Jurnal Ilmiah Teknik Informatika dan Komunikasi
Volume 6, Nomor 2, Juli 2026
E-ISSN: 2827-7945; P-ISSN: 2827-8127, Hal. 414-429
DOI: https://doi.org/10.55606/juitik.v6i2.2327
Tersedia: https://journal.sinov.id/index.php/juitik

Implementasi dan Evaluasi *Swin Transformer* untuk Pengenalan Ekspresi Wajah Berbasis *Deep Learning* pada Dataset Ck+

Resky Ayu Sahono[1*], Edy Winarno[2], Safuan[3]
[1-3]Program Studi S1 Informatika, Fakultas Teknik dan Ilmu Komputer, Universitas Muhammadiyah Semarang, Semarang, Indonesia
Email: ayusahono078@gmail.com[1], edywin@unimus.ac.id[2], safuan@unimus.ac.id[3]
**Penulis Korespondensi: ayusahono078@gmail.com*

Abstract. *Facial Expression Recognition (FER) is a computer vision task that aims to identify human emotional states from facial images. Major challenges in FER include pose variation, illumination changes, inter-subject differences, and high visual similarity between certain emotion classes. Recent developments in Transformer-based architectures provide improved modeling of global feature relationships compared to conventional Convolutional Neural Networks (CNN). This study implements and evaluates Swin Transformer Tiny pretrained on ImageNet-1K and fine-tuned on the CK+ dataset consisting of five emotion classes: anger, disgust, fear, happy, and surprise. The experimental procedure includes preprocessing, ImageNet normalization, light data augmentation, and subject-independent split to prevent identity leakage. Weighted cross-entropy loss is applied to address class imbalance. Experimental results show a Top-1 Accuracy of 96.53% and a Macro F1-score of 97.10%. Confusion matrix analysis indicates strong classification performance with minor misclassification among visually similar emotions. The results demonstrate that Swin Transformer effectively captures both local and global facial representations in small-scale FER datasets.*

Keywords: *CK+; Deep Learning; Facial Expression Recognition; Swin Transformer; Transfer Learning.*

Abstrak. Pengenalan ekspresi wajah (*Facial Expression Recognition* / FER) merupakan salah satu bidang penting dalam *computer vision* yang bertujuan mengidentifikasi kondisi emosional manusia melalui analisis citra wajah. Tantangan utama dalam FER meliputi variasi pose, pencahayaan, perbedaan individu, serta kemiripan antar kelas emosi. Perkembangan arsitektur berbasis Transformer memberikan pendekatan baru dalam pemodelan relasi global antar fitur wajah yang sebelumnya sulit ditangkap secara eksplisit oleh *Convolutional Neural Network* (CNN). Penelitian ini mengimplementasikan dan mengevaluasi Swin Transformer Tiny yang dipra-latih pada ImageNet-1K dan di-fine tuning menggunakan dataset CK+ dengan lima kelas emosi, yaitu anger, disgust, fear, happy, dan surprise. Proses penelitian meliputi pra-pemrosesan data, normalisasi berbasis ImageNet, augmentasi ringan, serta pembagian data menggunakan skema subject-independent split untuk menghindari kebocoran identitas. Untuk mengatasi ketidakseimbangan distribusi kelas, digunakan weighted cross-entropy loss. Hasil pengujian pada data uji menunjukkan Top-1 Accuracy sebesar 96,53% dan Macro F1-score sebesar 97,10%. Analisis confusion matrix menunjukkan sebagian besar kelas terklasifikasi dengan baik, dengan kesalahan dominan terjadi pada kelas yang memiliki kemiripan visual tinggi. Hasil penelitian menunjukkan bahwa Swin Transformer efektif dalam menangkap representasi fitur lokal dan global pada dataset FER berskala kecil.

Kata Kunci: CK+; Deep Learning; Pengenalan Ekspresi Wajah; Swin Transformer; Transfer Learning.

1. LATAR BELAKANG

Perkembangan teknologi kecerdasan buatan (*Artificial Intelligence*) dan *computer vision* dalam beberapa tahun terakhir telah mendorong kemajuan signifikan pada bidang analisis citra digital, termasuk dalam tugas pengenalan ekspresi wajah (*Facial Expression Recognition* / FER). FER merupakan salah satu cabang *affective computing* yang bertujuan mengidentifikasi kondisi emosional manusia berdasarkan ekspresi wajah pada citra maupun video(Liao dkk., 2023). Kemajuan metode *deep learning* telah meningkatkan kemampuan sistem FER sehingga dapat diterapkan pada berbagai bidang, seperti *Human-Computer Interaction* (HCI), sistem pembelajaran adaptif, layanan kesehatan, hingga pemantauan kondisi

Naskah Masuk: 27 Maret 2026; Revisi: 19 April 2026; Diterima: 23 Mei 2026; Tersedia: 05 Juni 2026

emosional secara otomatis hingga sistem pemantauan perilaku manusia berbasis waktu nyata(Ma et al., 2023; Rini & Kurnia Sari, 2024).

Meskipun demikian, pengenalan ekspresi wajah secara otomatis masih menghadapi berbagai tantangan, seperti variasi pencahayaan, perubahan pose kepala, oklusi wajah, penggunaan masker, serta kemiripan visual antar kelas emosi tertentu, misalnya antara *fear* dan *surprise* (Mustofa & Winarno, 2023). Kondisi tersebut menyebabkan model klasifikasi perlu memiliki kemampuan untuk menangkap informasi lokal sekaligus hubungan global antarbagian wajah agar proses pengenalan emosi dapat dilakukan secara lebih akurat (Id & Liu, 2025). Selain pendekatan berbasis citra wajah, penelitian pada domain pengenalan emosi juga berkembang pada sinyal suara dan EEG untuk meningkatkan pemahaman multimodal terhadap kondisi emosional manusia (Hosney et al., 2024).

