Klasifikasi Ekspresi Wajah Menggunakan HFT-CNN Dan Siamese Network Pada Citra Wajah

1st Elang Satria Putra Buana
Program Studi Informatika, Universitas Telkom, Kampus Surabaya, Jl. Ketintang No. 156, Surabaya 60231, Jawa Timur, Indonesia
elangspb@student.telkomuniversity.ac.id

2nd Ardian Yusuf Wicaksono, S.Kom., M.Kom.
Program Studi Informatika, Universitas Telkom, Kampus Surabaya, Jl. Ketintang No. 156, Surabaya 60231, Jawa Timur, Indonesia
ardianyw@telkomuniversity.ac.id

3rd Pima Hani Safitri, S.Kom., M.Kom.
Program Studi Informatika, Universitas Telkom, Kampus Surabaya, Jl. Ketintang No. 156, Surabaya 60231, Jawa Timur, Indonesia
phanisafitri@telkomuniversity.ac.id

Abstrak **—** Ekstraksi fitur yang kurang optimal merupakan salah satu kendala utama dalam klasifikasi ekspresi wajah menggunakan metode tradisional. Penelitian ini bertujuan untuk meningkatkan akurasi pengenalan ekspresi wajah dengan menerapkan pendekatan berbasis *deep learning* yang secara khusus menargetkan bagian-bagian penting wajah. Metode yang diusulkan menggabungkan arsitektur *Siamese Neural Network* (SNN) untuk mengukur kemiripan antar ekspresi, serta *multi-level feature extraction* (HFT-CNN) untuk melakukan ekstraksi fitur secara mendalam dan terfokus pada tiga area utama wajah, yaitu keseluruhan wajah, mata dan alis, serta mulut. Ketiga *channel* ini digabungkan untuk membentuk representasi fitur yang lebih kaya dan informatif. Hasil implementasi menunjukkan bahwa arsitektur HFT-CNN mampu mencapai akurasi hingga 99%, sedangkan model SNN *Triple* mencatatkan akurasi sebesar 91%. Meskipun demikian, hasil prediksi dari kedua model belum sepenuhnya stabil dalam berbagai kondisi pengujian, yang mengindikasikan masih adanya keterbatasan dalam hal generalisasi terhadap variasi ekspresi wajah. Selain itu, proses pengumpulan dan *preprocessing* data turut berpengaruh terhadap performa model, sehingga seleksi data secara manual tetap diperlukan guna memastikan kualitas dan relevansi data yang digunakan dalam pelatihan

Kata kunci— convolutional neural networks**, ekspresi wajah, pembelajaran mesin,** ***siamese networks*****.**

I. PENDAHULUAN

Ekspresi wajah memiliki peran yang sangat penting dalam interaksi manusia, baik secara verbal maupun non-verbal [1]. Banyak individu mengalami kesulitan dalam mengekspresikan emosi mereka secara jelas, yang sering kali menimbulkan kesalahpahaman atau konflik. Melalui ekspresi wajah, manusia dapat memperkuat atau mendukung pesan emosional yang ingin disampaikan, seperti kebahagiaan, kesedihan, kemarahan, ketakutan, atau rasa jijik. Ekspresi-emosi ini muncul melalui perubahan otot wajah dan tampilan visual wajah secara keseluruhan. Seiring perkembangan teknologi di era digital, pemanfaatan kecerdasan buatan (AI) untuk mengenali ekspresi wajah menjadi sangat relevan dan bermanfaat dalam membantu menganalisis emosi seseorang.

Namun, pengenalan ekspresi wajah secara otomatis oleh komputer menghadapi tantangan yang lebih kompleks dibandingkan pengamatan manusia secara langsung. Pemilihan fitur yang tepat sangat berperan dalam menentukan akurasi klasifikasi. Berdasarkan model komunikasi Mehrabian [2], diketahui bahwa ekspresi wajah menyumbang lebih dari setengah makna dalam komunikasi manusia, yaitu sebesar 55%, dibandingkan suara (38%) dan bahasa (7%). Dalam proses ekstraksi fitur, terdapat dua pendekatan utama: fitur geometris yang memanfaatkan titik-titik landmark wajah, serta fitur tampilan yang menyoroti tekstur otot wajah seperti kerutan dan lipatan. Penelitian terdahulu [3] menunjukkan bahwa mata dan mulut merupakan area wajah yang paling menentukan dalam mengenali emosi, sebagaimana ditunjukkan dalam eksperimen menggunakan teknologi *eye tracker.*

Dalam beberapa tahun terakhir, pendekatan berbasis *deep learning* telah banyak digunakan, salah satunya adalah arsitektur *Convolutional Neural Networks* (CNN). Pendekatan seperti Siamese Neural Network (SNN) memungkinkan pemrosesan dua *input* atau lebih secara paralel menggunakan lapisan *embedding* untuk mengukur kemiripan melalui perhitungan jarak Euclidean [4]. Di sisi lain, pendekatan *Hierarchical Feature with Three Channel CNN* (HFT-CNN) [5] memanfaatkan pembagian wajah ke dalam tiga area utama—seluruh wajah, mata dan alis, serta mulut—dan mengekstrak fitur dari masing-masing area secara independen pada *channel* berbeda. HFT-CNN memberikan hasil yang menjanjikan dalam meningkatkan akurasi pengenalan ekspresi wajah. Namun, metode ini cenderung menggunakan jumlah layer yang terbatas dan parameter yang relatif kecil, sehingga berisiko menghasilkan akurasi yang tidak sebaik arsitektur yang lebih kompleks. Sebaliknya, meskipun SNN unggul dalam mengukur kemiripan antara ekspresi wajah, arsitektur ini tidak dirancang khusus untuk klasifikasi langsung, sehingga tidak cukup andal bila digunakan sebagai acuan tunggal dalam pengukuran akurasi klasifikasi.

