FORMAT: Jurnal Ilmiah Teknik Informatika
Vol. 15, No. 1, Januari 2026, pp. 22-30
Akreditasi Sinta 5 RISTEKBRIN No. 164/E/KPT/2021
DOI: 10.22441/format.2026.v15.i1.003
E-ISSN 2722-7162
P-ISSN 2089-5615

Perbandingan Performa Xception dan InceptionV1 untuk Pengenalan Ekspresi Wajah

Ferdinand Defin Delio[1*], Diah Aryani[2], Habibullah Akbar[3], Muhamad Yusuf[4], Yaya Sudarya Triana[5]

[1,2,3] Fakultas Ilmu Komputer, Universitas Esa Unggul, Jalan Arjuna Utara No.9, Kebon Jeruk, Jakarta 11510
[4,5] Fakultas Ilmu Komputer, Universitas Mercu Buana, Jl. Raya, RT.4/RW.1, Meruya Sel., Kec. Kembangan, Jakarta, Daerah Khusus Ibukota Jakarta

[1]ferdinanddefin16@student.esaunggul.ac.id, [2]diah.aryani@esaunggul.ac.id, [3]habibullah.akbar@esaunggul.ac.id, [4]mhd.yusuf@mercubuana.ac.id, [5]yaya.sudarya@mercubuana.ac.id

ABSTRAK

Penelitian ini bertujuan untuk menganalisis dan membandingkan performa dua arsitektur Convolutional Neural Network (CNN) populer, yaitu Xception dan InceptionV1, dalam tugas pengenalan ekspresi wajah (Facial Expression Recognition/FER). Penelitian ini dilakukan dengan pendekatan transfer learning dan fine-tuning menggunakan dataset FER-2013 yang berisi 35.887 citra wajah grayscale berukuran 48×48 piksel yang diklasifikasikan ke dalam tujuh emosi dasar. Setiap citra diubah ukurannya menjadi 224×224 piksel, dinormalisasi, dan diproses dengan teknik augmentasi untuk meningkatkan generalisasi model terhadap variasi ekspresi wajah, pencahayaan, dan pose. Proses pelatihan dilakukan selama 30 epoch menggunakan optimizer Adam dengan learning rate 0.0001 dan batch size 64. Strategi fine-tuning dilakukan dengan membuka 30% lapisan atas model untuk mengoptimalkan bobot fitur yang telah dipelajari sebelumnya dari dataset ImageNet. Evaluasi kinerja dilakukan berdasarkan metrik akurasi, presisi, recall, F1-score, serta efisiensi komputasi yang diukur dari waktu pelatihan dan inferensi. Hasil eksperimen menunjukkan bahwa Xception mencapai akurasi validasi 70,69% dengan waktu inferensi rata-rata 20–25 ms, sedangkan InceptionV1 mencapai 65,80% dengan waktu inferensi 43–126 ms. Arsitektur Xception terbukti lebih efisien secara komputasi karena memanfaatkan depthwise separable convolution yang mengurangi jumlah parameter tanpa menurunkan akurasi. Temuan ini menunjukkan bahwa Xception lebih sesuai untuk aplikasi FER real-time dan perangkat dengan sumber daya terbatas, serta memberikan dasar yang kuat bagi penelitian lanjutan dalam pengembangan sistem pengenalan ekspresi wajah berbasis video dan lingkungan dunia nyata.

Article Info

Kata Kunci:
CNN
Pengenalan Ekspresi Wajah
FER-2013
Xception
InceptionV1

Riwayat artikel:
Submit 23 Sept 2025
Revisi 3 Nov 2025
Diterima 11 Nov 2025

1. PENDAHULUAN

Pengenalan Ekspresi Wajah (*Facial Expression Recognition/FER*) merupakan teknologi yang memungkinkan sistem komputer mengenali emosi manusia melalui analisis citra wajah. Teknologi ini berperan penting dalam berbagai bidang seperti kesehatan, keamanan, psikologi, serta interaksi manusia–komputer [1].. Kemajuan pesat dalam *deep learning*, khususnya *Convolutional Neural Network* (CNN), telah meningkatkan akurasi sistem FER secara signifikan. [2].

Dua arsitektur CNN yang populer dan banyak digunakan adalah Xception dan InceptionV1 [3]. Xception merupakan pengembangan dari Inception yang memanfaatkan *depthwise separable convolution* untuk mengefisienkan komputasi [4], sementara InceptionV1 menggunakan *Inception module* untuk mengekstraksi fitur multi-skala secara paralel [5].

E-ISSN 2722-7162
P-ISSN 2089-5615

Namun, tantangan masih muncul dalam penerapan FER pada dunia nyata karena variasi pose, pencahayaan, ekspresi halus, dan distribusi kelas yang tidak seimbang [6]. Oleh karena itu, diperlukan analisis komprehensif yang tidak hanya menilai akurasi tetapi juga efisiensi komputasi dari arsitektur CNN.

Tujuan penelitian ini adalah mengevaluasi dan membandingkan performa Xception dan InceptionV1 dalam hal akurasi dan efisiensi komputasi pada dataset FER-2013. Kontribusi utama penelitian ini adalah melakukan perbandingan langsung kedua arsitektur di bawah kondisi eksperimental yang seragam, menyoroti trade-off antara akurasi dan efisiensi komputasi yang penting untuk aplikasi dunia nyata seperti sistem real-time atau perangkat mobile [7].

2. METODE PENELITIAN

Pertumbuhan pesat teknologi *deep learning*, khususnya melalui arsitektur *Convolutional Neural Network* (CNN), telah memberikan kemajuan signifikan dalam pengenalan ekspresi wajah (*Facial Expression Recognition/FER*). Namun, sebagian besar model masih dilatih pada dataset laboratorium yang terstandarisasi, sehingga performanya menurun ketika diaplikasikan pada lingkungan nyata dengan variasi pose, pencahayaan, dan ekspresi yang beragam [2].