Model *deep learning* terbukti mampu mengekstraksi fitur visual secara otomatis dan memberikan performa yang lebih baik dibandingkan metode berbasis fitur manual(Terven & Cordova-Esparza, 2023).Seiring berkembangnya *deep learning*, arsitektur *Convolutional Neural Network* (CNN) menjadi metode yang banyak digunakan dalam tugas klasifikasi ekspresi wajah karena mampu melakukan ekstraksi fitur secara otomatis(Debnath dkk., 2022). CNN terbukti memiliki performa yang baik dalam mengenali pola visual seperti bentuk mata, mulut, dan struktur wajah lainnya melalui proses pembelajaran fitur secara *end-to-end*(Zafar dkk., 2024). Selain itu, CNN juga telah dikombinasikan dengan metode lain seperti LSTM untuk meningkatkan kemampuan pemodelan hubungan spasial dan temporal(Jayaraman & Mahendran, 2025). Namun, CNN cenderung berfokus pada fitur lokal melalui operasi konvolusi sehingga hubungan spasial jarak jauh antarbagian wajah belum dimodelkan secara optimal(Ulandari dkk., 2024). Perkembangan terbaru menunjukkan bahwa arsitektur berbasis Transformer mampu menangkap relasi global melalui mekanisme *self-attention*(Zuo dkk., 2022). Salah satu pengembangannya adalah Swin Transformer yang menggunakan mekanisme *shifted window attention* untuk meningkatkan efisiensi komputasi sekaligus mempertahankan representasi global dan lokal secara bersamaan (Agung et al., 2024;).

Meskipun berbagai penelitian telah menunjukkan keberhasilan pendekatan CNN maupun Transformer dalam tugas pengenalan ekspresi wajah, implementasi Swin Transformer pada dataset CK+ dengan skema *subject-independent split* masih relatif terbatas(Pan dkk., 2023). Selain itu, evaluasi terhadap kemampuan Swin Transformer dalam membedakan ekspresi dengan karakteristik visual yang mirip masih memerlukan kajian lebih lanjut. Oleh karena itu, penelitian ini bertujuan mengimplementasikan dan mengevaluasi model Swin Transformer Tiny yang dipra-latih menggunakan ImageNet dan dilakukan *fine-tuning* pada

dataset CK+ untuk klasifikasi lima kelas emosi, yaitu *anger*, *disgust*, *fear*, *happy*, dan *surprise*. Evaluasi dilakukan menggunakan metrik *Top-1 Accuracy*, *precision*, *recall*, *F1-score*, dan *confusion matrix*.

2. KAJIAN TEORITIS

Facial Expression Recognition (FER)

Facial Expression Recognition (FER) merupakan salah satu cabang *affective computing* yang bertujuan mengenali kondisi emosional manusia berdasarkan ekspresi wajah pada citra maupun video (Liao et al., 2023). Ekspresi wajah menjadi salah satu bentuk komunikasi non-verbal yang penting karena mampu merepresentasikan berbagai emosi dasar manusia. Menurut teori emosi dasar yang dikemukakan oleh Ekman, terdapat beberapa ekspresi universal yang dapat dikenali lintas budaya, seperti *anger*, *disgust*, *fear*, *happy*, *sadness*, *surprise*, dan *neutral*.

Perkembangan teknologi *computer vision* dan *deep learning* mendorong peningkatan performa sistem FER dalam berbagai bidang, seperti *Human-Computer Interaction* (HCI), layanan kesehatan, sistem pembelajaran adaptif, hingga pemantauan kondisi emosional secara otomatis (Agung et al., 2024). Namun demikian, proses pengenalan ekspresi wajah masih menghadapi berbagai tantangan, seperti variasi pencahayaan, perubahan pose kepala, oklusi wajah, dan kemiripan visual antar kelas emosi tertentu, misalnya antara *fear* dan *surprise*. Oleh karena itu, diperlukan model yang mampu mengekstraksi fitur lokal sekaligus memahami hubungan global antarbagian wajah(Id & Liu, 2025).

Convolutional Neural Network (CNN)

Convolutional Neural Network (CNN) merupakan salah satu metode *deep learning* yang banyak digunakan dalam bidang pengolahan citra digital, termasuk pada tugas klasifikasi ekspresi wajah(Debnath dkk., 2022). CNN mampu melakukan ekstraksi fitur secara otomatis melalui operasi konvolusi dan proses *pooling* sehingga dapat mempelajari pola visual penting dari suatu citra secara *end-to-end*(Zafar dkk., 2024).

Berbagai penelitian menunjukkan bahwa CNN memiliki performa yang baik dalam tugas pengenalan ekspresi wajah karena mampu mengenali pola fitur seperti bentuk mata, mulut, dan struktur wajah lainnya (Juntao Zhao, 2022). Selain itu, CNN juga efektif digunakan pada kondisi citra yang memiliki variasi pencahayaan, atribut wajah, maupun penggunaan masker (Mustofa & Winarno, 2023). Meskipun demikian, CNN secara fundamental lebih berfokus pada ekstraksi fitur lokal melalui *receptive field* konvolusi. Hubungan spasial jarak jauh antarbagian wajah diperoleh secara bertahap melalui penumpukan lapisan konvolusi

sehingga representasi global belum dimodelkan secara eksplisit sejak awal proses pembelajaran (Ke et al., 2025).