Penelitian ini mengusulkan metode klasifikasi ekspresi wajah dengan menggabungkan arsitektur HFT-CNN dan SNN. Penggabungan kedua metode ini diharapkan dapat mengoptimalkan keunggulan masing-masing, yaitu ketepatan dalam mengukur kemiripan melalui SNN dan kekuatan ekstraksi fitur secara mendetail melalui HFT-CNN. Tujuan dari penelitian ini adalah untuk mengimplementasikan gabungan kedua arsitektur tersebut

dalam mengenali dan mengklasifikasikan ekspresi wajah secara efektif, serta mengevaluasi tingkat ketepatan klasifikasinya.

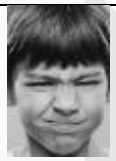
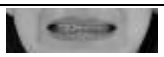
II. KAJIAN TEORI

Menyajikan dan menjelaskan teori-teori yang berkaitan dengan variabel-variabel penelitian. Poin subjudul ditulis dalam abjad.

A. Ekspresi Wajah

Ekspresi wajah sangat penting dalam interaksi sosial karena emosi yang ditampilkan seseorang dapat mempengaruhi perilaku orang yang melihat ekspresi wajah tersebut. Ekspresi emosi, baik dalam bentuk wajah, suara, bahasa tubuh, maupun simbol menyimpan petunjuk yang dapat diinterpretasikan oleh orang lain untuk memahami emosinya [6]. Dalam kasus pengenalan ekspresi wajah, biasanya arsitektur yang digunakan akan melihat bentuk keseluruhan wajah tanpa melihat bagian tertentu yang mungkin menyimpan fitur uniknya. Oleh karena itu diperlukan ekstraksi fitur dari bagian-bagian tersebut yang bisa dijadikan kunci untuk meningkatkan akurasi dalam pengenalan ekspresi wajah.

TABEL 1
(A) Sampel Ekspresi Dan Fitur

Ekspresi	Fitur	Citra
Angry	Rahang mengencang (kaku dan menegang).	
	Alis mata menyempit dan terdapat kerutan otot disekitar alis, Tatapan mata terbuka dan sorot mata tajam.	
	Mulut tertutup dan sedikit mengerucut.	
Disgust	Pipi terangkat hingga membentuk lekukan di area sekitar hidung dan mulut, Hidung seperti ditarik ke atas.	
	Kelopak mata bawah naik ke atas dan sejajar, Alis mengerucut.	
	Mulut tertutup dan ujung bibir sedikit tertarik / turun ke bawah.	
Fear	Wajah menegang, Otot wajah terlihat menurun.	
	Mata terbuka lebar,	
	Alis mata terangkat,	
	Mulut terbuka dan bibir tertarik kebelakang.	
Happy	Otot wajah tertarik / mengencang, Pipi terlihat terangkat,	
	Alis mata sedikit melengkung, Mata terbuka lebar, Muncul sedikit kerutan di bagian bawah kelopak mata.	
	Mulut terbuka dan gigi dapat terlihat, Ujung bibir tertarik kebelakang dan terangkat.	
Neutral	Keseluruhan wajah tidak menunjukkan penarikan otot apapun, Wajah simetris.	
	Mata dan alis memiliki posisi sama lurus,	
	Mulut tertutup rapat, Bibir atas dan bawah saling menyentuh.	
Sad	Kontur keseluruhan wajah terlihat turun.	
	Ujung alis bagian dalam sedikit terangkat kemudian mulai menurun hingga ujung luar, Kelopak mata atas sedikit terangkat.	
	Mulut sedikit turun.	
Surprise	Otot keseluruhan wajah tertarik kebelakang, Hampir tidak terlihat kerutan didaerah mata.	

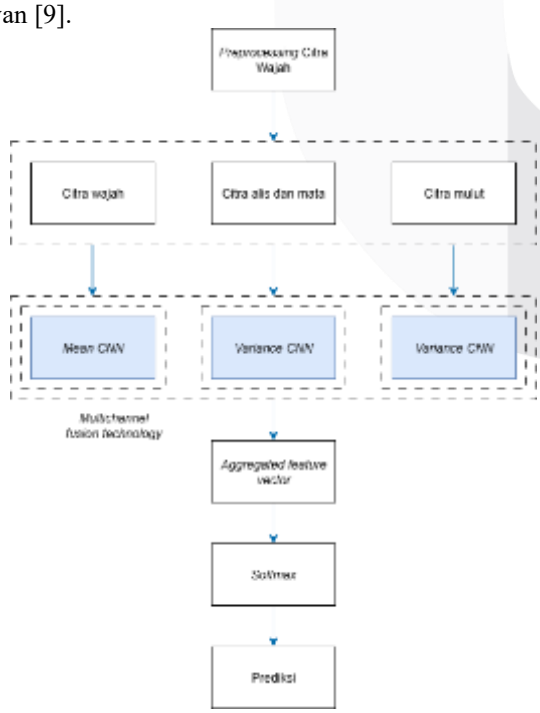
Ekspresi	Fitur	Citra
	Alis mata terangkat, Mata terbuka lebar.	
	Mulut terbuka lebar seperti mengucap huruf "O", Bibir dan gigi saling terpisah lebar.	

Pada TABEL 1 dipilih enam emosi dasar yang diekspresikan manusia yaitu *angry, disgust, fear, happy, sad, surprise*. Emosi ini secara luas dikenal sebagai emosi fundamental yang didefinisikan oleh Paul Ekman. Selain itu, disertakan emosi *neutral* yang merupakan emosi tanpa ekspresi dominan.

B. Convolutional Neural Networks

Convolutional Neural Networks (CNN) adalah teknologi dalam *deep learning* yang memiliki peran penting dalam pengenalan dan klasifikasi objek digital. CNN dirancang untuk meniru cara kerja otak manusia dalam memproses data visual, sehingga memungkinkan model untuk mengekstraksi fitur citra secara otomatis tanpa perlu intervensi manusia dalam proses *feature engineering* [7], [8].