Dataset FER-2013 merupakan *benchmark* penting dalam penelitian FER karena berisi 35.887 citra wajah grayscale berukuran 48×48 piksel yang dikelompokkan ke dalam tujuh ekspresi dasar. Walau banyak digunakan, FER-2013 memiliki keterbatasan berupa resolusi rendah, distribusi kelas tidak seimbang, dan variasi pencahayaan [10]. Keterbatasan ini justru menjadikannya representatif untuk kondisi dunia nyata.

Model Inception V1 dikenal efektif dalam mengekstraksi fitur multi-skala, tetapi memerlukan komputasi besar. Sebaliknya, Xception menggunakan *depthwise separable convolution* untuk mengefisienkan komputasi tanpa mengurangi akurasi [8]. Oleh karena itu, dibutuhkan analisis mendalam mengenai kompromi antara akurasi dan efisiensi komputasi dari kedua arsitektur tersebut untuk mendukung aplikasi FER pada sistem real-time atau perangkat dengan sumber daya terbatas.

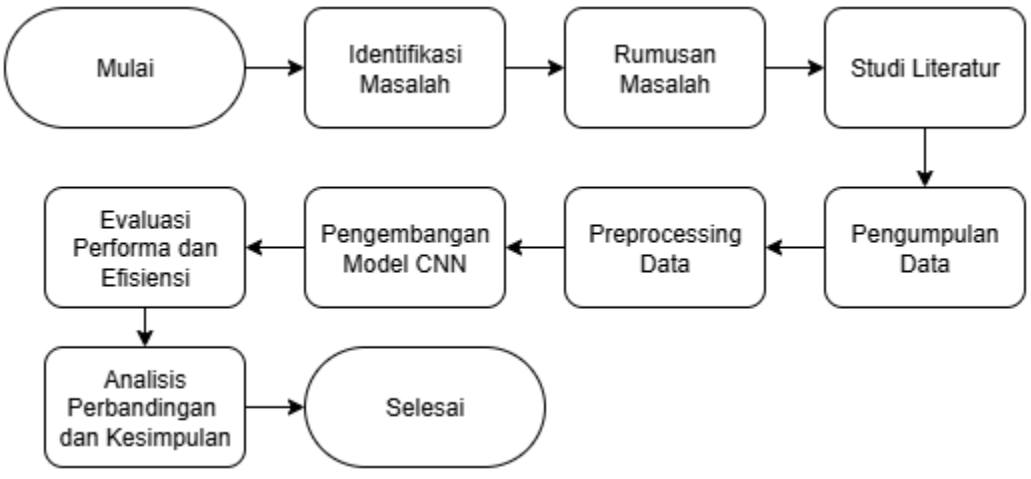
2.1. Riset Gap

Berdasarkan kajian literatur, masih terdapat kesenjangan penelitian (*research gap*) yang belum banyak dieksplorasi. Sebagian besar studi hanya menilai performa Xception atau Inception V1 secara terpisah tanpa perbandingan langsung menggunakan pengaturan eksperimen yang seragam [4], [6].

Selain itu, aspek efisiensi komputasi dan waktu inferensi jarang menjadi fokus utama, padahal faktor tersebut sangat penting untuk sistem FER yang digunakan pada perangkat bergerak dan aplikasi waktu nyata. Kurangnya penelitian yang menganalisis trade-off antara akurasi dan efisiensi menjadi celah yang perlu diisi.[9]

Penelitian ini berupaya menutup kesenjangan tersebut dengan membandingkan secara kuantitatif dan kualitatif kedua arsitektur (Xception dan Inception V1) pada dataset FER-2013, menggunakan pendekatan *transfer learning* dan *fine-tuning* dalam kondisi eksperimen yang seragam [3].

2.2. Tahapan Penelitian



Gambar 1. Tahapan Penelitian.

E-ISSN 2722-7162
P-ISSN 2089-5615

Pada Gambar 1 menunjukkan alur penelitian yang terdiri dari beberapa tahapan yang dilakukan secara sistematis mulai dari identifikasi masalah hingga analisis hasil dan kesimpulan.

Tahap pertama adalah identifikasi masalah, yaitu proses memahami isu utama dalam penelitian, seperti rendahnya efisiensi dan performa model CNN dalam mengenali ekspresi wajah pada kondisi dunia nyata. Hasil dari tahap ini digunakan untuk merumuskan rumusan masalah, yang menjadi dasar dalam penentuan arah penelitian, misalnya bagaimana membandingkan performa Xception dan Inception V1 secara komprehensif.

Selanjutnya dilakukan studi literatur untuk mengumpulkan teori dan hasil penelitian terdahulu yang berkaitan dengan Facial Expression Recognition (FER), arsitektur Convolutional Neural Network (CNN), serta pendekatan transfer learning dan fine-tuning. Tahapan ini membantu memperkuat landasan teoritis dan menemukan kesenjangan penelitian yang akan diisi.

Tahap berikutnya adalah pengumpulan data, yaitu memperoleh dataset FER-2013 yang berisi 35.887 citra wajah grayscale berukuran 48×48 piksel. Dataset ini dipilih karena sering digunakan sebagai benchmark pada penelitian FER, meskipun memiliki tantangan berupa resolusi rendah dan distribusi kelas yang tidak seimbang[10]. Setelah itu dilakukan preprocessing data, yang meliputi resizing citra ke ukuran 224×224 piksel, normalisasi nilai piksel, konversi citra ke format RGB, serta augmentasi seperti rotasi, zoom, dan perubahan kecerahan untuk meningkatkan variasi data pelatihan.

Tahapan selanjutnya adalah pengembangan model CNN, di mana dua arsitektur, yaitu Xception dan Inception V1, digunakan dengan pendekatan transfer learning menggunakan bobot awal dari ImageNet. Penelitian dilakukan pada dua skenario, yaitu *baseline* (melatih hanya lapisan atas) dan *fine-tuning* (membuka sebagian lapisan atas untuk pelatihan lanjutan). Pelatihan dilakukan menggunakan framework TensorFlow 2.10 (Keras API) pada lingkungan Google Colab dengan GPU Tesla T4, RAM 12 GB, dan prosesor Intel Xeon.

