15%), dan test set (15%) untuk memastikan evaluasi yang objektif. Tabel 1 menunjukkan distribusi dataset yang telah diseimbangkan, di mana setiap kelas memiliki jumlah data yang sama, yaitu 2000 data latih, 400 data

validasi, dan 400 data uji. Distribusi yang merata ini bertujuan untuk menghindari bias model serta memastikan proses pelatihan dan evaluasi berjalan secara adil pada setiap kelas ekspresi.

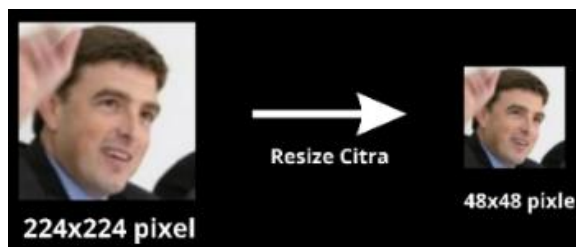
Tabel 1. Distribusi dataset dengan jumlah sama

Kelas Ekspresi	Data Latih	Data Validasi	Data Uji
Marah	2000	400	400
Jijik	2000	400	400
Takut	2000	400	400
Bahagia	2000	400	400
Terkejut	2000	400	400
Sedih	2000	400	400
Netral	2000	400	400

3.2. Pre-processing

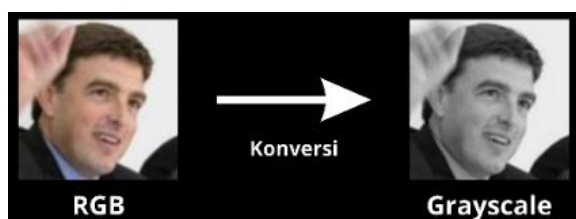
Sebelum citra digunakan dalam proses pelatihan, dilakukan tahap pra-pemrosesan untuk menstandarisasi data dan meningkatkan kualitas input. Tahapan ini penting karena data mentah memiliki variasi pencahayaan, resolusi, dan posisi wajah yang berbeda. Proses ini meliputi resize image, konversi warna, normalisasi, dan augmentasi data.

Pada tahap resize image, citra disesuaikan menggunakan interpolasi bilinear untuk menjaga kualitas visual saat perubahan ukuran. Gambar 6 menunjukkan proses resize citra. Resolusi 48×48 piksel digunakan pada CNN untuk efisiensi komputasi, sedangkan ResNet50 dan EfficientNet-B0 menggunakan resolusi 224×224 piksel agar sesuai dengan model pre-trained berbasis ImageNet [11].



Gambar 6. *Resize Image*

Pada tahap konversi warna, CNN menggunakan grayscale untuk mengurangi kompleksitas komputasi karena ekspresi wajah lebih bergantung pada tekstur dan bentuk. Gambar 7 menampilkan perbandingan citra RGB dan grayscale. Sementara itu, ResNet50 dan EfficientNet-B0 tetap menggunakan format RGB untuk menjaga kompatibilitas dengan arsitektur awal model [9].



Gambar 7. Contoh Citra RGB dan *Graycale*

Selanjutnya, dilakukan proses normalisasi menggunakan metode Min-Max untuk mentransformasikan nilai intensitas piksel dari rentang [0,255] ke dalam rentang [0,1]. Proses ini bertujuan untuk meningkatkan stabilitas numerik selama pelatihan serta mempercepat konvergensi model dalam proses optimasi [12]. Adapun normalisasi tersebut dinyatakan dengan persamaan berikut.

$$I_{norm} = \frac{I(x,y)}{255} \quad (5)$$

Augmentasi data digunakan untuk meningkatkan variasi dataset melalui transformasi citra secara sintetis tanpa penambahan data baru. Proses ini dilakukan secara acak dan dinamis selama pelatihan (runtime), sehingga setiap epoch menghasilkan variasi data yang berbeda. Transformasi yang diterapkan meliputi rotasi, pergeseran posisi, dan pembalikan horizontal. Gambar 8 menampilkan contoh hasil augmentasi citra.