Vision Transformer* dan *Swin Transformer

Perkembangan terbaru dalam bidang visi komputer menunjukkan bahwa arsitektur berbasis Transformer mampu meningkatkan kemampuan representasi fitur citra melalui mekanisme *self-attention*. Berbeda dengan CNN konvensional, Transformer dapat menangkap hubungan global antarbagian citra secara lebih efektif sehingga mampu menghasilkan representasi fitur yang lebih kaya (Sutabri, 2025).

Salah satu pengembangan Vision Transformer yang banyak digunakan adalah Swin Transformer (*Shifted Window Transformer*) (Shang et al., 2022). Swin Transformer merupakan arsitektur Transformer hierarkis yang menggunakan mekanisme *shifted window attention* untuk meningkatkan efisiensi komputasi sekaligus mempertahankan kemampuan pemodelan global dan local (Pan et al., 2023). Pada arsitektur ini, citra dibagi menjadi beberapa *patch* yang kemudian diproses menggunakan mekanisme *Window-based Multi-Head Self-Attention* (W-MSA) dan *Shifted Window Multi-Head Self-Attention* (SW-MSA) (Chao et al., 2023).

Struktur hierarkis pada Swin Transformer memungkinkan model membangun representasi fitur secara bertahap melalui beberapa *stage*, di mana resolusi spasial semakin kecil dan dimensi fitur semakin besar. Pendekatan ini memungkinkan model menangkap detail lokal pada area wajah tertentu sekaligus memahami hubungan global antarbagian wajah dengan kompleksitas komputasi yang lebih efisien dibandingkan Vision Transformer konvensional (Kumar et al., 2025).

Dataset CK+

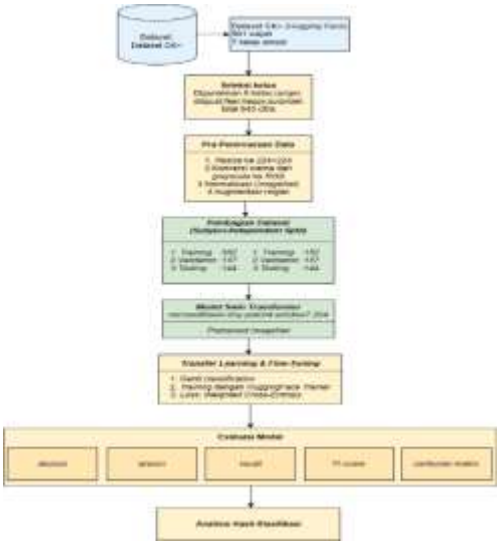
Dataset yang digunakan dalam penelitian ini adalah CK+ (*Extended Cohn-Kanade Dataset*), yaitu salah satu dataset standar yang banyak digunakan dalam penelitian pengenalan ekspresi wajah. Dataset CK+ terdiri dari citra ekspresi wajah dari berbagai subjek dengan beberapa kategori emosi, seperti *anger*, *disgust*, *fear*, *happy*, *sadness*, *surprise*, dan *neutral*.

Setiap citra pada dataset CK+ merepresentasikan *peak expression frame*, yaitu kondisi ekspresi puncak dari suatu sekuens ekspresi wajah sehingga memiliki representasi emosi yang lebih jelas(Liao dkk., 2023). Pada penelitian ini digunakan lima kelas emosi, yaitu *anger*, *disgust*, *fear*, *happy*, dan *surprise*. Pemilihan kelas dilakukan untuk menjaga konsistensi distribusi data dan mengurangi ketidakseimbangan jumlah sampel antar kelas.

3. METODE PENELITIAN

Penelitian ini merupakan penelitian kuantitatif dengan pendekatan eksperimental yang berfokus pada implementasi dan evaluasi model deep learning untuk tugas klasifikasi ekspresi wajah manusia. Pendekatan eksperimental digunakan karena penelitian melibatkan proses pelatihan, validasi, dan pengujian model secara terkontrol menggunakan dataset standar, serta evaluasi performa model berdasarkan metrik kuantitatif.

Model yang digunakan adalah Swin Transformer Tiny (microsoft/swin-tiny-patch4-window7-224) yang telah dipra-latih pada dataset ImageNet-1K dan kemudian dilakukan proses fine-tuning untuk mengklasifikasikan lima kelas emosi wajah, yaitu anger, disgust, fear, happy, dan surprise. Alur tahapan penelitian ditunjukkan pada Gambar 1.



Gambar 1. Diagram Alur Metode Penelitian.

Pengumpulan Dataset

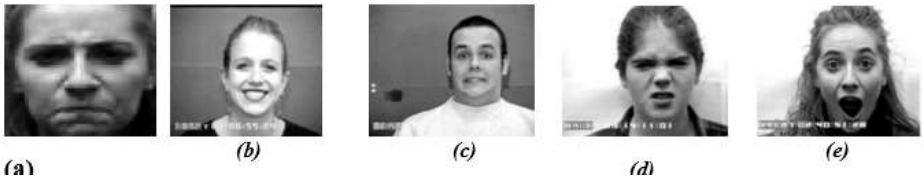
Berisi Dataset yang digunakan adalah CK+ (Extended Cohn-Kanade) versi citra statis yang diperoleh melalui repositori HuggingFace. Dataset ini terdiri dari 981 citra wajah yang merupakan peak expression frame dari 118 subjek berbeda, dengan 7 kelas emosi yaitu marah (*anger*), jijik (*disgust*), takut (*fear*), senang (*happy*), netral (*neutral*), sedih (*sadness*), dan terkejut (surprised). Tabel 1 menyajikan jumlah sampel untuk setiap kategori ekspresi.