Setelah model selesai dilatih, dilakukan evaluasi performa dan efisiensi berdasarkan metrik akurasi, precision, recall, F1-score, serta confusion matrix untuk melihat distribusi prediksi antar kelas. Selain itu, efisiensi komputasi juga dievaluasi melalui waktu pelatihan, rata-rata waktu inferensi, dan jumlah parameter model.

Tahap terakhir adalah analisis perbandingan dan kesimpulan, di mana hasil evaluasi kedua arsitektur dibandingkan secara kuantitatif dan kualitatif untuk menilai trade-off antara akurasi dan efisiensi komputasi. Analisis ini digunakan untuk menarik kesimpulan dan memberikan rekomendasi arsitektur terbaik untuk implementasi sistem FER pada perangkat dengan sumber daya terbatas.

2.3. Konfigurasi Ekperimen

Dalam penelitian ini, kedua arsitektur—Xception dan Inception V1—dilatih menggunakan pendekatan *transfer learning* dengan bobot awal dari ImageNet. Proses pelatihan dilakukan pada lingkungan Google Colab GPU (Tesla T4**)** dengan RAM 12 GB dan prosesor Intel Xeon [11]. Untuk menjaga konsistensi eksperimen, dataset dibagi menjadi 80% data pelatihan dan 20% data validasi, tanpa penerapan *k-fold cross-validation*. Strategi *fine-tuning* dilakukan dengan membuka sebagian lapisan atas (sekitar 30%) dan membekukan lapisan dasar guna menyesuaikan fitur tingkat tinggi terhadap dataset FER-2013. Rincian lengkap konfigurasi eksperimen ditunjukkan pada Tabel 1 berikut.

Table 1. Konfigurasi Eksperimen

Model	Optimizer	Learning Rate	Batch Size	Epoch	Fine-Tuning	Split Data	Framework
Inception V1	SGD (momentum=0.9)	0.01	32	30	Membekukan lapisan bawah (layers[:50]) dan melatih lapisan atas	80%–20%	TensorFlow 2.10 (Keras)
Xception	Adam	0.001	64	30	Membuka sekitar 30% lapisan atas untuk fine-tuning	80%–20%	TensorFlow 2.10 (Keras)

2.4. Justifikasi Pemilihan Metrik

Pemilihan metrik evaluasi dilakukan untuk memberikan penilaian performa model yang lebih adil dan representatif terhadap karakteristik dataset FER-2013 yang memiliki distribusi kelas tidak seimbang. Dengan demikian, evaluasi tidak hanya bergantung pada akurasi, tetapi juga mempertimbangkan keseimbangan antar kelas.

E-ISSN 2722-7162
P-ISSN 2089-5615

- Akurasi digunakan untuk menilai proporsi prediksi benar secara keseluruhan.
- Presisi dan recall menilai performa model terhadap kelas minoritas yang sering tidak seimbang dalam dataset FER-2013.
- F1-score menjadi ukuran harmonisasi antara presisi dan recall, sehingga memberikan evaluasi yang lebih adil.
- Confusion Matrix digunakan untuk memvisualisasikan pola kesalahan klasifikasi antar kelas.

Kombinasi metrik ini memberikan pemahaman yang komprehensif terhadap performa dan efisiensi masing-masing arsitektur, serta mendukung analisis trade-off antara akurasi dan waktu inferensi sebagaimana menjadi fokus utama penelitian ini.

2.5. State Of art

Penelitian-penelitian terdahulu telah berkontribusi besar dalam pengembangan model Facial Expression Recognition (FER) berbasis Convolutional Neural Network (CNN). Beberapa studi awal menunjukkan bahwa arsitektur CNN konvensional mampu mengenali ekspresi dasar dengan akurasi tinggi, namun masih menghadapi kendala pada efisiensi dan kemampuan generalisasi terhadap data dunia nyata.[12]

Penelitian oleh [5] membandingkan CNN tradisional dengan model berbasis Inception, dan menemukan bahwa arsitektur Inception memberikan peningkatan akurasi pada dataset FER-2013 meskipun dengan beban komputasi yang lebih tinggi. Sunil & Hariprasad (2023) mengusulkan model hibrida yang mengintegrasikan Xception dan Inception-v2, menunjukkan bahwa kombinasi kedua arsitektur dapat saling melengkapi dan menghasilkan akurasi yang lebih tinggi dengan efisiensi yang relatif terjaga [8].

Sementara itu, Shanimol & Charles (2024) menggabungkan CNN berbasis Inception dengan Gated Recurrent Unit (GRU) dan mencapai akurasi hingga 95,56% pada dataset CK+. Integrasi CNN–RNN ini membantu model menangkap dinamika temporal ekspresi wajah, namun kompleksitas dan waktu inferensi yang tinggi membuat pendekatan tersebut kurang cocok untuk aplikasi real-time berbasis citra diam [13].

Penelitian Li & Deng (2019) dengan Deep Locality-Preserving CNN (DLP-CNN) menekankan pentingnya representasi fitur lokal yang diskriminatif untuk mendeteksi emosi kompleks pada kondisi dunia nyata [2]. Di sisi lain, Sheng & Lau (2024) menunjukkan bahwa arsitektur ringan seperti Inception-V1 dan Inception-V3 dapat memberikan keseimbangan optimal antara akurasi dan efisiensi waktu komputasi, yang menjadi faktor penting untuk aplikasi praktis [7].