Gambar 8. Contoh Augmentasi Rotasi Acak

Penerapan augmentasi ini bertujuan untuk mensimulasikan berbagai kondisi pengambilan gambar, seperti perubahan sudut, posisi, dan orientasi wajah, sehingga mampu meningkatkan kemampuan generalisasi model dalam mengenali pola ekspresi yang lebih beragam [10].

3.3. Arsitektur Model

Setelah tahap pra-pemrosesan, citra digunakan sebagai input untuk proses pelatihan model. Penelitian ini membandingkan tiga arsitektur deep learning, yaitu CNN, ResNet50, dan EfficientNet-B0 untuk menganalisis kinerja dalam klasifikasi ekspresi wajah.

3.4. Convolutional Neural Network (CNN)

Convolutional Neural Network (CNN) digunakan sebagai model dasar yang berfungsi mengekstraksi fitur citra melalui kombinasi lapisan konvolusi dan *pooling*. Fitur yang dihasilkan kemudian diproses oleh *fully connected layer* untuk menghasilkan prediksi kelas. Arsitektur ini dipilih karena sederhana dan efisien sebagai pembanding awal.

3.5. ResNet50

ResNet50 digunakan dengan pendekatan *transfer learning* sebagai *feature extractor*. Model ini

memanfaatkan bobot *pre-trained* dari ImageNet, sehingga mampu menangkap fitur kompleks dengan lebih baik. Lapisan akhir dimodifikasi menggunakan *fully connected layer* sesuai jumlah kelas pada penelitian ini.

3.6. EfficientNet-B0

EfficientNet-B0 juga digunakan dengan pendekatan *transfer learning*. Model ini memiliki efisiensi parameter yang lebih baik melalui metode *compound scaling*, sehingga mampu memberikan performa tinggi dengan kompleksitas yang relatif rendah.

3.7. Konfigurasi Pelatihan

Proses pelatihan dilakukan dengan membagi dataset menjadi data latih, validasi, dan uji. Model dilatih menggunakan optimizer Adam karena kemampuannya dalam menyesuaikan laju pembelajaran secara adaptif. Fungsi loss yang digunakan adalah *categorical crossentropy*, yang dirumuskan sebagai:

$$L = \sum_{i=1}^n y_i \log(\hat{y}_i) \quad (6)$$

di mana y_i merupakan label sebenarnya dan $\hat{y}_i$ merupakan probabilitas hasil prediksi model. Parameter pelatihan yang digunakan meliputi jumlah epoch, batch size, dan learning rate yang disesuaikan untuk memperoleh performa optimal. Selama proses pelatihan, dilakukan pemantauan terhadap nilai *training loss* dan *validation loss* untuk memastikan stabilitas model.

3.8. Proses Pelatihan Model

Pelatihan model dilakukan menggunakan data latih yang telah melalui proses augmentasi secara real-time. Pada setiap epoch, model memperbarui bobot melalui mekanisme backpropagation berdasarkan error yang dihasilkan, sementara data validasi digunakan untuk mengevaluasi performa model selama pelatihan. Model terbaik kemudian dipilih berdasarkan nilai validation accuracy tertinggi guna menghindari terjadinya overfitting.

3.9. Evaluasi Model

Evaluasi performa kinerja model diukur menggunakan beberapa metrik, yaitu accuracy, precision, recall, dan F1-score. Accuracy mengukur proporsi prediksi benar terhadap seluruh data uji, namun cenderung bias pada dataset tidak seimbang [18]. Oleh karena itu, digunakan precision untuk mengukur ketepatan prediksi positif, recall untuk menilai kemampuan model dalam mendeteksi seluruh data positif, serta F1-score sebagai rata-rata harmonik antara precision dan recall guna memberikan evaluasi yang lebih seimbang [19]. Secara matematis, metrik evaluasi dirumuskan sebagai berikut:

Akurasi (*Accuracy*) digunakan untuk mengukur proporsi keseluruhan prediksi yang benar terhadap total data uji. Metrik ini dirumuskan sebagai berikut.

