Jurnal Teknologi Informasi dan Ilmu Komputer (JTIIK) DOI: 10.25126/jtiik.202295269
Vol. 9, No. 5, Oktober 2022, hlm. 1047-1054 p-ISSN: 2355-7699
Akreditasi KEMENRISTEKDIKTI, No. 36/E/KPT/2019 e-ISSN: 2528-6579

IDENTIFIKASI EMOSI WAJAH PENGGUNA KONFERENSI VIDEO MENGGUNAKAN CONVOLUTIONAL NEURAL NETWORK DENGAN ARSITEKTUR VGG-16

Lina[*1], Arthur Adhitya Marunduh[2], Wasino[3], Daniel Ajienegoro[4]

[1,2,3,4]Universitas Tarumanagara, Jakarta Barat
Email: [1]lina@untar.ac.id, [2]marunduh17@gmail.com, [3]wasino@fti.untar.ac.id, [4]daniel.a71e@gmail.com
*Penulis Korespondensi

(Naskah masuk: 09 Juli 2021, diterima untuk diterbitkan: 24 Oktober 2022)

Abstrak

Pandemi COVID-19 telah menyebabkan pergeseran bentuk komunikasi antar individu maupun antar kelompok. Komunikasi lisan antar individu dilakukan dengan bantuan teknologi secara daring. Berbagai teknologi konferensi video secara daring banyak dikembangkan, seperti Zoom, MS Teams, Google Meet, dan lain sebagainya. Dengan penggunaan teknologi tersebut, pengguna dapat saling berinteraksi secara visual dan verbal. Umumnya peserta konferensi video dengan mudah menangkap sinyal verbal dari lawan bicara. Namun untuk menginterpretasikan sinyal visual, peserta memerlukan proses analisis yang lebih kompleks. Makalah ini membahas tentang sistem identifikasi emosi secara otomatis dari peserta konferensi video. Sistem yang dikembangkan dapat diterapkan pada semua jenis teknologi konferensi video, baik versi rekaman maupun versi langsung. Aplikasi yang dikembangkan dapat digunakan untuk keperluan umum, seperti dalam rapat kelompok, kuliah, webinar, dan jenis rapat lainnya. Dalam eksperimen yang dilakukan, terdapat empat kelas emosi manusia yang digunakan, yaitu netral, senang, sedih, dan marah. Sistem melakukan deteksi area wajah dalam citra dari video masukan dengan algoritma Viola-Jones, dan melakukan identifikasi emosi pada wajah yang terdeteksi menggunakan metode Convolutional Neural Network (CNN) dengan arsitektur VGG-16. Hasil eksperimen menunjukkan bahwa sistem mampu secara otomatis melakukan pendeteksian area wajah dengan tingkat akurasi sebesar 93.15% dan melakukan identifikasi emosi dengan akurasi sebesar 88.39% untuk data latih dan 70.17% untuk data pengujian.

Kata kunci: *Identifikasi emosi, konferensi video, Convolutional Neural Network, Viola-Jones*

EMOTION IDENTIFICATION OF VIDEO CONFERENCE USERS USING CONVOLUTIONAL NEURAL NETWORK

Abstract

The COVID-19 pandemic has caused a shift in the form of communication between individuals and between groups. Oral communication between individuals is carried out with the help of online technology. Various online video conferencing technologies have been developed, such as Zoom, MS Teams, Google Meet, etc. With the use of this technology, users can interact with each other visually and verbally. Generally, video conference participants can easily receive verbal signals from the speaker. However, a more complex analysis is required to interpret other signals such as visual signals from face, gesture, etc. This paper discusses an automatic emotion identification system of video conferencing participants. The developed system can be applied to all types of video conference technology, both recorded or live versions, The developed application can be used for general purpose, such as in group meetings, lectures, webinars, and other types of meetings. In the experiments, four classes of human emotion are examined, such as neutral, happy, angry, and sad. The system detects facial areas in video frames using the Viola-Jones algorithm and identifies the emotions on the detected faces using the Convolutional Neural Network (CNN) method with the VGG-16 architecture. The experimental results show that the proposed system is able to automatically detect facial areas with an accuracy of 93.15% and identify emotions from the detected faces with an accuracy of 88.39% for training data and 70.17% for test data.