Dari hasil kajian tersebut, terlihat bahwa belum terdapat konsensus mengenai arsitektur CNN paling efektif untuk FER, khususnya pada kondisi dunia nyata (in-the-wild). Sebagian penelitian menekankan akurasi tinggi, sementara yang lain menyoroti efisiensi dan kecepatan inferensi. Oleh karena itu, penelitian ini menempati posisi unik dalam literatur dengan melakukan perbandingan langsung antara arsitektur Xception dan Inception V1 di bawah pengaturan eksperimen yang seragam, untuk menilai trade-off antara akurasi dan efisiensi komputasi dalam konteks aplikasi real-time.[14]

3. HASIL DAN DISKUSI

Bagian ini menjelaskan secara rinci hasil penelitian yang dilakukan untuk membandingkan performa dua arsitektur CNN populer, yaitu Xception dan Inception V1, dalam mengklasifikasikan ekspresi wajah manusia pada dataset FER-2013. Penjelasan mencakup tahapan *preprocessing*, pelatihan model pada tahap Baseline dan Fine-Tuned selama 30 *epoch*, evaluasi menggunakan metrik klasifikasi (akurasi, presisi, recall, dan F1-score), serta analisis kualitatif melalui confusion matrix dan grafik performa.

3.1. Data dan Preprocessing

Dataset yang digunakan dalam penelitian ini adalah FER-2013, yang terdiri dari 35.887 citra wajah grayscale beresolusi 48×48 piksel. Seluruh citra dikategorikan ke dalam tujuh kelas emosi dasar, yaitu marah (angry), jijik (disgust), takut (fear), bahagia (happy), sedih (sad), terkejut (surprise), dan netral (neutral).[10]

Sebelum digunakan untuk pelatihan, seluruh gambar dikonversi ke format RGB dan di-resize menjadi 224×224 piksel agar sesuai dengan kebutuhan input model Xception dan Inception V1. Proses normalisasi dilakukan dengan membagi nilai piksel dengan 255 (rescale=1./255) untuk menyamakan skala input antar citra.
Perbedaan utama terletak pada strategi augmentasi citra:

- Baseline: tidak menggunakan augmentasi (plain input), hanya dilakukan resizing dan normalisasi.
- Fine-Tuned: menggunakan ImageDataGenerator dari Keras dengan intensitas augmentasi berbeda untuk tiap model. Xception menggunakan augmentasi agresif, meliputi rotasi hingga 45°, perubahan kecerahan, pergeseran horizontal/vertikal, zoom, shear, dan horizontal flip. Inception V1 menggunakan augmentasi lebih ringan untuk menjaga stabilitas pelatihan.

E-ISSN 2722-7162
P-ISSN 2089-5615

Pendekatan ini bertujuan untuk membandingkan performa awal masing-masing arsitektur dalam kondisi murni (baseline) serta mengevaluasi peningkatan kinerja yang dihasilkan melalui fine-tuning dan augmentasi data.

3.2. Pelatihan Model dan Evaluasi Awal

Proses pelatihan dilakukan secara terpisah untuk Xception dan Inception V1 dengan konfigurasi yang konsisten, yaitu penggunaan optimizer Adam, fungsi loss sparse_categorical_crossentropy, dan batch size sebesar 64. Jumlah epoch ditetapkan sebanyak 30 baik untuk tahap baseline maupun setelah dilakukan fine-tuning.

Pada tahap baseline, Xception mencapai akurasi sebesar 52,76% dengan waktu pelatihan sekitar ±75 menit, sedangkan Inception V1 memperoleh akurasi awal sebesar 63.37% dengan waktu pelatihan ±61.5 menit. Perbedaan ini menunjukkan bahwa Xception memiliki kemampuan representasi awal yang lebih baik dibandingkan Inception V1 dalam kondisi tanpa fine-tuning.

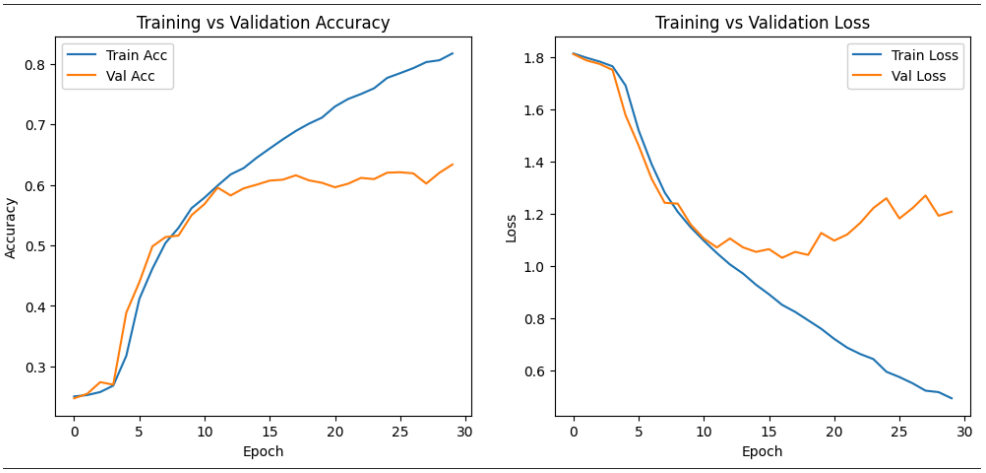
Table 2. Perbandingan Model

Model	Tahap	Akurasi	Waktu
Xception	Baseline	52,76%	±75 Menit
	Fine-Tuned	70,69%	±195 Menit
Inception V.1	Baseline	63.37%	±62.5 Menit
	Fine-Tuned	65.80%	±180 Menit

Setelah dilakukan fine-tuning, performa kedua model meningkat secara signifikan. Xception berhasil mencapai akurasi tertinggi sebesar 70,69% dengan waktu pelatihan ±195 menit, sementara Inception V1 meraih akurasi sebesar 65.80% dengan waktu pelatihan ±180 menit.

Hasil ini juga melampaui penelitian Kim et al. (2021) yang menggunakan *Inception-V1* dengan metode *Facial Image Threshing (FIT)* dan hanya memperoleh akurasi 63,99%, serta penelitian lain yang melaporkan 61,12% untuk Inception V1 [15]. Dengan demikian, penerapan *fine-tuning* dan augmentasi agresif pada Xception terbukti lebih efektif, sejalan dengan temuan Sunil & Hariprasad (2023) bahwa arsitektur lebih dalam memberikan efisiensi dan akurasi lebih baik dalam *Facial Expression Recognition* [8].