CNN terdiri dari beberapa komponen utama, yaitu *convolutional layers* yang mengekstraksi fitur penting citra, *pooling layers* untuk mengurangi dimensi data, *activation functions* yang mensimulasikan respon *neuron*, dan *fully connected layers* yang berfungsi untuk membuat prediksi berdasarkan fitur yang telah di ekstraksi. Setiap *neuron* dalam CNN saling terhubung dan membuat jaringan ini efektif dalam mengurangi parameter dan mempercepat proses konvergensi. Dengan fitur *weight sharing*, CNN mampu mengurangi jumlah parameter yang diperlukan, dan melalui *downsampling* atau *pooling*, CNN dapat mempertahankan informasi penting sembari menghapus fitur yang kurang relevan [9].



GAMBAR 1
(A) Arsitektur HFT-CNN

Salah satu varian CNN yang lebih kompleks adalah *Hierarchical Feature with Three Channel Convolutional Neural Networks* (HFT-CNN) yang dirancang khusus untuk meningkatkan akurasi pengenalan ekspresi wajah melalui *multi-level feature extraction*.

Pada GAMBAR 1 memperlihatkan arsitektur HFT-CNN dengan cara membagi gambar wajah menjadi tiga *channel* berbeda yang masing-masing melatih area spesifik dari wajah seperti keseluruhan wajah, mata dan alis, dan mulut. Setiap *channel* menggunakan kernel CNN yang berbeda untuk menyesuaikan ekstraksi fitur dengan area wajah tertentu.

Channel untuk keseluruhan wajah menggunakan pendekatan *mean-controlled* CNN, yang dirancang untuk menangkap karakteristik global seperti kontur wajah. Rata-rata kernel dihitung dengan formula berikut:

$$\mathrm{m} = \frac{1}{\mathrm{h} \times \mathrm{w} \times \mathrm{n}} \sum_{\mathrm{i}=0}^{\mathrm{h}} \sum_{\mathrm{j}=0}^{\mathrm{w}} \sum_{\mathrm{k}=0}^{\mathrm{n}} \left(\mathrm{C}(\mathrm{i},\mathrm{j},\mathrm{k})\right) \quad (1)$$

Di mana h, w dan n masing-masing adalah tinggi, lebar, dan jumlah kernel; sementara $C(i, j, k)$ adalah nilai kernel pada posisi tertentu.

Sementara itu, *channel* untuk area mata dan alis serta mulut menggunakan pendekatan *variance-controlled* CNN, yang dioptimalkan untuk mendeteksi detail perubahan halus pada elemen seperti tepi, garis, atau tesktur di area tersebut. Varians kernel dihitung dengan formula berikut:

$$\mathrm{s} = \sqrt{\frac{1}{\mathrm{h} \times \mathrm{w} \times \mathrm{n}} \sum_{\mathrm{i}=0}^{\mathrm{h}} \sum_{\mathrm{j}=0}^{\mathrm{w}} \sum_{\mathrm{k}=0}^{\mathrm{n}} (\mathrm{C}(\mathrm{i},\mathrm{j},\mathrm{k}) - \mathrm{m})^2} \quad (2)$$

Ketiga *channel* ini dapat dilatih dengan gambar yang berasal dari individu berbeda, selama gambar tersebut berasal dari satu kelas ekspresi emosi yang sama.

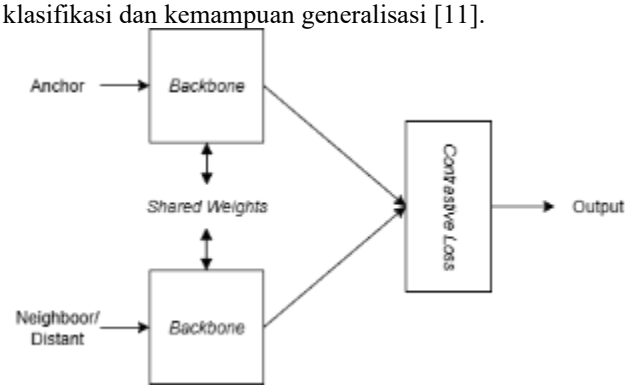
Setelah fitur dari ketiga *channel* diekstraksi, teknik *multi-channel fusion* digunakan untuk menggabungkan fitur tersebut untuk kemudian diproses melalui *layer Softmax* untuk klasifikasi ekspresi [5]. Pendekatan ini memungkinkan HFT-CNN mengekspresikan fitur wajah secara menyeluruh dan meningkatkan tingkat pengenalan ekspresi dibandingkan dengan CNN tradisional yang hanya berfokus pada fitur global

C. Siamese Neural Networks

Siamese Neural Networks (SNN) merupakan arsitektur *neural network* (NN) yang dirancang untuk belajar dengan cara membandingkan kesamaan antara dua input atau lebih. Jaringan ini terdiri dari dua jaringan kembar yang berbagi bobot dan parameter, menerima dua input berbeda, dan menghasilkan representasi dari input tersebut. Kemudian, jaringan tersebut akan menghitung metrik jarak probabilitas antara kedua representasi yang menunjukkan kesamaan atau perbedaan antara dua *input* [10].

Dalam proses pelatihannya, SNN menggunakan dua *input* dengan menerapkan fungsi *contrastive loss*. Dalam pendekatan ini, pasangan data terdiri dari *anchor* dan *neighboor* untuk pasangan dengan kelas yang sama (positif), atau *anchor* dan *distant* yaitu pasangan dari kelas yang berbeda (negatif). Fungsi *loss* ini dirancang untuk menarik pasangan dari kelas sama agar memiliki representasi yang lebih dekat di *feature space*, serta mendorong pasangan dari kelas yang berbeda agar saling menjauh dengan jarak minimal yang ditentukan. Dengan cara ini, SNN belajar menghasilkan representasi fitur yang mampu membedakan

antar kelas secara efektif, sehingga meningkatkan performa klasifikasi dan kemampuan generalisasi [11].



GAMBAR 2
(B) Arsitektur SNN Dua Input

Konsep SNN bersifat *topology-agnostic*, yang berarti dapat menggunakan berbagai jenis arsitektur NN sebagai *backbone*-nya. Struktur jaringan SNN dapat dilihat pada GAMBAR 2.