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \quad (7)$$

Presisi (*Precision*) mengukur tingkat ketepatan prediksi positif, yaitu proporsi data yang diprediksi sebagai kelas positif dan benar-benar termasuk dalam kelas tersebut. Secara matematis, *Precision* dirumuskan sebagai berikut.

$$Precision = \frac{TP}{TP + FP} \quad (8)$$

Sensitivitas (*Recall*) mengukur kemampuan model dalam mendeteksi seluruh data positif yang sebenarnya. Recall dirumuskan sebagai

$$Recall = \frac{TP}{TP + FN} \quad (9)$$

F1-Score merupakan rata-rata harmonik antara *Precision* dan *Recall* yang digunakan untuk menyeimbangkan kedua metrik tersebut. *F1-Score* dirumuskan sebagai:

$$F1\ Score = 2 \times \frac{Precision \times Recall}{Precision + Recall} \quad (10)$$

Pada kasus klasifikasi multi-kelas seperti pengenalan ekspresi wajah, metrik-metrik tersebut dihitung untuk setiap kelas dan dirata-ratakan menggunakan pendekatan *macro-average* atau *weighted-average* untuk memperoleh gambaran performa model secara menyeluruh.

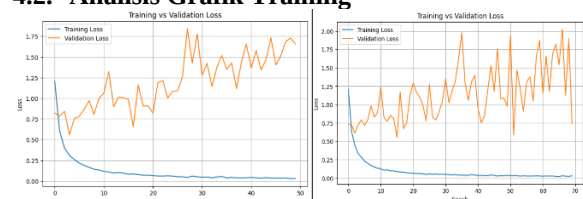
4. HASIL DAN PEMBAHASAN

4.1. Training dan Pengujian Model

Penelitian ini menguji tiga model deep learning, yaitu CNN, ResNet50, dan EfficientNet-B0 dalam klasifikasi ekspresi wajah. Setiap model dilatih menggunakan dua skenario jumlah epoch, yaitu 50 dan 70 epoch, untuk mengevaluasi pengaruh jumlah epoch terhadap kinerja model.

Evaluasi performa dilakukan menggunakan data uji yang tidak terlibat dalam proses pelatihan, sehingga hasil yang diperoleh mencerminkan kemampuan generalisasi model.

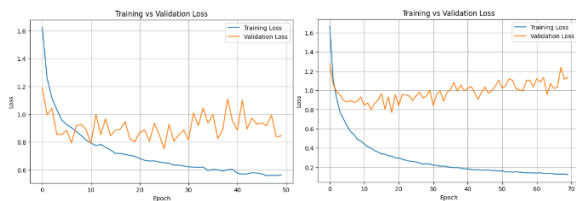
4.2. Analisis Grafik Training



Gambar 9. Grafik Training CNN epoch 50 dan 70

Berdasarkan gambar 9, training loss pada model CNN menurun drastis hingga mendekati nol pada kedua skenario (50 dan 70 epoch), menunjukkan model mampu mempelajari data latih dengan baik. Namun, validation loss terlihat berfluktuasi dan cenderung meningkat, terutama pada skenario 70 epoch.

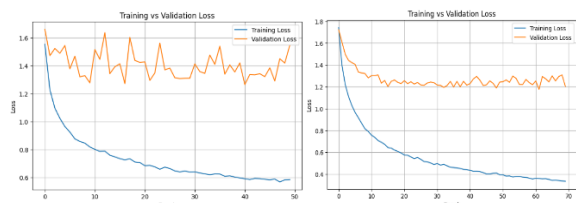
Perbedaan pola ini menunjukkan adanya overfitting, di mana model terlalu menyesuaikan data latih tetapi kurang mampu melakukan generalisasi. Oleh karena itu, pelatihan pada 50 epoch lebih optimal dibandingkan 70 epoch karena lebih stabil dan memiliki risiko overfitting yang lebih rendah.