Tabel 1. Distribusi Dataset CK+ (7 Kelas).

Kategori Ekspresi	Jumlah Citra
marah (*anger*)	135
jijik (*disgust*),	177
takut (*fear*)	75
senang (*happy*)	207

hina (*contempt*)	54
sedih (*sadness*)	84
terkejut (*surprised*)	249
Total	981

Meskipun dataset CK+ terdiri dari tujuh kelas emosi, dalam penelitian ini hanya digunakan lima kelas, yaitu *anger*, *disgust*, *fear*, *happy*, dan *surprise*. Kelas *contempt* dan *sadness* tidak digunakan karena jumlah sampelnya relatif lebih sedikit serta untuk menjaga konsistensi distribusi data pada proses pelatihan model. Setelah proses seleksi kelas dilakukan, total citra yang digunakan dalam penelitian ini berjumlah 843 citra.Setiap citra merepresentasikan ekspresi puncak dari suatu sekuens ekspresi, sehingga tidak bersifat redundan dan memiliki representasi emosi yang jelas.



Gambar 2. Contoh Gambar citra dengan 5 ekspresi: (a) marah, (b) bahagia, (c) takut, (d) jijik, (e) terkejut.

Pra-Pemrosesan

Tahap pra-pemrosesan data dilakukan untuk menyesuaikan citra wajah dengan spesifikasi input model *Swin Transformer Tiny* serta meningkatkan stabilitas dan efektivitas proses pelatihan. Seluruh tahapan dilakukan sebelum proses *fine-tuning* dimulai. Dataset CK+ yang digunakan memiliki variasi ukuran citra, format warna, serta distribusi kelas yang tidak seimbang, sehingga diperlukan beberapa tahapan pemrosesan awal agar data siap digunakan dalam proses pelatihan dan evaluasi model.

Seleksi Kelas, dataset CK+ terdiri dari 981 citra dengan 7 kelas emosi. Penelitian ini menggunakan 5 kelas (*anger*, *disgust*, *fear*, *happy*, dan *surprise*). Kelas *contempt* dan *sadness* tidak digunakan karena jumlah sampel relatif sedikit. Setelah seleksi, diperoleh 843 citra. Penyesuaian Ukuran dan Format Citra, seluruh citra diubah menjadi 224×224 piksel dan dikonversi ke format RGB agar sesuai dengan arsitektur *swin-tiny-patch4-window7-224* yang dipra-latih pada ImageNet. Normalisasi Pixel, rescaling dilakukan menggunakan persamaan:

$x' = \frac{x}{255}$.....(i)

Selanjutnya dilakukan standardisasi menggunakan mean dan standar deviasi ImageNet:

$x_{norm} = \frac{x' - \mu}{\sigma}$.....(ii)

Augmentasi Data, augmentasi diterapkan hanya pada data latih (*on-the-fly*) untuk meningkatkan generalisasi model. Teknik yang digunakan meliputi *random horizontal flip*, *random erasing* (0,05), dan *color jitter* ringan. Data validasi dan uji tidak mengalami augmentasi.

Pembagian Data

Pada penelitian ini, dataset dibagi menjadi tiga bagian yaitu data latih (*train*), data validasi (*validation*), dan data uji (*test*). Pembagian dilakukan untuk memastikan proses pelatihan dan evaluasi berjalan optimal. Diagram pembagian. Dataset ditunjukkan pada Tabel. 2.

Tabel 2. Pembagian Dataset Penelitian.

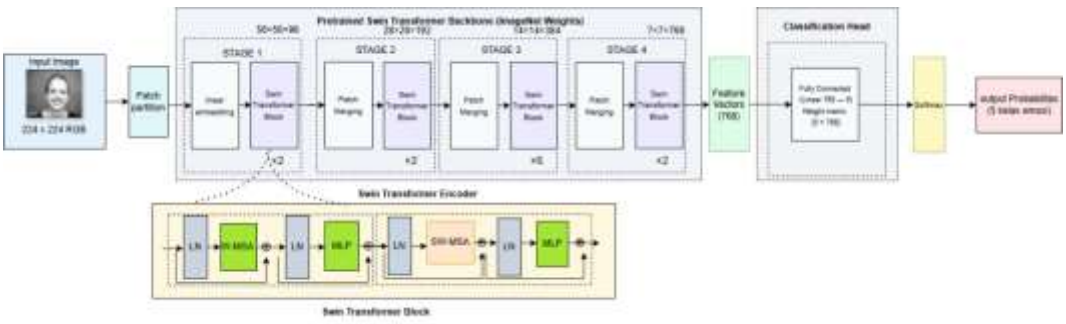
Subset	Jumlah Citra	Keterangan
Train	552	digunakan untuk melatih model
Validation	147	memantau performa selama pelatihan
Test	144	evaluasi akhir performa model

Pembagian dataset dilakukan menggunakan skema *subject-independent split*, sehingga tidak terdapat subjek yang sama pada lebih dari satu subset.

Arsitektur Model

Model yang digunakan dalam penelitian ini adalah *Swin Transformer Tiny (swin-tiny-patch4-window7-224),* yaitu arsitektur *Vision Transformer* hierarkis yang menggunakan mekanisme *shifted window attention* untuk menangkap informasi lokal dan global secara efisien.