3.3. Visualisasi Grafik Accuracy dan Loss

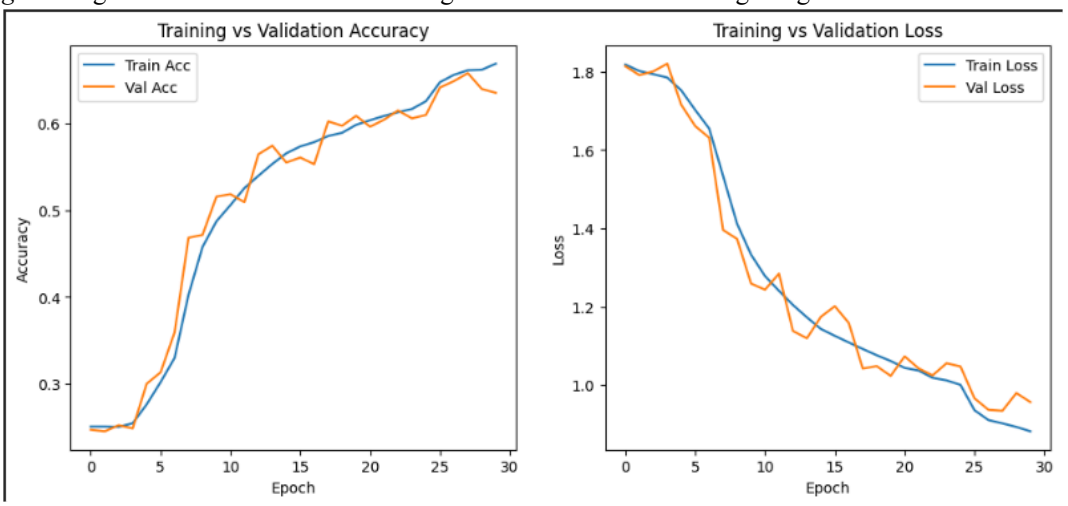


Gambar 2. Visualisasi Grafik Accuracy/loss InceptionV1 Baseline

Pada Gambar 2 menunjukkan perbandingan kurva *training* dan *validation* accuracy serta loss pada model InceptionV1 baseline selama 30 epoch. Terlihat bahwa akurasi pelatihan meningkat stabil hingga mencapai sekitar 0,83, sedangkan akurasi validasi hanya mencapai sekitar 0,6 dan mulai stagnan setelah epoch ke-10. Hal ini menandakan adanya kesenjangan performa antara data pelatihan dan validasi. Pada grafik loss, nilai *training loss* terus menurun hingga di bawah 0,5, sementara *validation loss* menurun pada awal pelatihan namun kembali naik setelah epoch ke-15. Pola ini menunjukkan bahwa model mulai mengalami overfitting, di mana performa pelatihan terus meningkat tetapi kemampuan generalisasi menurun. Secara keseluruhan, model InceptionV1 baseline mampu

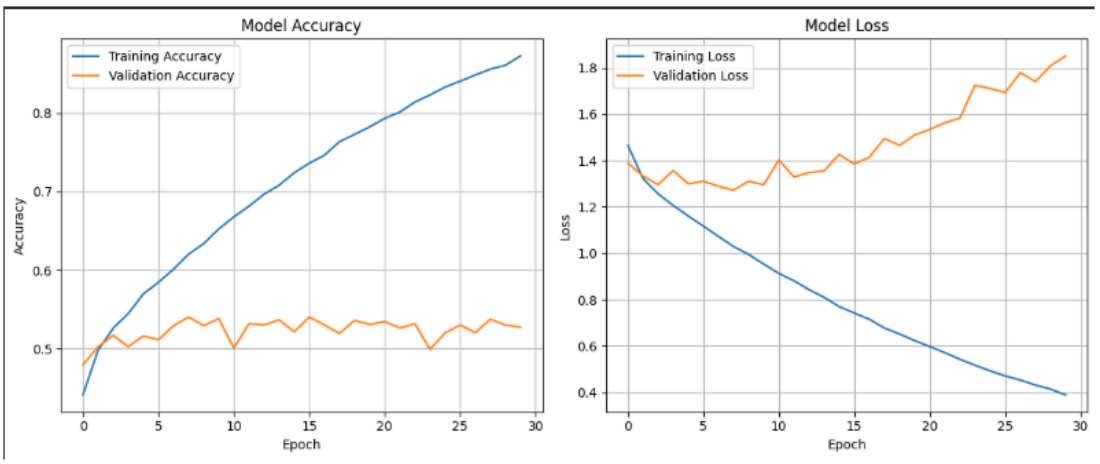
E-ISSN 2722-7162
P-ISSN 2089-5615

belajar dengan baik di data pelatihan, namun belum optimal pada data validasi, sehingga diperlukan proses *fine-tuning* dan augmentasi tambahan untuk meningkatkan stabilitas serta mengurangi *validation loss*.