SNN umumnya menggunakan dua jenis fungsi *loss* yaitu *triplet loss* dan *contrastive loss*. *Contrastive loss* didefinisikan sebagai:

$$\ell_c = \sum_{i=1}^{b} \left[(1-y) \parallel f(x_1^i) - f(x_2^i) \parallel_2^2 + y\left[- \parallel f(x_1^i) - f(x_2^i) \parallel_2^2 + a \right]_+ \right] \quad (3)$$

Keterangan:

ℓ_c	=	*loss function*
b	=	ukuran *mini-batch* (jumlah pasangan dalam satu *batch*)
i	=	indeks setiap *mini-batch*
$f(x)$	=	fungsi yang memetakan input (x) ke *latent space*
x_1^i, x_2^i	=	dua input (pasangan) dalam *batch* ke-i
a	=	*margin* (batas minimum pemisahan untuk *negative pairs*)
y^i	=	label pasangan ke-i $y^i = 0$ (kelas sama / *positive pairs*) $y^i = 1$ (kelas berbeda / *negative pairs*)
$\|\cdot\|_2^2$	=	*norm Euclidean Distance*
$[\cdot]_+$	=	operator maksimum $[z]_+ = \max(z, 0)$ menghasilkan 0 jika $z < 0$, dan menghasilkan z jika $z \geq 0$. Digunakan untuk memastikan kerugian hanya dihitung saat kondisi belum terpenuhi (misalnya jarak antar *embedding* belum cukup besar).

Di mana y bernilai nol jika pasangan $\{x_1^i, x_2^i\}$ berasal dari kelas yang sama, dan bernilai satu jika pasangan berasal dari kelas berbeda.

D. Akurasi

Dalam analisis klasifikasi terdapat metode yang dapat digunakan untuk mengetahui hasil prediksi yang telah dilakukan yaitu dengan menggunakan *confusion matrix* [12]. Metode ini bekerja dengan cara membandingkan hasil prediksi terhadap label sebenarnya. Komponen utama dari *confusion matrix* adalah:

- **True Positive (TP)**: sampel yang terdeteksi sebagai positif dan memang memiliki label positif.
- **True Negative (TN)**: sampel yang terdeteksi sebagai negative dan meman memiliki label negatif
- **False Positive (FP)**: sampel yang terdeteksi sebagai positif namun sebenarnya adalah negatif.
- **False Negative (FN)**: sampel yang terdeteksi sebagai negatif namun sebenarnya adalah positif.

TABEL 2
(B) Confusion Matrix Untuk Klasifikasi Biner

Class	As Positive	As Negative
Positive	TP	FN
Negative	FP	TN

F1 score merupakan metrik yang digunakan untuk mengevaluasi kinerja dari klasifikasi biner [13]. *F1 score* dihitung sebagai nilai rata-rata antara *precision* dan *recall*. *F1 score* didefinisikan sebagai:

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

Dimana:

- ***Precision*** mengukur proporsi dari prediksi positif.

$$\mathrm{F1} = 2 \times \frac{\mathrm{Precision} \times \mathrm{Recall}}{\mathrm{Precision} + \mathrm{Recall}} \quad (5)$$

- ***Recall*** mengukur proporsi sampel positif yang berhasil dideteksi.

$$REC = \frac{TP}{TP + FN} \quad (6)$$

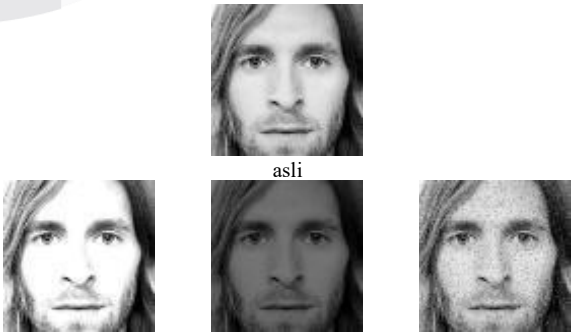
III. METODE

A. Pengumpulan Data

Pada tahap ini, data yang digunakan berupa citra wajah diperoleh melalui proses pengunduhan menggunakan *repository open-source* "*image-downloader*" yang dikembangkan oleh Qianyan-Tech. Dalam prosesnya, kueri yang digunakan adalah "***[nama label] facial expression***" digunakan untuk memastikan hasil pencarian yang sesuai dengan kebutuhan. Selain itu pencarian data secara manual juga dilakukan karena keterbatasan dari *tools* yang digunakan tidak selalu mendapatkan citra berupa wajah manusia. Untuk memastikan keseimbangan jumlah *dataset*, maka pada tahap pengumpulan data setidaknya setiap kelas paling sedikit harus memiliki 40 data gambar yang berbeda.

B. Preprocessing Data

Gambar yang telah dikumpulkan akan melalui tahap *preprocessing* menggunakan algoritma *haar-cascade classifier* untuk mendeteksi bagian wajah tertentu seperti keseluruhan wajah, area mata dan alis, serta mulut. Setelah bagian-bagian tersebut berhasil diidentifikasi dan dilakukan *crop* hasilnya akan diubah ukurannya menjadi 38x38 piksel. Untuk memastikan *dataset* memiliki variasi yang memadai, dilakukan *augmentasi* data berupa penyesuaian tingkat kecerahan (*brightness*) sebanyak 30% lebih gelap dan 50% lebih terang, serta penambahan *noise* (*salt & pepper*). Selanjutnya, semua data yang dihasilkan akan dikonversi ke format *grayscale*. Hasil augmentasi data dapat dilihat pada GAMBAR 3.