Gambar 10. Grafik Training Resnet50 epoch 50 dan 70

Sedangkan Gambar 10, terlihat bahwa pada model ResNet50 nilai *training loss* mengalami penurunan yang signifikan pada kedua skenario (50 dan 70 epoch), menunjukkan bahwa model mampu mempelajari data latih dengan sangat baik. Namun, *validation loss* cenderung berfluktuasi dan tidak menunjukkan penurunan yang stabil. Pada skenario 50 epoch, nilai *validation loss* berada pada rentang yang relatif stabil meskipun terjadi naik-turun, sedangkan pada 70 epoch terlihat kecenderungan meningkat.

Perbedaan pola antara *training loss* dan *validation loss* ini menunjukkan adanya overfitting, terutama pada skenario 70 epoch di mana gap antar keduanya semakin besar. Hal ini mengindikasikan bahwa model terlalu menyesuaikan data latih namun kurang mampu melakukan generalisasi. Oleh karena itu, pelatihan dengan 50 epoch dinilai lebih optimal dibandingkan 70 epoch karena lebih stabil dan memiliki risiko overfitting yang lebih rendah.



Gambar 11. Grafik Training EfficientNetB0 epoch 50 dan 70

Kemudian pada gambar 11, terlihat bahwa pada model EfficientNet-B0 nilai *training loss* mengalami penurunan secara signifikan pada kedua skenario (50 dan 70 epoch), yang menunjukkan bahwa model mampu mempelajari data latih dengan baik. Namun, *validation loss* berada pada nilai yang relatif tinggi dan cenderung berfluktuasi tanpa adanya penurunan yang

berarti. Pada skenario 50 epoch, *validation loss* berada di kisaran yang cukup tinggi, sedangkan pada 70 epoch terlihat lebih stabil namun tetap tidak menunjukkan peningkatan performa yang signifikan.

Perbedaan yang cukup besar antara *training loss* dan *validation loss* menunjukkan adanya overfitting yang cukup kuat pada model EfficientNet-B0. Model cenderung menghafal data latih tetapi kurang mampu melakukan generalisasi terhadap data validasi. Oleh karena itu, baik pada 50 maupun 70 epoch, model ini tidak menunjukkan performa optimal. Hal ini mengindikasikan bahwa EfficientNet-B0 kurang sesuai untuk dataset dengan resolusi rendah dan jumlah data terbatas dibandingkan model yang lebih sederhana seperti CNN.

4.3. Evaluasi Matrik

Setelah proses pelatihan model selesai dilakukan, langkah selanjutnya adalah melakukan evaluasi terhadap kinerja masing-masing model menggunakan data uji. Evaluasi ini bertujuan untuk mengukur kemampuan generalisasi model dalam mengklasifikasikan ekspresi wajah yang belum pernah dilihat sebelumnya. Hasil evaluasi disajikan dalam bentuk metrik kuantitatif yang mencakup accuracy, precision, recall, dan F1-score untuk setiap kelas pada masing-masing model dan skenario pelatihan.

Pada tabel 5 yang menunjukkan hasil kinerja CNN Secara lebih rinci, CNN menunjukkan performa yang sangat baik pada kelas Happiness, Surprise, dan Fear dengan nilai precision, recall, dan F1-score yang mendekati sempurna. Hal ini mengindikasikan bahwa pola ekspresi pada kelas-kelas tersebut memiliki karakteristik visual yang lebih mudah dikenali oleh model. Sebaliknya, kelas Sadness menunjukkan performa terendah, khususnya pada nilai recall yang sangat kecil, yang menandakan bahwa model masih kesulitan dalam mendeteksi ekspresi tersebut secara konsisten.

Tabel 2. Hasil evaluasi kinerja CNN

Skenario	Kelas	Precision	Recall	F1-Score
Epoch 50	Anger	0,65	0,62	0,64
	Disgust	0,86	0,99	0,92
	Fear	0,99	1,00	0,99
	Happiness	1,00	0,99	1,00
	Neutral	0,60	0,94	0,73
	Sadness	0,76	0,25	0,38
	Surprise	1,00	0,99	0,99
Accuracy		0,83		
Epoch 70	Anger	0,65	0,61	0,63
	Disgust	0,87	1,00	0,93
	Fear	0,99	0,99	0,99
	Happiness	0,98	0,99	0,99
	Neutral	0,60	0,96	0,73
	Sadness	0,90	0,25	0,39
	Surprise	1,00	1,00	1,00
Accuracy		0,83		