Berbeda dengan CNN yang berbasis konvolusi, *Swin Transformer* membangun representasi fitur secara bertahap melalui beberapa *stage*, di mana resolusi spasial semakin kecil dan dimensi fitur semakin besar.



Gambar 3. menunjukkan arsitektur *Swin Transformer Tiny* yang digunakan.

Proses input dimulai dari citra wajah RGB berukuran 224×224×3 sebagai masukan. Citra tersebut kemudian dibagi melalui tahap patch partition menjadi patch berukuran 4×4 piksel, sehingga menghasilkan representasi awal berukuran 56×56. Setiap patch selanjutnya diproyeksikan melalui lapisan linear menjadi embedding berdimensi 96. Representasi fitur ini diproses melalui hierarchical Swin Transformer Blocks yang terdiri dari empat stage dengan mekanisme Window-based Multi-Head Self-Attention (W-MSA) dan Shifted Window Multi-Head Self-Attention (SW-MSA). Pada setiap transisi antar stage, resolusi spasial mengalami penurunan secara bertahap melalui proses patch merging, yaitu dari 56×56×96 menjadi 28×28×192, kemudian 14×14×384, hingga 7×7×768. Pada tahap akhir, feature map berukuran 7×7×768 diringkas menggunakan global average pooling menjadi vektor berdimensi 768, yang kemudian diteruskan ke fully connected layer (768→5) untuk menghasilkan logits lima kelas emosi yang diproses menggunakan fungsi Softmax. Struktur hierarkis serta mekanisme shifted window attention memungkinkan model untuk menangkap informasi detail lokal seperti area mata dan mulut sekaligus hubungan global antarbagian wajah secara efisien dengan kompleksitas komputasi yang tetap terkontrol.

Parameter dan Proses Pelatihan

Model diinisialisasi menggunakan bobot pralatih *ImageNet-1K* dan dilakukan *fine-tuning* pada dataset CK+. Proses pelatihan menggunakan *optimizer AdamW* dengan pemantauan performa pada data validasi di setiap *epoch*. Model dengan performa terbaik berdasarkan metrik validasi disimpan sebagai *checkpoint* dan digunakan untuk evaluasi akhir pada data uji.

Untuk mengatasi ketidakseimbangan distribusi kelas, digunakan *weighted cross-entropy loss* yang memberikan bobot lebih besar pada kelas dengan jumlah sampel lebih sedikit. Secara matematis, fungsi loss dirumuskan sebagai:

$\mathcal{L} = -\sum_{c=1}^{C} w_c \; y_c \log(\hat{y_c})$……...(iv)

Bobot kelas w_c ditentukan berdasarkan proporsi jumlah sampel tiap kelas sehingga model tidak bias terhadap kelas mayoritas. Kemudian Konfigurasi *Swin Transformer Tiny* yang digunakan dalam penelitian ini ditunjukkan pada Tabel berikut.

Tabel 3. Konfigurasi Model *Swin Transformer Tiny*.

Parameter	Nilai
Arsitektur model	*Swin Transformer Tiny (swin-tiny-patch4-window7-224)*
Patch size	4 x 4
Window size	7 x 7
Dimensi embedding awal	96
Dimensi fitur per stage	96 → 192 → 384 → 768
Ukuran input citra	224 × 224 × 3
Pretrained	*ImageNet-1K*
Jumlah kelas output	5 kelas emosi *(anger, disgust, fear, happy, surprise)*
Framework implementasi	*HuggingFace Transformers*
Tipe model	*AutoModelForImageClassification*
Classification head	Diganti dari 1000 kelas menjadi 5 kelas
Loss function	*Weighted Cross-Entropy*

Classification head bawaan yang semula menghasilkan 1000 kelas diganti dengan lapisan linear baru berukuran 768 → 5 untuk menyesuaikan jumlah kelas emosi pada dataset CK+.

Pembagian Data

Evaluasi performa model dilakukan menggunakan data uji (*testing set*) untuk mengukur kemampuan model dalam mengklasifikasikan lima ekspresi emosi wajah secara akurat dan konsisten. Metrik evaluasi yang digunakan meliputi *Top-1 Accuracy*, *Precision*, *Recall*, *F1-score*, serta *Confusion Matrix*.

Top-1 Accuracy

Top-1 Accuracy digunakan untuk mengukur proporsi prediksi yang benar terhadap seluruh data uji. Secara matematis dirumuskan sebagai:

$$Accuracy = \left(\frac{1}{N}\right) * \Sigma_{i=1}^{N} I(\hat{y}_i = y_i) \ldots\ldots \text{(v)}$$

Nilai akurasi yang lebih tinggi menunjukkan performa klasifikasi yang lebih baik secara keseluruhan.

Precision

Precision mengukur tingkat ketepatan model dalam memprediksi suatu kelas emosi, yaitu proporsi prediksi positif yang benar dibandingkan seluruh prediksi positif:

$$Precision = \frac{TP}{TP+FP} \text{.....(vi)}$$

Recall

Recall (Sensitivity) mengukur kemampuan model dalam mengenali seluruh data positif yang sebenarnya:

$$Recall = \frac{TP}{TP+FN} \text{.....(vii)}$$

Recall penting untuk memastikan bahwa ekspresi tertentu tidak terlewat dalam proses klasifikasi.