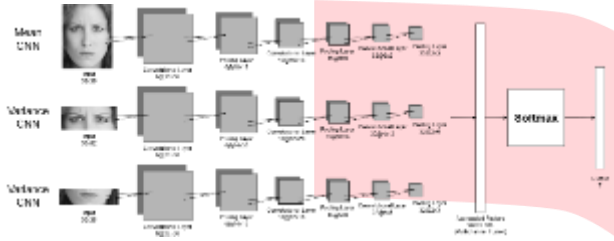


GAMBAR 3
(C) Hasil Augmentasi Citra

Hasil akhir dari keseluruhan gambar yang didapatkan berjumlah 480 data yang dibagi ke 7 kelas dengan masing-masing 160 data per bagian wajah. *Dataset* yang sudah dihasilkan akan dibagi menjadi *training sets* dan *test sets* dengan rasio 8:2.

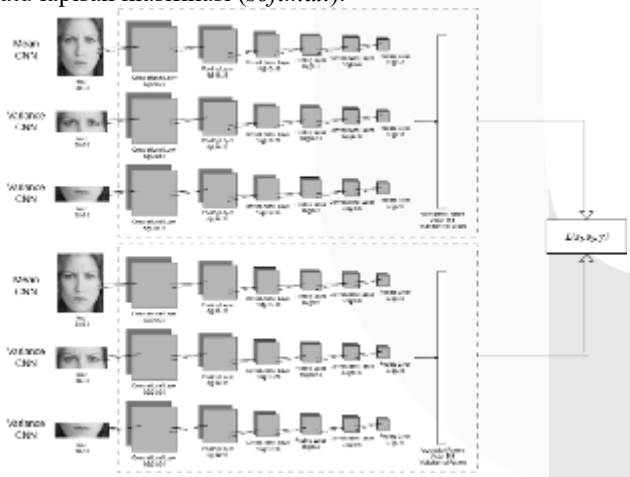
C. Siamese Network dan HFT-CNN

Penelitian ini menggabungkan dua metode utama, yaitu *Hierarchical Feature with Three Channel Convolutional Neural Networks* (HFT-CNN) dan *Siamese Neural Networks* (SNN), dengan variasi *Double* dan *Triplet Network* pada arsitektur SNN. *Siamese double network* dirancang untuk membandingkan dua input sekaligus, yaitu *anchor*, dan *comparison image* (gambar pembanding). Setiap *input* tersebut akan diproses melalui arsitektur HFT-CNN sebagai *backbone*.



GAMBAR 4
(D) Model HFT-CNN

GAMBAR 4 menunjukkan *output layer* dari model yang dibuat. Angka sebelum simbol "@" merupakan jumlah *channel* dalam *layer* CNN dan angka setelahnya merupakan ukuran dari *feature map* (*neuron*). Arsitektur dibagi menjadi 3 jaringan konvolusi yaitu satu lapisan *input*, dua lapisan CNN, dua lapisan *pooling*, satu lapisan *feature fusion*, dan satu lapisan klasifikasi (*softmax*).

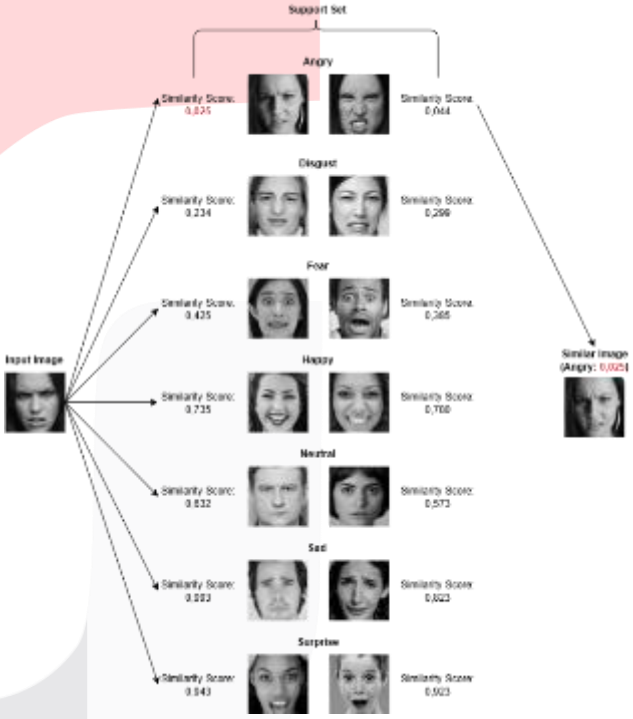


GAMBAR 5
(E) Model Siamese Dengan Backbone HFT-CNN

GAMBAR 5 menunjukkan arsitektur lengkap *siamese* dengan HFT-CNN sebagai *backbone*. Setiap cabang menerima satu gambar bagian wajah yang diproses melalui tiga jalur CNN: wajah utuh (*mean-controlled*), mata–alis, dan mulut (keduanya *variance-controlled*). Masing-masing jalur terdiri dari lapisan *convolutional* dan *pooling* yang mengekstraksi fitur bagian wajah. Hasil dari ketiga jalur digabung menjadi satu vektor melalui proses *multi-channel fusion*. Dua vektor dari pasangan gambar kemudian dibandingkan menggunakan fungsi jarak, dan nilai perbedaan diproses oleh fungsi *loss* $L = (S_1, S_2, y)$ untuk menentukan apakah keduanya berasal dari kelas ekspresi yang sama atau berbeda.

Dalam skema pengujian SNN, model dievaluasi dengan cara memberikan sebuah *input image* yang kemudian dicocokkan dengan kumpulan gambar pembanding (*support-set*). Baik *input image* maupun gambar-gambar dalam *support-set* sepenuhnya berasal dari data yang berbeda dengan *dataset* yang digunakan selama proses pelatihan. Model akan menghitung tingkat kemiripan atau jarak antara *input image* dan setiap gambar dalam *support-set* untuk menentukan pasangan yang paling sesuai. Gambar dalam *support-set* yang memiliki jarak terkecil terhadap *input image* dianggap sebagai prediksi model terhadap kelas yang diwakili oleh gambar tersebut.