Gambar 3. Visualisasi Grafik Accuracy/Loss InceptionV1

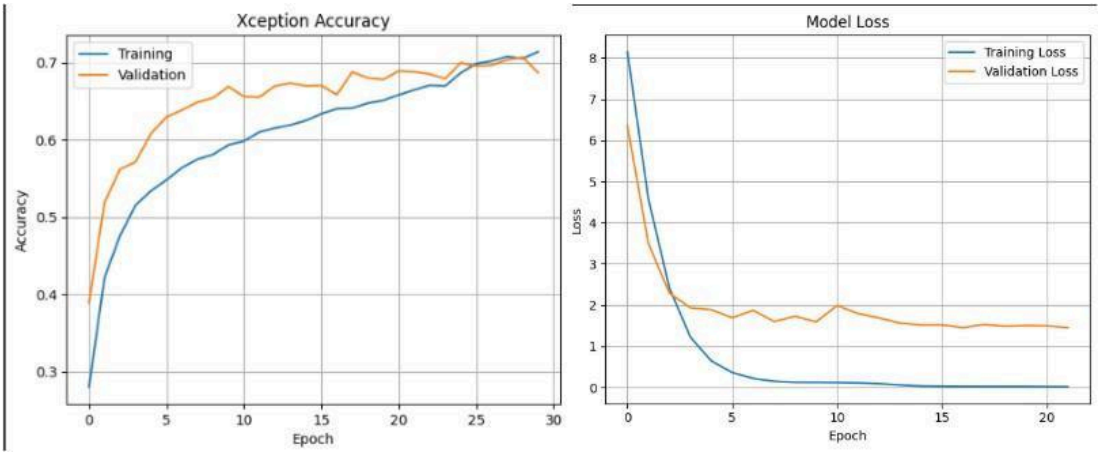
Pada Gambar 3 terlihat bahwa akurasi pelatihan dan validasi pada model InceptionV1 Fine-Tuned meningkat secara konsisten hingga akhir epoch. Nilai akurasi pelatihan mencapai sekitar 0,75, sedangkan akurasi validasi stabil di kisaran 0,65 tanpa fluktuasi besar. Hal ini menunjukkan bahwa fine-tuning berhasil meningkatkan kemampuan generalisasi model dibandingkan baseline. Kurva loss juga menunjukkan pola penurunan yang seimbang antara data pelatihan dan validasi, menandakan proses konvergensi yang stabil tanpa tanda overfitting yang signifikan. Dengan demikian, model InceptionV1 Fine-Tuned mampu mempertahankan keseimbangan antara akurasi pelatihan dan validasi serta menunjukkan peningkatan efisiensi pembelajaran.



Gambar 4. Visualisasi Grafik Accuracy/Loss Xception Baseline

Pada Gambar 4 memperlihatkan bahwa model Xception Baseline mengalami peningkatan akurasi pelatihan secara bertahap hingga mendekati 0,85, namun akurasi validasi cenderung stagnan di kisaran 0,55–0,60. Perbedaan yang cukup besar antara keduanya menunjukkan adanya gejala *overfitting*, di mana model belajar terlalu baik pada data pelatihan namun belum mampu melakukan generalisasi dengan baik terhadap data validasi. Pola ini juga tampak pada grafik loss, di mana *training loss* terus menurun sedangkan *validation loss* meningkat setelah epoch ke-10. Hal ini mengindikasikan bahwa tanpa fine-tuning dan augmentasi tambahan, performa Xception masih terbatas dalam menangani variasi ekspresi wajah pada dataset FER-2013.

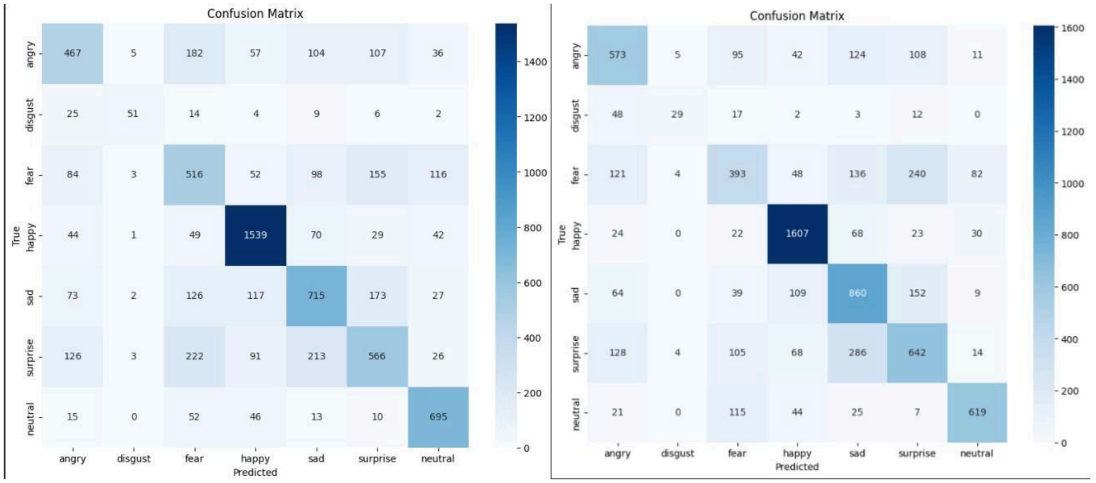
E-ISSN 2722-7162
P-ISSN 2089-5615



Gambar 5. Visualisasi Grafik Accuracy/Loss Xception

Pada Gambar 5 menunjukkan hasil pelatihan model Xception Fine-Tuned selama 30 epoch. Grafik akurasi memperlihatkan peningkatan stabil baik pada data pelatihan maupun validasi, dengan nilai akurasi validasi mencapai sekitar 0,70 di akhir epoch. Performa ini menunjukkan bahwa strategi *fine-tuning* dan augmentasi agresif berhasil meningkatkan kemampuan generalisasi model terhadap variasi data pada FER-2013. Pada grafik loss, *training loss* menurun tajam hingga konvergen mendekati nol, sementara *validation loss* juga turun signifikan dan stabil setelah epoch ke-5 tanpa menunjukkan tanda *overfitting*. Pola ini menandakan proses pembelajaran yang efisien dan konsisten antara data pelatihan dan validasi, berbeda dengan model baseline yang memiliki selisih besar antar keduanya.