Keywords: *Emotion detection, video conference, Convolutional Neural Network, Viola-Jones*

1. PENDAHULUAN

Perkembangan teknologi telah berkembang dengan sangat pesat, namun pandemi COVID-19 telah menyebabkan pergeseran bentuk komunikasi antar individu maupun antar kelompok secara lebih signifikan. Dengan terbatasnya aktivitas yang dapat dilakukan seseorang pada masa pandemi, maka

komunikasi umumnya dilakukan dengan bantuan teknologi secara *online*. Berbagai teknologi kolaborasi dan konferensi video secara *online* banyak bermunculan dengan melibatkan unsur audio dan video, seperti Zoom, Microsoft Teams, Google Meet, dan lain sebagainya. Dengan penggunaan teknologi video konferensi tersebut, pengguna dapat saling berinteraksi melampaui dimensi lokasi.

Manusia umumnya mempunyai dua cara dalam mengekspresikan emosinya yaitu secara verbal dan nonverbal. Kemampuan verbal adalah mengungkapkan segala sesuatu yang dirasakan secara sadar dengan kata-kata, sedangkan kemampuan nonverbal adalah mengekspresikan apa yang dirasakan tubuh menggunakan media yang ada pada tubuh, seperti gerakan tangan, mengangkat bahu, ekspresi wajah, dan lainnya. Banyak hal yang dapat dilihat dari ekspresi wajah seseorang, termasuk niat, hubungan sosial, dan kepribadiannya dari ekspresi wajah yang ditunjukkan (Batoteng, Pasiak, & Ticoalu, 2015). Manusia dapat dengan sengaja memiliki ekspresi wajah tertentu, namun biasanya ekspresi wajah tersebut terjadi di luar kontrol diakibatkan luapan emosi manusia secara alamiah. Emosi manusia terutama tampil dalam ekspresi wajah yang merupakan hasil dari gerakan otot wajah yang mencerminkan perasaannya. Misalnya, tersenyum berarti kehangatan, mengangkat alis mencerminkan keterkejutan, dan sebagainya.

Walaupun secara umum ekspresi wajah dapat diartikan secara langsung pada saat kejadian, namun sesungguhnya wajah adalah sebuah sistem multi sinyal yang dapat mengandung banyak pesan. Hal ini mengakibatkan penerima untuk salah mengartikan ekspresi wajah seseorang karena banyaknya makna dibalik satu ekspresi wajah. Tidak mudah untuk mengetahui ekspresi wajah yang merupakan fenomena visual, sehingga diperlukan alat bantu untuk mendefinisikan ekspresi wajah manusia. Sistem pengenalan emosi secara otomatis dapat dibuat berdasarkan asumsi bahwa emosi memiliki pola tertentu yang tercermin dari ekspresi wajah (Liebold, Richter, Teichmann, Hamker & Ohler, 2015).

Berbagai metode untuk pengenalan wajah maupun pengenalan emosi berbasiskan citra wajah telah banyak dilakukan oleh para peneliti. Beberapa diantaranya adalah menggunakan metode statistik (Lina, Takahashi, Ide, & Murase, 2009), Convolutional Neural Network (Wang, Wu, Zhang, Xu, Zhang, Wu, & Coleman, 2020), metode Convolutional Neural Network dengan pengolahan ciri fisiologis (Lee, Lee, Lim, & Kang, 2020), dan metode You Look Only Once (Yu & Zhang, 2021). Dari berbagai metodologi yang telah dikembangkan sebelumnya, metode Convolutional Neural Network (CNN) merupakan salah satu algoritma yang diklaim paling akurat untuk mengolah data citra digital untuk berbagai aplikasi pengenalan. Pada penelitian yang dikembangkan oleh Oh (Oh, Ryu, Jeong, Yang, Hwang, Lee, & Lim, 2021) digunakan jumlah hidden layers yang sangat banyak pada algoritma CNN sehingga mampu melakukan pengenalan terhadap multi obyek secara baik. Sedangkan penelitian dari Andika (Andika, Pratiwi, & Handajani, 2019) menggunakan algoritma CNN serupa untuk pengenalan terhadap jenis penyakit pneumonia. Selain itu, kombinasi antara metode CNN dengan berbagai metode lainnya dalam konsep deep learning juga telah banyak dikembangkan dalam pengenalan emosi manusia seperti pada penelitian Giannopoulos (Giannopoulos, Perikos, & Hatzilygeroudis, 2018), kombinasi teknik Hierarchical Committee dan CNN pada penelitian Kim (Kim, Roh, Dong, & Lee, 2016), teknik kombinasi ekspresi pada penelitian Du (Du & Martinez, 2015), serta penggunaan fitur audio dan visual untuk pengenalan emosi pada penelitian Ouyang (Ouyang, Kawaii, Goh, Shen, Ding, Ming, & Huang, 2017). Penelitian terkait lainnya untuk pengenalan emosi wajah dengan algoritma CNN adalah penelitian dari Kusuma (Kusuma, Jonathan, & Lim, 2020) untuk basis data FER-13. Selain itu, penelitian lainnya menerapkan metode Deep Transfer Learning dan Multiple Temporal Models untuk pengenalan wajah yang diterapkan pada basis data Wild 2017 (Ouyang, Kawaii, Goh, Shen, Ding, Ming, & Huang, 2017). Perbedaan utama dari penelitian ini dibandingkan dengan penelitian sejenis lainnya terletak pada penerapan arsitektur VGG-16 untuk pengenalan emosi wajah yang diterapkan pada basis data FER-13 dan juga basis data orang Indonesia.