GAMBAR 6 menggambarkan alur pengujian SNN dalam mencocokkan *input image* dengan *support-set. Cara* ini bertujuan untuk menguji kemampuan model dalam membedakan representasi visual pada data yang belum pernah dilihat sebelumnya, serta menilai keakuratan prediksi berdasarkan kedekatan representasi fitur.



GAMBAR 6
(F) Skema Uji SNN

IV. HASIL DAN PEMBAHASAN

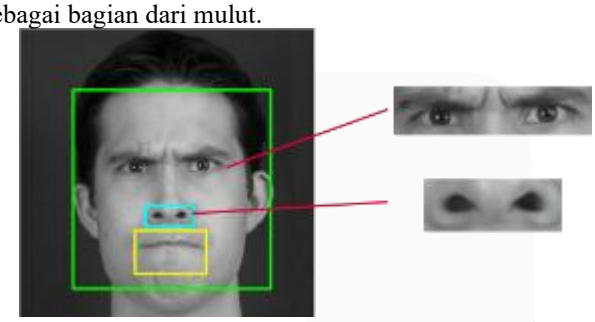
A. Akuisisi Citra Wajah

Dalam proses ekstraksi fitur untuk keperluan pelatihan model, metode *Haar Cascade Classifier* digunakan untuk mendeteksi dan mengekstraksi area pada bagian wajah. Dengan menggunakan model *pre-train* dari OpenCV, bagian keseluruhan wajah di deteksi terlebih dahulu kemudian dilakukan deteksi lebih lanjut pada bagian yang lain seperti mata dan alis, serta mulut. Setiap bagian yang berhasil terdeteksi kemudian disimpan sebagai *input dataset* yang akan digunakan dalam proses pelatihan klasifikasi ekspresi wajah. Sampel dari gambar yang telah terdeteksi dapat dilihat pada TABEL 3.

TABEL 3
(C) Sampel Ekstraksi Wajah

Citra	Ekstraksi Fitur
	

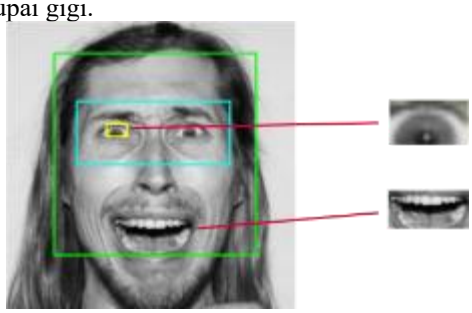
Dalam implementasi algoritma *Viola-Jones* untuk mendeteksi fitur-fitur wajah, seperti wajah secara keseluruhan, mata dan alis, serta mulut, ditemukan beberapa permasalahan yang berkaitan dengan ketepatan deteksi. Salah satu permasalahan yang sering muncul adalah ketidaktepatan dalam mendeteksi area wajah secara utuh, khususnya ketika subjek membuka mulutnya secara lebar. Pada kondisi tersebut, wajah tetap terdeteksi, namun bagian mulut hanya terambil sebagian sehingga area bibir bawah terpotong. Selain itu, pada proses deteksi mata, sering terjadi kesalahan di mana area hidung atau lubang hidung teridentifikasi sebagai mata. Kesalahan serupa juga terjadi pada deteksi mulut, di mana salah satu mata terkadang keliru dikenali sebagai bagian dari mulut.



GAMBAR 7
(G) Kesalahan Deteksi Mata

Sebagai contoh pada GAMBAR 7, *bounding box* biru seharusnya mendeteksi mata namun malah mengenali hidung. Kesalahan ini kemungkinan terjadi karena algoritma menganggap bentuk dan gradasi abu-abu pada hidung menyerupai kelopak mata, sementara lubang hidung yang gelap tampak seperti iris, terutama pada citra *grayscale*.

Selanjutnya, pada GAMBAR 8 *bounding box* berwarna kuning seharusnya mendeteksi mulut bukan mata. Kemiripan bentuk dan pola antara kedua bagian bisa saja menjadi penyebab utama. Bibir dapat disalahartikan sebagai kelopak mata, dan sklera (bagian putih pada mata) dapat dianggap menyerupai gigi.



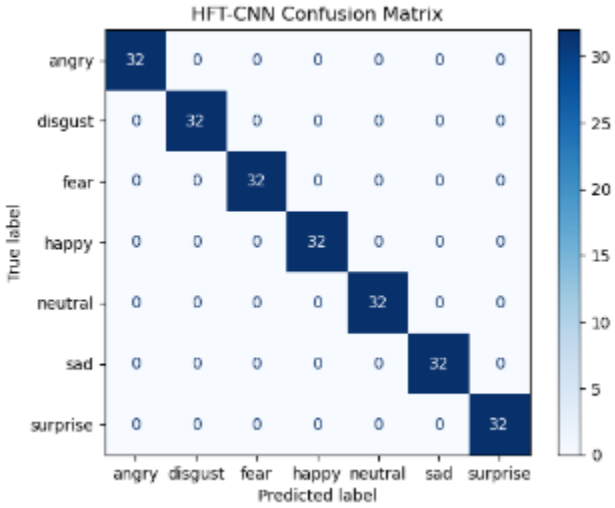
GAMBAR 8
(H) Kesalahan Deteksi Mata

Kesalahan deteksi tersebut merupakan akibat dari keterbatasan algoritma Viola-Jones yang menggunakan pendekatan *sliding window* untuk mencocokkan *Haar-like feature* pada setiap bagian, di mana pola-pola sederhana seperti garis, tepi, atau perbedaan blok gelap dan terang yang dibandingkan oleh *Haar-feature* dapat muncul di berbagai bagian wajah dan menyerupai pola dari bagian lain, sehingga mengakibatkan deteksi yang tidak akurat.

Meskipun tingkat kesalahan ini cukup mengganggu, hasil deteksi tetap digunakan sebagai *input* untuk model yang dibuat, dengan mempertimbangkan adanya toleransi terhadap ketidaksempurnaan pada hasil deteksi.