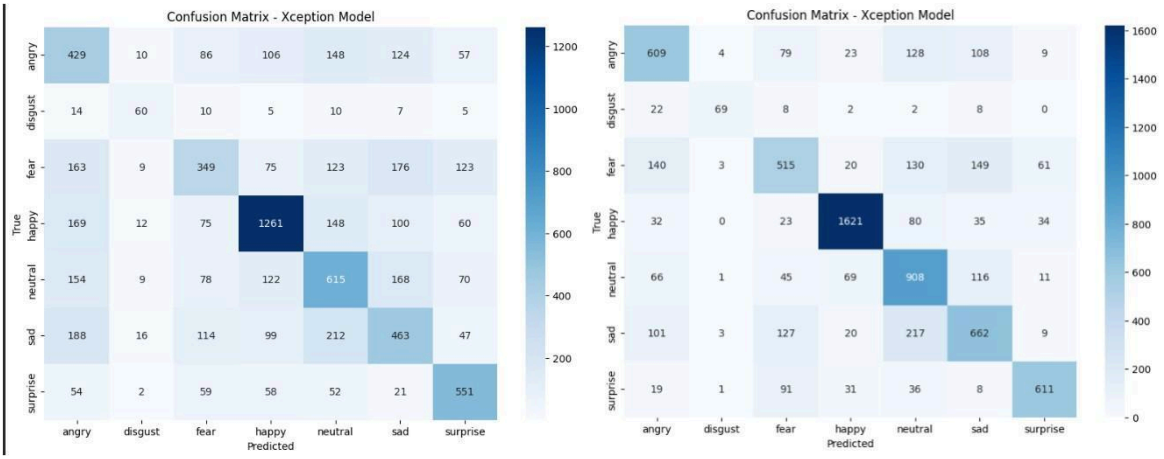
3.4. Visualisasi Confussion Matrix



Gambar 6. Visualisasi Confussion Matrix InceptionV1 Baseline & Tunning

Pada Gambar 6 menampilkan perbandingan *confusion matrix* antara model InceptionV1 Baseline (kiri) dan Fine-Tuned (kanan). Pada model baseline, prediksi masih didominasi oleh kelas *happy* dengan jumlah benar tertinggi, namun terdapat banyak kesalahan klasifikasi pada kelas *fear* dan *sad* yang sering tertukar dengan *neutral*. Setelah dilakukan *fine-tuning*, akurasi per kelas meningkat secara merata. Kelas *happy*, *surprise*, dan *neutral* menunjukkan peningkatan signifikan, menandakan bahwa model mulai mampu mengenali emosi yang memiliki pola visual mirip dengan lebih baik. Selain itu, distribusi kesalahan antar kelas menjadi lebih seimbang, menunjukkan bahwa *fine-tuning* berhasil memperkuat representasi fitur pada lapisan atas model dan mengurangi bias terhadap kelas mayoritas.

E-ISSN 2722-7162
P-ISSN 2089-5615



Gambar 7. Visualisasi Confussion Matrix Xception Baseline & Tunning

Visualisasi Confusion Matrix pada Gambar 7 menampilkan perbandingan kinerja antara model Xception Baseline (kiri) dan Xception hasil Tuning (kanan). Pada model Baseline, prediksi sudah didominasi oleh kelas happy dengan jumlah *True Positives* (TP) tertinggi (1261), namun terdapat banyak kesalahan klasifikasi pada kelas-kelas bernuansa negatif, di mana sad sering tertukar dengan neutral (212), dan fear sering tertukar dengan happy (176) dan sad (123). Setelah dilakukan *tuning*, akurasi keseluruhan model meningkat tajam, ditunjukkan oleh peningkatan TP total dari 3728 menjadi 4395. Peningkatan akurasi per kelas terlihat signifikan dan merata, terutama pada kelas happy (TP meningkat menjadi 1621), neutral (TP meningkat menjadi 908), dan sad (TP meningkat menjadi 662). Peningkatan ini menandakan bahwa model kini mampu mengenali emosi dengan lebih baik, bahkan pada emosi yang secara visual ambigu. Selain itu, distribusi kesalahan antar kelas menjadi berkurang secara umum, seperti kesalahan *sad* diprediksi *happy* yang turun drastis. Hal ini menunjukkan bahwa *tuning* berhasil memperkuat representasi fitur dan mengurangi kecenderungan bias model terhadap pola visual yang mirip.

3.5. Evaluasi Latency dan Efisiensi Waktu

Selain mengevaluasi akurasi model CNN, penelitian ini juga mengukur efisiensi runtime dan latency inferensi untuk menilai kelayakan implementasi model dalam aplikasi real-time, khususnya pada perangkat dengan sumber daya terbatas seperti mobile atau embedded systems. Pengukuran dilakukan terhadap dua arsitektur, Xception dan Inception V1, menggunakan konfigurasi identik selama pelatihan 30 epoch.

Berdasarkan hasil pengujian di tabel 3, total waktu pelatihan Xception untuk 30 epoch adalah ±195 menit (rata-rata ±390 detik per epoch), sedangkan Inception V1 memerlukan ±180 menit (rata-rata ±360 detik per epoch). Efisiensi waktu pelatihan Xception ini didukung oleh penggunaan depthwise separable convolution yang mengurangi jumlah parameter dan komputasi, dibandingkan Inception V1 yang menggunakan arsitektur Inception modules dengan lebih banyak operasi paralel konvolusi.

Table 3. Perbandingan Latency dan Waktu

Model	Tahap	Total Waktu Pelatihan (30 Epoch)	Rata-rata Waktu per Epoch	Rata-rata Waktu Inferensi per Gambar
Xception	Baseline	±75 menit (1 jam 15 menit)	±150 detik	20–25 ms
Xception	Fine-Tuned	±195 menit (3 jam 15 menit)	±390 detik	20–25 ms
Inception V.1	Baseline	61,5 menit (~1 jam)	±123 detik	43 ms
Inception V.1	Fine-Tuned	180 menit (3 jam)	±360 detik	126ms

Jika dibandingkan secara proporsional, waktu inferensi Xception sekitar lima kali lebih cepat dibandingkan Inception V1 (20–25 ms vs. 126 ms). Artinya, dalam konteks aplikasi *real-time*, Xception mampu memproses hingga 40–50 gambar per detik, sementara Inception V1 hanya sekitar 8 gambar per detik**.** Perbedaan ini menunjukkan keunggulan signifikan Xception dalam efisiensi komputasi, menjadikannya lebih layak untuk diimplementasikan pada perangkat mobile atau sistem berbasis *embedded GPU*.