Dan pada table 3 Pada model ResNet50, performa yang dihasilkan berada di bawah CNN dengan accuracy meningkat dari 0,63 pada 50 epoch

menjadi 0,67 pada 70 epoch. Meskipun terjadi peningkatan, nilai metrik pada sebagian besar kelas masih menunjukkan ketidakseimbangan antara precision dan recall. Kelas Surprise tetap menjadi kelas dengan performa terbaik, sedangkan Sadness kembali menjadi kelas dengan performa terendah. Hal ini menunjukkan bahwa meskipun ResNet50 mampu menangkap fitur yang lebih kompleks, model ini belum mampu beradaptasi secara optimal terhadap karakteristik dataset.

Tabel 3. Hasil Evaluasi Kinerja RESNET50

Skenario	Kelas	Precision	Recall	F1-Score
Epoch 50	Anger	0,68	0,50	0,58
	Disgust	0,67	0,69	0,68
	Fear	0,66	0,54	0,59
	Happiness	0,62	0,77	0,69
	Neutral	0,50	0,68	0,58
	Sadness	0,36	0,27	0,31
	Surprise	0,88	1,00	0,93
Accuracy		0,63		
Epoch 70	Anger	0,87	0,47	0,61
	Disgust	0,69	0,77	0,73
	Fear	0,88	0,57	0,69
	Happiness	0,63	0,94	0,76
	Neutral	0,42	0,78	0,54
	Sadness	0,64	0,15	0,24
	Surprise	0,90	1,00	0,95
Accuracy		0,67		

Sementara itu, di tabel 4 EfficientNet-B0 menghasilkan performa paling rendah di antara ketiga model, dengan accuracy hanya mencapai 0,47 pada 50 epoch dan meningkat menjadi 0,50 pada 70 epoch. Nilai precision, recall, dan F1-score pada sebagian besar kelas relatif rendah dan tidak stabil. Kelas Sadness menjadi yang paling sulit dikenali dengan nilai metrik yang sangat kecil, sedangkan Surprise masih menunjukkan performa relatif lebih baik dibandingkan kelas lainnya. Hal ini mengindikasikan bahwa EfficientNet-B0 mengalami kesulitan dalam melakukan generalisasi pada dataset dengan resolusi rendah dan jumlah data terbatas.

Tabel 4. Hasil Evaluasi Kinerja EFFICIENTNET-B0

Skenario	Kelas	Precision	Recall	F1-Score
Epoch 50	Anger	0,44	0,32	0,37
	Disgust	0,39	0,80	0,53
	Fear	0,99	0,21	0,34
	Happiness	0,44	0,59	0,53
	Neutral	0,42	0,48	0,45
	Sadness	0,09	0,07	0,08
	Surprise	0,91	0,81	0,85
Accuracy		0,47		
Epoch 70	Anger	0,39	0,56	0,46
	Disgust	0,46	0,72	0,56
	Fear	0,84	0,38	0,52
	Happiness	0,56	0,59	0,57
	Neutral	0,50	0,42	0,46
	Sadness	0,13	0,11	0,12
	Surprise	0,96	0,73	0,83
Accuracy		0,50		

Secara keseluruhan, hasil evaluasi metrik menunjukkan bahwa kompleksitas arsitektur tidak selalu berbanding lurus dengan peningkatan performa. Model CNN yang lebih sederhana justru mampu memberikan hasil yang lebih stabil dan akurat, terutama dalam menangani karakteristik data ekspresi wajah beresolusi rendah.

4.4. Analisis Perbandingan 3 Model CNN, ResNet50, dan EfficientNetB0

Berdasarkan serangkaian eksperimen yang dilakukan pada dataset CK+ dengan skenario 50 dan 70 epoch, hasil evaluasi menunjukkan perbedaan performa yang signifikan di antara ketiga arsitektur. Secara keseluruhan, arsitektur CNN sederhana memberikan hasil terbaik dibandingkan model yang lebih kompleks seperti ResNet50 dan EfficientNet-B0. Ringkasan perbandingan metrik kinerja untuk ketiga model pada skenario terbaiknya (70 epoch) disajikan pada Tabel 6.