B. HFT-CNN

Selama proses pelatihan, model *Hierarchical Feature with Three Channel Convolutional Neural Networks* (HFT-CNN) menunjukkan peningkatan akurasi secara bertahap seiring bertambahnya jumlah epoch, disertai dengan penurunan nilai *loss* yang konsisten. Setelah menyelesaikan seluruh 100 *epoch* pelatihan, model berhasil mencapai akurasi pelatihan (*train accuracy*) sebesar 99% dan akurasi validasi (*validation accuracy*) juga sebesar 99%.



GAMBAR 9
(I) Confusion Matrix HFT-CNN

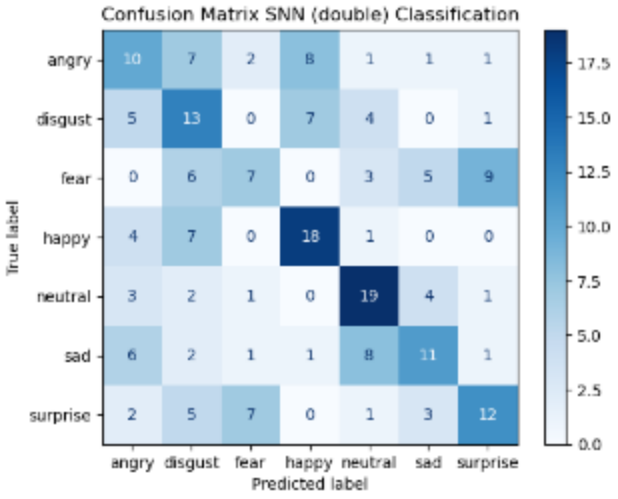
Hasil klasifikasi untuk tugas pengenalan emosi menunjukkan bahwa model HFT-CNN mampu membedakan setiap kategori emosi dengan cukup baik. Berdasarkan *confusion matrix* pada GAMBAR 9 mayoritas prediksi berada pada diagonal *confusion matrix*, yang mengindikasikan bahwa sebagian besar emosi berhasil diklasifikasikan secara akurat oleh model.

C. Siamese Neural Networks (Double)

Pada model *Siamese Neural Network* (SNN) yang diterapkan, arsitektur HFT-CNN digunakan sebagai *backbone* dan dilatih selama 100 epoch menggunakan pendekatan *contrastive loss* untuk mendeteksi kemiripan ekspresi wajah. Selama pelatihan, akurasi data pelatihan meningkat secara signifikan dan mencapai hampir 100% setelah sekitar 30 *epoch*, kemudian tetap stabil hingga akhir pelatihan. Sementara itu, akurasi validasi berada pada kisaran 92–95% dan tidak mengalami peningkatan seiring bertambahnya *epoch*.

Dari sisi nilai *loss*, *training loss* menurun tajam hingga mendekati nol, sementara *validation loss* cenderung stabil setelah beberapa *epoch*. Evaluasi akhir pada data uji menunjukkan bahwa model menghasilkan nilai *loss* sebesar

0.041 dan akurasi sebesar 0.955. Meskipun terdapat indikasi *overfitting*, model tetap mampu memberikan performa yang cukup baik pada data uji, dengan akurasi mencapai sekitar 95%. Hasil klasifikasi menunjukkan bahwa model berhasil mengidentifikasi sebagian besar pasangan gambar secara tepat.

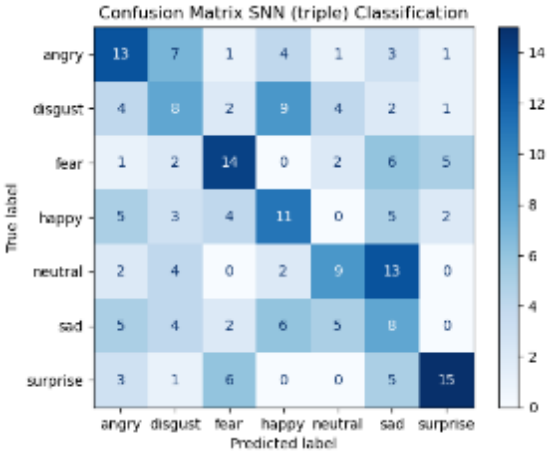


GAMBAR 10
(J) Confusion Matrix SNN Double

Hasil *confusion matrix* (GAMBAR 10) menunjukkan bahwa model SNN Double memiliki performa yang cukup baik dalam mengenali emosi "*neutral*", dengan 19 prediksi benar dari 30 data uji. Namun, masih terdapat kesalahan klasifikasi yang cukup signifikan pada beberapa kelas, terutama antara "*surprise*" dan "*fear*". Kesalahan ini kemungkinan disebabkan oleh kemiripan bentuk ekspresi wajah pada kedua emosi tersebut, yang menyulitkan model dalam membedakan fitur secara akurat.

D. Siamese Neural Networks (Triple)

Model Siamese Neural Network (SNN) dengan penerapan *Triplet Loss* dilatih selama 100 *epoch* dan menunjukkan performa yang cukup baik pada pelatihan, namun mengalami kendala saat diuji. Akurasi pelatihan meningkat dengan cepat dan mencapai nilai maksimum sekitar *epoch* ke-30, kemudian stabil hingga akhir pelatihan. Nilai *loss* sempat naik turun pada awal pelatihan, namun akhirnya menurun dan stabil di angka 17%. Meskipun nilai akurasi pelatihan mencapai 91% dan *loss* terbilang kecil, terdapat indikasi *overfitting* yang cukup kuat karena perbedaan signifikan antara performa pelatihan dan validasi.



GAMBAR 11
(K) Confusion Matrix SNN Triple

Dalam hasil pengujian (GAMBAR 11) terhadap klasifikasi ekspresi emosi, model mengevaluasi tujuh kelas: *angry*, *disgust*, *fear*, *happy*, *neutral*, *sad*, dan *surprise*. Model menunjukkan performa yang cukup baik pada beberapa kelas, seperti *fear* (14 prediksi benar), *surprise* (15), dan *angry* (13), namun masih mengalami kesalahan pada kelas lain seperti *disgust*, yang hanya memiliki 8 prediksi benar dan sisanya tersebar ke kelas lain seperti *happy* dan *angry*. Prediksi pada kelas *neutral* dan *sad* juga menunjukkan pencampuran antar kelas, dengan masing-masing 13 dan 8 prediksi benar.