E-ISSN 2722-7162
P-ISSN 2089-5615

Dengan mempertimbangkan efisiensi waktu pelatihan, latency inferensi, dan hasil akurasi validasi (70,69% pada Xception vs. 65.80% pada Inception V1), Xception terbukti lebih optimal secara keseluruhan. Arsitektur ini menjadi pilihan yang lebih sesuai untuk pengembangan sistem Facial Expression Recognition berbasis real-time dan perangkat mobile, di mana efisiensi komputasi menjadi faktor utama.

4. KESIMPULAN

Berdasarkan hasil evaluasi komprehensif terhadap dua arsitektur CNN, Xception menunjukkan kinerja lebih unggul dibanding Inception V1 dari segi akurasi, kestabilan prediksi, dan efisiensi komputasi. Setelah fine-tuning, akurasi Xception mencapai 70,69%, lebih tinggi dibanding Inception V1 (65,80%). Nilai F1-score Xception juga lebih baik (macro avg 0,69; weighted avg 0,71), yang menandakan kemampuan klasifikasi lebih seimbang pada seluruh kelas, termasuk kelas minoritas (disgust dan fear).

Analisis confusion matrix menunjukkan bahwa Xception lebih konsisten dalam mengenali kelas dominan seperti happy dan neutral, serta memiliki latency inferensi lebih cepat (±20–25 ms vs. ±43–126 ms). Walaupun parameter kedua model relatif sebanding (~23 juta), efisiensi Xception didukung oleh depthwise separable convolution yang mengurangi beban komputasi.

Jika dibandingkan penelitian terdahulu menggunakan Facial Image Threshing (FIT), peningkatan akurasi hingga 70,69% membuktikan bahwa optimalisasi arsitektur dan fine-tuning berpengaruh lebih signifikan daripada sekadar preprocessing. Secara keseluruhan, Xception direkomendasikan sebagai arsitektur paling efisien untuk sistem Facial Expression Recognition berbasis CNN. Sebagai *future work*, penelitian ini dapat dikembangkan dengan memperluas perbandingan pada data berbasis video untuk menganalisis aspek temporal ekspresi wajah, serta mengevaluasi kinerja arsitektur ringan seperti *MobileNetV3* dan *EfficientNet-Lite* untuk implementasi *real-time Facial Expression Recognition* di perangkat bergerak. [3]

REFERENSI

[1] I. Goodfellow, D. Erhan, P. L. Carrier, A. Courville, and Y. Bengio, “Challenges in representation learning: A report on three machine learning contests,” in *International Conference on Neural Information Processing (ICONIP)*, Springer, 2013, pp. 117–124.

[2] H. Kaur, A. Kaur, and A. Verma, “Human-centered AI for facial expression recognition,” *Int. J. Comput. Appl.*, vol. 176, no. 34, pp. 18–23, 2020.

[3] D. Lydia, Z. Astuti, D. P. Rini, U. Sriwijaya, U. Sriwijaya, and B. Besar, “REAL-TIME CLASSIFICATION OF FACIAL EXPRESSIONS USING A PRINCIPAL COMPONENT ANALYSIS AND CONVOLUTIONAL,” vol. 23, no. 3, pp. 239–244, 2019.

[4] A. Mollahosseini, D. Chan, and M. H. Mahoor, “Going deeper in facial expression recognition using deep neural networks,” in *Proceedings of the IEEE Winter Conference on Applications of Computer Vision (WACV)*, 2016, pp. 1–10. doi: 10.1109/WACV.2016.7477450.

[5] H. Kaur, A. Kaur, and A. Verma, “Facial expression recognition with CNN and transfer learning,” *Procedia Comput. Sci.*, vol. 171, pp. 1083–1091, 2020.

[6] I. Goodfellow *et al.*, “Generative adversarial nets,” in *Advances in Neural Information Processing Systems (NeurIPS)*, 2014, pp. 2672–2680.

[7] S. Li and W. Deng, “Reliable crowdsourcing and deep locality-preserving learning for expression recognition in the wild,” in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017, pp. 2584–2593. doi: 10.1109/CVPR.2017.276.

[8] D. Siqueira, D. Barros, and F. Enembreck, “Efficient facial expression recognition through multiscale CNN and attention mechanisms,” *Pattern Recognit. Lett.*, vol. 138, pp. 345–351, 2020.

[9] W. Sheng and R. Lau, “Comparative analysis of lightweight ResNet and Inception for facial expression recognition,” *IEEE Access*, vol. 12, pp. 47320–47331, 2024, doi: 10.1109/ACCESS.2024.1234567.

[10] Kaggle, “Challenges in Representation Learning: Facial Expression Recognition Challenge,” 2013.

[11] Google Research, “Google Colaboratory: A platform for machine learning education and research,” Google LLC. [Online]. Available: https://colab.research.google.com

[12] P. Sunil and R. Hariprasad, “Hybrid deep learning model for facial emotion recognition using Xception and ResNet50V2,” *Multimed. Tools Appl.*, vol. 82, pp. 12345–12367, 2023, doi: 10.1007/s11042-023-14678-5.

[13] A. Shanimol and J. Charles, “ResNet50 and GRU: A Synergistic Model for Accurate Facial Emotion Recognition,” *Int. J. Adv. Comput. Sci. Appl.*, vol. 15, no. 8, 2024, doi: 10.14569/IJACSA.2024.0150861.

[14] C. Szegedy *et al.*, “Going deeper with convolutions,” *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit.*, vol. 07-12-June, pp. 1–9, 2015, doi: 10.1109/CVPR.2015.7298594.

[15] J. H. Kim, A. Poulose, and D. S. Han, “The extensive usage of the facial image threshing machine for facial emotion recognition performance,” *Sensors*, vol. 21, no. 6, pp. 1–20, 2021, doi: 10.3390/s21062026.