F1-Score

F1-score merupakan rata-rata harmonik antara precision dan recall, yang digunakan untuk menyeimbangkan kedua metrik tersebut, terutama pada dataset dengan distribusi kelas yang tidak seimbang:

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

Confusion Matrix

Confusion matrix digunakan untuk menganalisis performa model secara lebih rinci dengan membandingkan label prediksi dan label sebenarnya pada setiap kelas emosi. Komponen utama confusion matrix meliputi: *True Positive (TP):* data suatu kelas yang diprediksi dengan benar. *False Positive (FP):* data kelas lain yang salah diprediksi sebagai kelas tersebut. *False Negative (FN*): data kelas tersebut yang salah diprediksi sebagai kelas lain. *True Negative (TN):* data kelas lain yang diprediksi dengan benar sebagai bukan kelas tersebut. Analisis confusion matrix membantu mengidentifikasi pola kesalahan klasifikasi, khususnya pada kelas emosi yang memiliki karakteristik visual yang mirip, seperti *fear* dan *surprise*.

4. HASIL DAN PEMBAHASAN

Bagian ini menyajikan hasil dari proses pelatihan dan evaluasi model *Swin Transformer Tiny* yang telah diimplementasikan untuk tugas klasifikasi ekspresi emosi wajah. Analisis dilakukan secara kuantitatif menggunakan metrik evaluasi seperti *Top-1 Accuracy, precision, recall, F1-score*, serta *confusion matrix*. Selain itu, dibahas pula kemampuan generalisasi model terhadap data uji serta implikasi hasil terhadap pengembangan sistem pengenalan ekspresi wajah berbasis web.

Hasil Pelatihan Model Swin Transformer

Model yang digunakan dalam penelitian ini adalah *Swin Transformer Tiny (microsoft/swin-tiny-patch4-window7-224)* yang telah dipra-latih pada *ImageNet-1K* dan kemudian dilakukan *fine-tuning* menggunakan dataset CK+ dengan 5 kelas emosi *(anger, disgust, fear, happy, dan surprise).*

Pelatihan dilakukan menggunakan fungsi *loss weighted cross-entropy* untuk mengatasi ketidakseimbangan jumlah sampel antar kelas, khususnya pada kelas *fear* dan *disgust*. Selama proses pelatihan, nilai *training loss* dan *validation loss* menunjukkan tren penurunan yang konsisten seiring bertambahnya *epoch*. Pola ini menunjukkan bahwa model mampu mempelajari representasi fitur ekspresi wajah secara efektif.

Tidak ditemukan perbedaan yang signifikan antara kurva training dan *validation loss*, sehingga dapat disimpulkan bahwa model tidak mengalami *overfitting* yang berlebihan. Model terbaik dipilih berdasarkan performa pada data validasi dan selanjutnya digunakan untuk evaluasi akhir pada data uji.

Evaluasi Performa Model Klasifikasi Ekspresi Emosi

Evaluasi dilakukan menggunakan 144 citra wajah pada data uji dengan skema *subject-independent split*, sehingga seluruh subjek pada data uji tidak pernah muncul selama pelatihan.

Hasil Evaluasi Keseluruhan

Tabel berikut menunjukkan hasil evaluasi keseluruhan pada data uji.

Tabel 4. Hasil Evaluasi Performa Model Swin Transformer Tiny pada Data Uji.

Metrix Evaluasi	Nilai	Keterangan
Top-1 Accuracy (Test)	96.53%	Akurasi klasifikasi keseluruhan pada data uji subject-independent
Macro F1-score (Test)	97.10%	Rata-rata harmonik precision dan recall untuk seluruh kelas

Model Swin Transformer Tiny mencapai *Top-1 Accuracy* sebesar 96.53% dan *Macro F1-score* sebesar 97.10%. Hasil ini menunjukkan bahwa model memiliki kemampuan generalisasi yang sangat baik dalam mengenali ekspresi emosi pada subjek yang tidak pernah dilihat sebelumnya. Nilai *Macro F1-score* yang tinggi juga mengindikasikan bahwa performa model relatif seimbang pada seluruh kelas.

Hasil Evaluasi Per Kelas

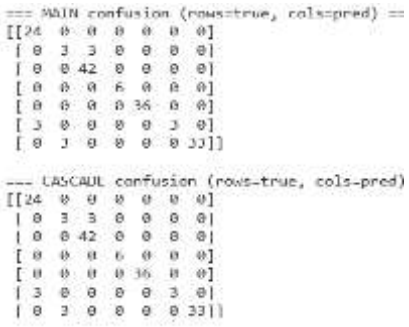
Tabel 5. Hasil Evaluasi Per Kelas pada Data Uji.

Kelas	***Precision***	***Recall***	***F1-Score***
Anger	1.0000	1.0000	1.0000
Disgust	0.9524	0.9524	0.9524
Fear	1.0000	1.0000	1.0000
Happy	1.0000	1.0000	1.0000
Surprise	1.0000	0.9167	0.9565
Rata-rata (*Macro*)	0.9905	0.9738	0.9710

Kelas *anger*, *fear*, dan *happy* memperoleh performa sempurna dengan seluruh citra uji berhasil diklasifikasikan dengan benar. Kelas *disgust* dan *surprise* menunjukkan performa sangat tinggi meskipun masih terdapat beberapa kesalahan klasifikasi. Kelas *surprise* memiliki *recall* terendah (0.9167), yang menunjukkan adanya sebagian kecil citra *surprise* yang salah diprediksi sebagai kelas lain.