Tabel 5. Perbandingan Kinerja 3 Model

Arsitekt ur	epoc h	Akura si	Presi si	Reca ll	F1- Scor e
CNN	50	0,82	0,84	0,82	0,80
	70	0,83	0,85	0,83	0,81
ResNet5 0	50	0,82	0,84	0,82	0,80
	70	0,83	0,85	0,83	0,81
Efficient -NetB0	50	0,47	0,53	0,47	0,45
	70	0,50	0,55	0,50	0,50

Berdasarkan Tabel 6, terlihat bahwa model CNN memberikan performa terbaik dibandingkan ResNet50 dan EfficientNet-B0 pada kedua skenario pelatihan. Meskipun ResNet50 dan EfficientNet-B0 memiliki arsitektur yang lebih dalam dan kompleks, hasil eksperimen menunjukkan bahwa peningkatan kompleksitas tidak secara langsung meningkatkan akurasi maupun F1-score. CNN justru menunjukkan konsistensi performa yang lebih tinggi serta peningkatan yang stabil ketika epoch ditambah dari 50 menjadi 70.

Fenomena ini menunjukkan bahwa kesesuaian arsitektur terhadap karakteristik dataset memiliki peran yang lebih dominan dibandingkan kedalaman jaringan semata. Pada citra ekspresi wajah resolusi rendah (48×48 grayscale), fitur spasial sederhana sudah cukup representatif sehingga arsitektur ringan seperti CNN mampu melakukan separabilitas kelas secara efektif. Sebaliknya, model berbasis transfer learning menunjukkan generalization gap yang lebih besar akibat perbedaan domain antara fitur pre-trained dan distribusi data ekspresi wajah yang digunakan dalam penelitian ini.

5. KESIMPULAN DAN SARAN

Berdasarkan hasil penelitian, dapat disimpulkan bahwa arsitektur CNN memberikan performa terbaik dalam klasifikasi ekspresi wajah dibandingkan

ResNet50 dan EfficientNet-B0 pada dataset CK+ yang telah diseimbangkan. CNN mampu mencapai akurasi tertinggi sebesar 0,83 dengan performa yang stabil pada kedua skenario pelatihan, yaitu 50 dan 70 epoch. Sementara itu, ResNet50 menunjukkan peningkatan performa yang terbatas, dan EfficientNet-B0 menghasilkan performa paling rendah dengan kecenderungan overfitting yang lebih tinggi.

Hasil penelitian ini menunjukkan bahwa tingkat kompleksitas arsitektur tidak selalu menjamin peningkatan kinerja model, terutama pada dataset dengan resolusi rendah dan jumlah data terbatas. Model sederhana seperti CNN justru lebih mampu melakukan generalisasi dengan baik dibandingkan model berbasis transfer learning yang mengalami kesenjangan distribusi fitur.

Untuk penelitian selanjutnya, disarankan melakukan fine-tuning yang lebih optimal pada model pre-trained, menambahkan teknik regularisasi, serta menggunakan dataset dengan resolusi lebih tinggi agar potensi arsitektur yang lebih kompleks dapat dimaksimalkan.