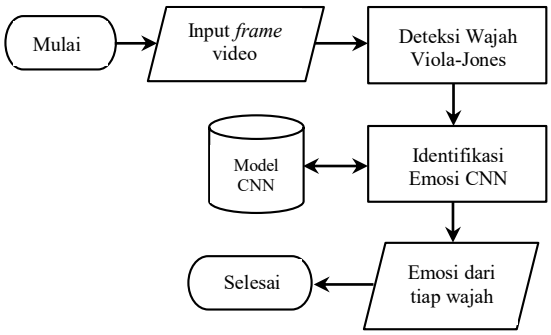
Makalah ini membahas pengembangan aplikasi untuk mengidentifikasi emosi manusia yang menggunakan video konferensi sebagai media komunikasi. Aplikasi ini dapat diaplikasikan terhadap semua jenis video konferensi baik secara *live* maupun rekaman, seperti pada rapat, perkuliahan, webinar, dan jenis pertemuan lainnya. Aplikasi yang dirancang menerapkan algoritma CNN untuk mengenali emosi manusia berdasarkan ekspresi wajah peserta dari konferensi video secara *online*. Aplikasi dapat melakukan deteksi wajah secara multi objek yaitu dapat mengenali emosi lebih dari satu wajah pada satu tampilan konferensi video secara *online*. Beberapa batasan yang diterapkan adalah pengelompokkan emosi dikategorikan hanya menjadi 4 kelas, yaitu netral, senang, marah, dan sedih. Selain itu citra wajah juga harus memiliki pencahayaan yang baik dengan posisi wajah frontal menghadap kamera. Perbedaan penelitian yang diusulkan ini dengan berbagai penelitian yang telah ada sebelumnya adalah pada penggunaan arsitektur VGG-16 pada algoritma CNN dengan data masukan berupa *frame* dari aplikasi konferensi video secara *online*. Aplikasi yang dikembangkan memiliki dua tahapan, yaitu pendeteksian area wajah (multi objek) secara bersamaan dan pengenalan emosi terhadap keseluruhan wajah yang terdeteksi pada *frame* konferensi video *online* tersebut. Data yang digunakan dalam eksperimen adalah rekaman

konferensi video *online* yang dikumpulkan oleh tim peneliti dengan variasi berupa jumlah orang yang terlibat dan variasi ekspresi wajah dari para peserta. Luaran dari aplikasi adalah pemberian tanda (*mark*) pada area wajah yang terdeteksi, serta jenis emosi yang teridentifikasi dari setiap wajah yang terdeteksi.

Makalah ini disusun dengan empat struktur, yaitu bab 1 berisi pendahuluan, bab 2 berisi penjelasan mengenai metodologi yang digunakan, bab 3 berisi hasil eksperimen dan analisis, serta bab 4 berisi kesimpulan dan saran serta rencana untuk penelitian selanjutnya.