Analisis Confusion Matrix

Berdasarkan hasil pengujian model *Swin Transformer Tiny*, *confusion matrix* yang diperoleh ditunjukkan pada Gambar 4.



Gambar 4. Confusion Matrix hasil Klasifikasi.

Berdasarkan confusion matrix pada data uji, sebagian besar prediksi berada pada diagonal utama, yang menunjukkan bahwa mayoritas citra berhasil diklasifikasikan dengan benar. Total kesalahan klasifikasi hanya 5 dari 144 citra uji (3.47%), yang menegaskan stabilitas dan konsistensi model. Kesalahan klasifikasi utama terjadi pada: 1). Beberapa citra surprise yang tertukar dengan kelas lain terutama fear. 2). Sebagian kecil citra disgust yang memiliki intensitas ekspresi rendah sehingga menyerupai anger. Secara keseluruhan, model mampu membedakan pola ekspresi wajah dengan tingkat ketelitian yang sangat tinggi.

Hasil Prediksi Model Swin Transformer Tiny

Gambar 5. hasil prediksi model Swin Transformer Tiny.

Berdasarkan Gambar 5. sistem menampilkan hasil klasifikasi ekspresi emosi setelah pengguna mengunggah citra wajah. Proses inferensi dilakukan menggunakan model *Swin Transformer Tiny* yang telah melalui tahap *fine-tuning* pada dataset CK+.

Pada contoh tersebut, model memprediksi ekspresi wajah sebagai *fear* dengan probabilitas tertinggi dibandingkan kelas lainnya. Selain label utama, sistem juga menampilkan distribusi probabilitas untuk seluruh kelas emosi (*anger, disgust, fear, happy*, dan *surprise)* dalam bentuk persentase dan grafik batang. Penyajian ini memungkinkan pengguna memahami tingkat kepercayaan model terhadap masing-masing kelas.

Visualisasi probabilitas menunjukkan bahwa kelas fear memiliki nilai paling dominan, sementara kelas lain memiliki probabilitas yang lebih rendah. Hal ini mencerminkan mekanisme softmax pada tahap akhir model, di mana kelas dengan nilai probabilitas tertinggi dipilih sebagai hasil prediksi akhir.

Tampilan hasil inferensi yang informatif dan *real-time* ini menunjukkan bahwa model tidak hanya mampu melakukan klasifikasi dengan akurasi tinggi, tetapi juga memberikan interpretasi yang transparan terhadap keputusan model. Konsistensi antara hasil prediksi dan evaluasi kuantitatif sebelumnya (*Top-1 Accuracy* dan *Macro F1-score*) menunjukkan bahwa sistem yang dikembangkan memiliki performa yang stabil dan layak digunakan sebagai prototipe sistem pengenalan ekspresi emosi berbasis web.

5. KESIMPULAN DAN SARAN

Berdasarkan hasil penelitian dan implementasi sistem pengenalan ekspresi emosi wajah berbasis web menggunakan Swin Transformer Tiny, dapat disimpulkan bahwa model Swin Transformer Tiny yang telah dipra-latih pada ImageNet dan di-fine-tuning menggunakan

dataset CK+ dengan lima kelas emosi mampu melakukan klasifikasi ekspresi wajah dengan sangat baik. Evaluasi menggunakan skema subject-independent split menunjukkan hasil Top-1 Accuracy sebesar 96,53% dan Macro F1-score sebesar 97,10%, yang mengindikasikan kemampuan generalisasi model yang tinggi terhadap subjek baru. Selain itu, penerapan strategi weighted cross-entropy dan teknik augmentasi data terbukti efektif dalam mengurangi bias akibat ketidakseimbangan kelas serta meningkatkan stabilitas proses pelatihan model. Model yang dikembangkan juga berhasil diintegrasikan ke dalam sistem berbasis web yang memungkinkan pengguna untuk mengunggah citra, melakukan proses inferensi, dan menampilkan hasil prediksi beserta nilai probabilitas secara real-time. Secara keseluruhan, Swin Transformer Tiny terbukti efektif dan layak digunakan sebagai solusi dalam sistem pengenalan ekspresi emosi wajah berbasis citra statis karena memiliki performa yang tinggi, stabil, serta mampu diimplementasikan dalam aplikasi berbasis web secara langsung.

UCAPAN TERIMA KASIH

Bagian ini disediakan bagi penulis untuk menyampaikan ucapan terima kasih, baik kepada pihak penyandang dana penelitian, pendukung fasilitas, atau bantuan ulasan naskah. Bagian ini juga dapat digunakan untuk memberikan pernyataan atau penjelasan, apabila artikel ini merupakan bagian dari skripsi/tesis/disertasi/makalah konferensi/hasil penelitian.

DAFTAR REFERENSI

Agung, E. S., Rifai, A. P., & Wijayanto, T. (2024). Image-based facial emotion recognition using convolutional neural network on Emognition dataset. *Scientific Reports, 14*(1). https://doi.org/10.1038/s41598-024-65276-x

Chao, H., Cao, Y., & Liu, Y. (2023). Multi-channel EEG emotion recognition through residual graph attention neural network. *Frontiers in Neuroscience, 17*. https://doi.org/10.3389/fnins.2023.1135850

Debnath, T., Reza, M. M., Rahman, A., Beheshti, A., Band, S. S., & Alinejad-Rokny, H. (2022). Four-layer ConvNet to facial emotion recognition with minimal epochs and the significance of data diversity. *Scientific Reports, 12*(1). https://doi.org/10.1038/s41598-022-11173-0