E. Hasil

Dalam penelitian ini, dilakukan perbandingan kinerja antara tiga metode berbeda untuk mengukur ketepatan dalam mendeteksi emosi manusia, yaitu *Hierarchical Feature with Three Channel Convolutional Neural Networks* (HFT-CNN) serta *Siamese Neural Networks* (SNN) dengan dua variasi arsitektur, yaitu *double* dan *triple networks*. Ketiga model diuji menggunakan *dataset* yang sama dan skenario pengujian yang seragam untuk memastikan hasil evaluasi yang sama. Berdasarkan hasil pengujian yang dilakukan, masing-masing model menunjukkan nilai akurasi yang berbeda. Nilai-nilai akurasi tersebut dapat dilihat pada TABEL 4

TABEL 4
(D) Perbandingan Akurasi Model

Akurasi Model		
HFT-CNN	SNN (double)	SNN (triple)
0,996 (99%)	0,955 (95%)	0,919 (91%)

V. KESIMPULAN

Penelitian ini berhasil mengimplementasikan arsitektur *Hierarchical Features with Three-Channel Convolutional Neural Network* (HFT-CNN) dan *Siamese Neural Network* (SNN) untuk tugas pengenalan dan klasifikasi ekspresi wajah. Proses pengumpulan dan pengolahan data dilakukan dengan cermat, termasuk seleksi manual untuk memastikan kualitas data yang digunakan. Hasil pelatihan menunjukkan bahwa model HFT-CNN mampu mencapai akurasi tinggi hingga 99%, sedangkan SNN *Triple* mencatatkan akurasi sebesar 91%. Meskipun demikian, performa kedua model belum sepenuhnya stabil di semua kondisi pengujian, yang mengindikasikan adanya keterbatasan dalam kemampuan generalisasi terhadap variasi ekspresi wajah. Temuan ini menunjukkan bahwa meskipun pendekatan yang digunakan efektif, pengembangan lebih lanjut masih diperlukan untuk meningkatkan akurasi dan konsistensi model dalam pengenalan ekspresi wajah yang lebih kompleks dan beragam.

REFERENSI

[1] C. Dewi, L. S. Gunawan, S. G. Hastoko, and H. J. Christanto, "Real-Time Facial Expression Recognition: Advances, Challenges, and Future Directions," May 01, 2024, *World Scientific*. doi: 10.1142/S219688882330003X.

[2] M. Ramzani Shahrestani, S. Motamed, and M. Yamaghani, "Recognition of facial emotion based on SOAR model," *Front Neurosci*, vol. 18, 2024, doi: 10.3389/fnins.2024.1374112.

[3] A. Esté Jaloveckas and R. Granero, “The eyes as the exclamation mark of the face: exploring the relationship between eye size, intensity of female facial expressions and attractiveness in a range of emotions,” *Front Psychol*, vol. 15, Jul. 2024, doi: 10.3389/fpsyg.2024.1421707.

[4] F. Xu, J. Gao, and X. Pan, “Cow Face Recognition for a Small Sample Based on Siamese DB Capsule Network,” *IEEE Access*, vol. 10, pp. 63189–63198, 2022, doi: 10.1109/ACCESS.2022.3182806.

[5] Y. He, Y. Zhang, S. Chen, and Y. Hu, “Facial Expression Recognition Using Hierarchical Features with Three-Channel Convolutional Neural Network,” *IEEE Access*, vol. 11, pp. 84785–84794, 2023, doi: 10.1109/ACCESS.2023.3303402.

[6] G. A. Van Kleef and S. Côté, “The Social Effects of Emotions,” *Annu Rev Psychol*, vol. 73, p. 30, Jan. 2022, doi: 10.1146/annurev-psych-020821.

[7] A. Ajit, K. Acharya, and A. Samanta, “A Review of Convolutional Neural Networks,” in *International Conference on Emerging Trends in Information Technology and Engineering, ic-ETITE 2020*, Institute of Electrical and Electronics Engineers Inc., Feb. 2020. doi: 10.1109/ic-ETITE47903.2020.049.

[8] M. M. Taye, “Theoretical Understanding of Convolutional Neural Network: Concepts, Architectures, Applications, Future Directions,” Mar. 01, 2023, *MDPI*. doi: 10.3390/computation11030052.

[9] Z. Li, F. Liu, W. Yang, S. Peng, and J. Zhou, “A Survey of Convolutional Neural Networks: Analysis, Applications, and Prospects,” *IEEE Trans Neural Netw Learn Syst*, vol. 33, no. 12, pp. 6999–7019, Dec. 2022, doi: 10.1109/TNNLS.2021.3084827.

[10] A. Mahajan, J. D. Dormer, Q. Li, D. Chen, Z. Zhang, and B. Fei, “Siamese neural networks for the classification of high-dimensional radiomic features,” SPIE-Intl Soc Optical Eng, Mar. 2020, p. 131. doi: 10.1117/12.2549389.

[11] B. Ghojogh, M. Sikaroudi, S. Shafiei, H. R. Tizhoosh, F. Karray, and M. Crowley, *Fisher Discriminant Triplet and Contrastive Losses for Training Siamese Networks*. Glasgow: IEEE, 2020. doi: 10.1109/IJCNN48605.2020.9206833.

[12] M. Sokolova, N. Japkowicz, and S. Szpakowicz, “Beyond accuracy, F-score and ROC: A family of discriminant measures for performance evaluation,” in *AAAI Workshop - Technical Report*, 2006, pp. 24–29. doi: 10.1007/11941439_114.

[13] M. Sitarz, “Extending F1 metric, probabilistic approach,” Oct. 2022, doi: 10.54364/AAIML.2023.1161.