2. METODE PENELITIAN

Sistem aplikasi yang dirancang merupakan aplikasi yang dapat mengenali emosi wajah dari *frame* suatu konferensi video secara *online*. Masukan dari sistem berupa rekaman video yang berisi wajah peserta konferensi, kemudian sistem secara otomatis melakukan pendeteksian area wajah serta mengenali emosi yang terkandung pada setiap wajah yang ditemukan. Terdapat dua tahapan utama dalam sistem yang dirancang, yaitu 1) tahapan pendeteksian wajah dan 2) tahapan pengenalan emosi. Proses pendeteksian wajah manusia yang terdapat dalam rekaman video menggunakan metode Viola-Jones, sedangkan proses pengenalan emosi pada wajah yang terdeteksi menggunakan metode Convolutional Neural Network (CNN). Diagram alir sistem yang dirancang untuk identifikasi emosi wajah dari konferensi video dapat dilihat pada Gambar 1.



Gambar 1 Skema Perancangan Sistem Identifikasi Emosi

2.1 Tahap Pendeteksian Wajah Viola-Jones

Tahap pertama pada sistem yang dirancang adalah pendeteksian wajah dengan metode Viola-Jones. Metode Viola-Jones diciptakan pada tahun 2001 oleh Paul Viola dan Michael Jones dengan menggabungkan konsep Haar, gambar integral dan fungsi AdaBoost, kemudian memproses seluruh fungsi tersebut menjadi pengklasifikasi Cascade. Algoritma dari metode Viola-Jones yang dilakukan pada proses pendeteksian wajah adalah sebagai berikut (Viola & Jones, 2001):

1. Pembacaan sampel citra wajah dari suatu frame.
2. Pembacaan fitur Haar dengan mengklasifikasikan data berdasarkan perbedaan nilai dari daerah gelap dan daerah terang. Jika perbedaan nilai antara daerah terang dan daerah gelap diatas nilai ambang, maka dapat disimpulkan bahwa fitur tersebut ada.
3. Perhitungan Integral Image pada sebuah citra dengan menambahkan nilai piksel citra secara bersamaan. Nilai integral masing-masing piksel adalah jumlah dari semua piksel yang ada pada citra tersebut.
4. Pemilihan fitur Haar sebagai nilai ambang menggunakan metode Adaboost. Metode ini menggabungkan banyak *classifier* lemah untuk membuat sebuah *classifier* kuat dengan cara mengelompokkan daerah. Selama proses filter, jika terdapat salah satu filter gagal buat melewatkan sebuah wilayah citra, maka wilayah itu langsung digolongkan menjadi bukan wajah. Hanya jika seluruh proses filter dapat dilalui, maka wilayah tersebut digolongkan daerah wajah.
5. Pengurutan Cascade *classifier* yaitu dengan bobot paling besar diletakkan dalam proses pertama kali, bertujuan buat menghapus wilayah citra bukan wajah secepat mungkin.

Setelah proses pendeteksian wajah selesai dilakukan, sistem telah memiliki satu atau beberapa area yang terdeteksi sebagai wajah, untuk kemudian dilanjutkan pemrosesannya pada tahap identifikasi emosi dari seluruh area wajah yang terdeteksi tersebut.

2.2 Tahap Identifikasi Emosi dengan Convolutional Neural Network (CNN)

Tahapan identifikasi emosi dilakukan terhadap *frame* wajah yang berhasil terdeteksi dari tahap sebelumnya. Proses pengenalan emosi dilakukan menggunakan metode Convolutional Neural Network (CNN). Metode CNN terdiri dari dua tahapan utama, yaitu tahap pelatihan fitur yang terdiri dari *convolutional layer*, ReLU (fungsi aktivasi) dan *pooling layer*, sedangkan tahap klasifikasi terdiri dari proses *flattening*, *fully connected layer*, dan prediksi. Setiap bagian yang ada pada CNN memiliki dua proses utama yaitu *feedforward* dan *backpropagation*.

Tahap pertama pada metode CNN adalah *convolutional layer* yang merupakan lapisan yang melakukan ekstraksi ciri yang terhubung ke area lokal citra masukan. Persamaan yang digunakan dalam proses perhitungan *convolutional layer* adalah sebagai berikut:

$$x(i,j) = \sum_m \sum_n w_{m,n}^{l} * o_{i+m,j+n}^{l-1} + b \quad (1)$$

dengan *x(i, j)* merupakan hasil perhitungan konvolusi pada posisi *(i, j)*, *l* adalah lapisan, *o(i,j)* merupakan nilai masukan, *w(m,n)* merupakan filter, *b* merupakan bias, serta *i* dan *j* masing-masing mewakili baris dan kolom piksel pada citra.