DAFTAR PUSTAKA

- [1] P. Ekman, "Universals and cultural differences in facial expressions of emotion.," *Nebraska Symp. Motiv.*, vol. 19, pp. 207–283, 1971.
- [2] S. Li and W. Deng, "Deep Facial Expression Recognition : A Survey," vol. 3045, no. c, pp. 1–20, 2020, doi: 10.1109/TAFFC.2020.2981446.
- [3] J. Wan, B. Li, K. Wang, X. Teng, T. Wang, and B. Mao, "ScienceDirect ScienceDirect An Improved ResNet50 for Environment Image Classification An Improved ResNet50 for Environment Image Classification," *Procedia Comput. Sci.*, vol. 242, pp. 1000–1007, 2024, doi: 10.1016/j.procs.2024.08.246.
- [4] M. Tan and Q. V Le, "EfficientNet : Rethinking Model Scaling for Convolutional Neural Networks," 2019.
- [5] D. Goleman, *Emotional Intelligence: Why It Can Matter More Than IQ*. London: Bloomsbury Publishing Plc, 1996.
- [6] S. Almoshity and S. A. Al-asbaily, "Face Expressions Recognition by using Deep Learning," *Int. J. Comput. Appl.*, vol. 186, no. 8, pp. 40–44, 2024.
- [7] N. Rodrigues, N. Costa, and A. Pereira, "Systematic Review of Emotion Detection with Computer Vision and DeepLearning," *MDPI*, pp. 1–29, 2024.
- [8] D. Hartanto and R. Herawati, "COMPARATIVE ANALYSIS OF EFFICIENTNET AND RESNET MODELS IN THE CLASSIFICATION OF SKIN CANCER," *J. Inform. PROXIES*, vol. 7, no. 2, pp. 69–84, 2024.
- [9] N. Mohamed and S. Kailani, "Comparison between ResNet and EfficientNet for Image Classification – An Analytical Study of Performance and Efficiency," *North African J. Sci. Publ.*, vol. 3, no. 4, pp. 18–25, 2025.
- [10] F. I. Ilmawati and T. Hidayat, "Optimizing Facial Expression Recognition with Image Augmentation Techniques : VGG19 Approach on FERC Dataset," vol. 8, no. 2, pp. 632–640, 2024.
- [11] A. A. Istiqomah, C. A. Sari, A. Susanto, and E. H. Rachmawanto, "Facial Expression Recognition using Convolutional Neural Networks with Transfer Learning Resnet-50," vol. 8, no. 2, pp. 257–264, 2024.
- [12] jeff heaton, "Ian Goodfellow, Yoshua Bengio, and Aaron Courville: Deep learning," *Genet. Program. Evolvable Mach.*, vol. 19, no. 1, pp. 305–307, 2018, doi: 10.1007/s10710-017-9314-z.
- [13] A. L. S. Guntero, E. Julianto, and D. Budiyo, "Pengenalan Ekspresi Wajah Menggunakan Convolutional Neural Network," *J. Inform. Atma Jogja*, vol. 3, pp. 155–160, 2022.
- [14] I. Maulana, N. Khairunisa, and R. Mufidah, "DETEKSI BENTUK WAJAH MENGGUNAKAN CONVOLUTIONAL NEURAL NETWORK (CNN)," vol. 7, no. 6, pp. 3348–3355, 2023.
- [15] N. Behar and M. Shrivastava, "ResNet50-Based Effective Model for Breast Cancer Classification Using Histopathology Images," *Tech Sci. Press*, vol. 130, no. 2, 2021, doi: 10.32604/cmes.2022.017030.
- [16] B. Hossain, S. M. H. Sazzad, M. Islam, and N. Akhtar, "Informatics in Medicine Unlocked Transfer learning with fine-tuned deep CNN ResNet50 model for classifying COVID-19 from chest X-ray images," *Informatics Med. Unlocked*, vol. 30, no. March, p. 100916, 2022, doi: 10.1016/j.imu.2022.100916.
- [17] A. Deshpande, V. V Estrela, and P. Patavardhan, "The DCT-CNN-ResNet50 architecture to classify brain tumors with super-resolution , convolutional neural network , and the ResNet50," *Neurosci. Informatics*, vol. 1, no. 4, p. 8, 2021, doi: 10.1016/j.neuri.2021.100013.
- [18] D. Berrar, "Performance Measures for Binary Classification 1," *Elsevier*, vol. 1, pp. 1–31, 2019, doi: <https://doi.org/10.1016/B978-0-12-809633-8.20351-8>.
- [19] D. Chicco and G. Jurman, "The advantages of the Matthews correlation coefficient (MCC) over F1 score and accuracy in binary classification evaluation," pp. 1–13, 2020.