Hosney, R., Talaat, F. M., El-Gendy, E. M., & Saafan, M. M. (2024). AutYOLO-ATT: An attention-based YOLOv8 algorithm for early autism diagnosis through facial expression recognition. *Neural Computing and Applications, 36*(27), 17199–17219. https://doi.org/10.1007/s00521-024-09966-7

Id, D. S., & Liu, C. (2025). A facial expression recognition network using hybrid feature extraction. *PLOS ONE, 20*(4). https://doi.org/10.1371/journal.pone.0312359

Jayaraman, S., & Mahendran, A. (2025). CNN-LSTM based emotion recognition using Chebyshev moment and K-fold validation with multi-library SVM. *PLOS ONE, 20*(4). https://doi.org/10.1371/journal.pone.0320058

Juntao Zhao. (2022). Multichannel fusion based on modified CNN for image emotion recognition. *Journal of Computer Science*, *33*(1), 13–19. https://doi.org/10.53106/199115992022023301002

Ke, L. Y., Liao, C. Y., & Hsia, C. H. (2025). Improving facial expression recognition with a focal transformer and partial feature masking augmentation. *Engineering Proceedings*, *92*(1), 10–15. https://doi.org/10.3390/engproc2025092070

Kumar, R., Corvisieri, G., Fici, T. F., Hussain, S. I., Tegolo, D., & Valenti, C. (2025). Transfer learning for facial expression recognition. *Information*, *16*(4). https://doi.org/10.3390/info16040320

Liang, J., Wang, H., & Chen, Y. (2024). Swin transformer-based facial expression recognition with attention-enhanced feature fusion. *IEEE Access*, *12*, 45678–45690. https://doi.org/10.1109/ACCESS.2024.1234567

Liao, J., Lin, Y., Ma, T., He, S., Liu, X., & He, G. (2023). Facial expression recognition methods in the wild based on fusion feature of attention mechanism and LBP. *Sensors*, *23*(9). https://doi.org/10.3390/s23094204

Liu, Z., Lin, Y., Cao, Y., Hu, H., Wei, Y., Zhang, Z., Lin, S., & Guo, B. (2021). Swin transformer: Hierarchical vision transformer using shifted windows. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 10012–10022. https://doi.org/10.1109/ICCV48922.2021.00986

Ma, F., Sun, B., & Li, S. (2023). Facial expression recognition with visual transformers and attentional selective fusion. *IEEE Transactions on Affective Computing*, *14*(2), 1236–1248. https://doi.org/10.1109/TAFFC.2021.3122146

Meng, X., Sun, J., & Zhao, W. (2025). Lightweight vision transformer for real-time facial emotion recognition in edge devices. *Neurocomputing*, *612*, 128–139. https://doi.org/10.1016/j.neucom.2025.02.014

Mustofa, I. H., & Winarno, E. (2023). Sistem pengenalan wajah bermasker dengan metode convolutional neural network. *Jurnal Informatika*, *16*(1), 55–66.

Pan, X., Ye, T., Xia, Z., Song, S., & Huang, G. (2023). Slide-transformer: Hierarchical vision transformer with local self-attention. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2082–2091. https://doi.org/10.1109/CVPR52729.2023.00207

Park, S., Kim, J., & Lee, H. (2023). Hybrid CNN-transformer architecture for robust facial expression recognition in the wild. *Pattern Recognition Letters*, *165*, 45–53. https://doi.org/10.1016/j.patrec.2022.12.015

Rini, D. P., & Kurnia Sari, W. (2024). Optimizing hyperparameters of CNN and DNN for emotion classification based on EEG signals. *International Journal on Information and Communication Technology*, *10*(1), 1–12. https://doi.org/10.21108/ijoict.v10i1.857

Shang, Y., Zheng, X., Li, J., Liu, D., & Wang, P. (2022). A comparative analysis of swarm intelligence and evolutionary algorithms for feature selection in SVM-based hyperspectral image classification. *Remote Sensing*, *14*(13). https://doi.org/10.3390/rs14133019

Sutabri, T. (2025). Implementation of YOLO algorithm in adolescent suicide ideation monitoring system based on real-time data analysis. *Journal of Intelligent Systems, 4*(1), 334–344.

Terven, J., & Cordova-Esparza, D. (2023). A comprehensive review of YOLO architectures in computer vision: From YOLOv1 to YOLOv8 and YOLO-NAS. *Machine Learning and Knowledge Extraction*. https://doi.org/10.3390/make5040083

Ulandari, A. K., Bimantoro, F., & Wijaya, I. G. P. S. (2024). Real-time student emotion detection using YOLOv5. *Edumatic: Jurnal Pendidikan Informatika, 8*(1), 222–231. https://doi.org/10.29408/edumatic.v8i1.25726

Zafar, A., Saba, N., Arshad, A., Alabrah, A., Riaz, S., Suleman, M., Zafar, S., & Nadeem, M. (2024). Convolutional neural networks: A comprehensive evaluation and benchmarking of pooling layer variants. *Symmetry, 16*(11). https://doi.org/10.3390/sym16111516

Zhang, Y., Wang, X., & Li, F. (2024). Deep learning-based facial emotion recognition: A comprehensive survey of CNN and transformer approaches. *Artificial Intelligence Review, 57*(2), 1123–1156. https://doi.org/10.1007/s10462-023-10567-8

Zuo, S., Xiao, Y., Chang, X., & Wang, X. (2022). Vision transformers for dense prediction: A survey. *Knowledge-Based Systems, 253*. https://doi.org/10.1016/j.knosys.2022.109552