Tahapan kedua yaitu *pooling layer* yang merupakan lapisan yang berfungsi untuk memperkecil ukuran lapisan sebelumnya

(*downsampling*) pada dimensi spasial (lebar, tinggi). Terdapat 2 jenis *pooling layer*, yaitu *max pooling* dan *avarage pooling*. Proses *max pooling* dilakukan dengan pencarian dan penerapan nilai maksimum dari setiap bagian yang berada pada *feature map,* sedangkan proses *avarage pooling* melakukan perhitungan nilai rata-rata dari setiap bagian yang ada pada *feature map*. *Max pooling* merupakan metode yang paling sering digunakan termasuk digunakan pada penelitian ini. Selanjutnya, ReLU adalah fungsi aktivasi yang bertanggung jawab untuk dapat melakukan normalisasi pada nilai yang dihasilkan dari *convolutional layer*. Tahapan ReLU merupakan tahapan normalisasi nilai. ReLU akan menampilkan nilai secara langsung jika nilai positif sedangkan untuk nilai yang negatif akan diberi nilai 0.

Pada tahapan *classification*, proses pertama yang terjadi adalah *flattening* yang akan mengubah *feature map* pada layer sebelumnya menjadi vektor satu dimensi. Selanjutnya, tahapan berikutnya adalah pembentukan *fully-connected layer* untuk klasifikasi linier pada CNN. Pada *fully-connected layer* setiap neuron memiliki koneksi penuh ke semua neuron yang ada pada lapisan sebelumnya. Hasil luaran dari *fully-connected layer* berupa nilai *y* dengan parameter bobot *W* dan bias *b* dari suatu input *x* dapat dilihat pada rumus berikut:

$$y = \sum_i x_i \times W_i + b \quad (2)$$

Softmax merupakan sebuah fungsi aktivasi yang akan digunakan pada layer *output*. Fungsi dari *softmax* adalah mengambil angka dari vektor input yang telah melalui proses *fully-connected layer* dan mengubah angka tersebut menjadi dalam cakupan 0-1. Layer *output* memiliki banyak kesamaan dengan *fully-connected layer*, yang membedakan keduanya adalah pada layer *output* menggunakan fungsi aktivasi *softmax* dan pada *fully-connected layer* menggunakan fungsi aktifasi ReLU. Persamaan fungsi aktivasi *softmax* dapat dilihat pada rumus berikut:

$$p_i = \frac{e^{z_i}}{\sum_{j=1}^{N} e^{z_j}} \quad (3)$$

dengan p_i merupakan probabilitas pada vektor ke-i, z_i merupakan vektor masukan ke-i, dan z merupakan vektor masukan.

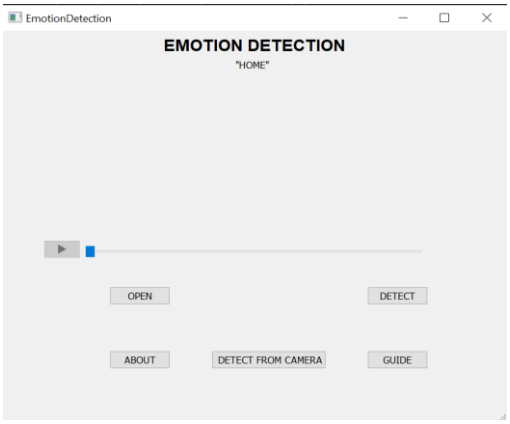
3. HASIL PERCOBAAN

3.1 Metode Pengujian

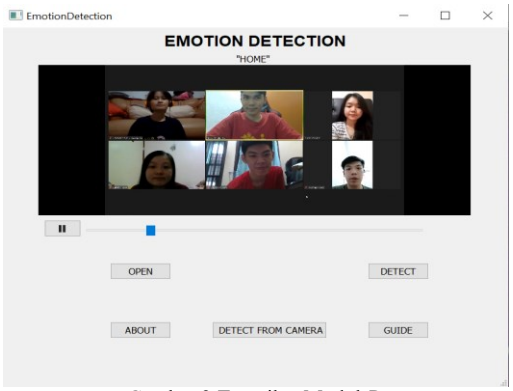
Pengujian dilakukan terhadap aplikasi yang dibuat untuk memastikan bahwa program dapat melakukan klasifikasi emosi pada wajah manusia dengan baik dan benar. Pengujian dilakukan dalam dua tahap, yaitu tahap pengujian modul yang bertujuan untuk menguji apakah modul yang dirancang telah menjalankan fungsi yang sesuai dengan spesifikasi, dan tahap kedua yaitu pengujian pada model CNN dengan tujuan untuk mengetahui seberapa efisien model dan seberapa akurat data luaran dari program.

3.2 Pengujian Terhadap Modul

Pada tahap pertama dilakukan pengujian terhadap modul untuk memeriksa apakah setiap tombol, menu, dan fungsinya telah berjalan dengan baik dan benar. Dalam program ini terdapat tiga buah modul yang terintegrasi, yaitu modul *Home*, modul *Detect*, dan modul *Help*.



Gambar 2 Tampilan Modul *Home*



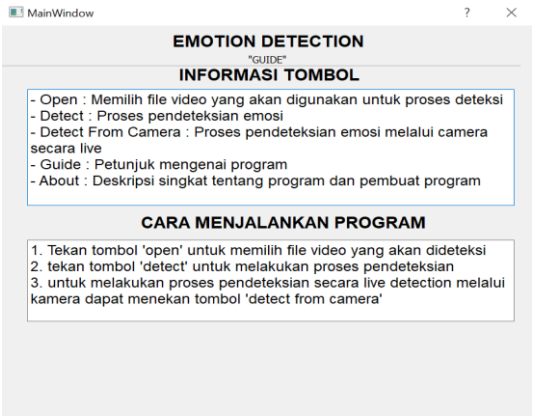
Gambar 3 Tampilan Modul *Detect*

Modul *Home* merupakan tampilan awal saat program dijalankan. Modul ini memiliki lima tombol dengan fungsi *link* untuk menampilkan modul lainnya serta terdapat satu *frame* untuk melakukan proses review terhadap citra yang telah dipilih untuk dilakukan proses pendeteksian. Tombol yang dimiliki pada modul ini antara lain adalah tombol *Open* yang memungkinkan pengguna memilih video rekaman dari suatu konferensi, tombol *Detect* yang berfungsi untuk melakukan proses pendeteksian wajah, serta tombol *Help* untuk membaca panduan. Tampilan dari modul *Home* yang dirancang dapat dilihat pada Gambar 2.

Modul kedua adalah modul *Detect* yang digunakan untuk melakukan proses pendeteksian. Pada modul ini, terdapat sebuah panel sebagai media untuk melihat proses klasifikasi berjalan. Program melakukan proses klasifikasi dari input citra (*frame* dari video) yang telah dipilih sebelumnya. Tombol

Detect from Camera juga tersedia dan dapat digunakan untuk melakukan proses pendeteksian wajah secara *live detection.* Pada menu ini, aplikasi akan membuka kamera dari perangkat yang menjalankan program dan panel akan muncul dan menampilkan proses klasifikasi. Tampilan dari modul *Detect* yang dirancang dapat dilihat pada Gambar 3.

Modul terakhir adalah modul *Help* yang meliputi informasi tombol-tombol yang terdapat pada program dan petunjuk penggunaan program. Tampilan dari modul *Help* yang dirancang dapat dilihat pada Gambar 4.



Gambar 4 Tampilan Modul *Help*

3.3 Pengujian Model CNN

Selanjutnya pengujian dilakukan terhadap model CNN yang terbentuk. Data yang digunakan dalam pengujian aplikasi berupa data dalam bentuk citra dua dimensi yang merupakan hasil dari cacahan video dari konferensi (*frame* dari video). Data latih berupa citra wajah didapatkan dari website *Kaggle* yang bernama FER2013 (Melinte & Vladareanu, 2020). Format bobot hasil pelatihan dari model CNN berformat .h5. Sedangkan data uji merupakan data yang dikumpulkan dan diperoleh oleh tim peneliti secara mandiri melalui variasi aplikasi konferensi video seperti Zoom dan Microsoft Teams dengan variasi jumlah orang dari 2 hingga 10 orang. Data latih dan data uji dalam percobaan ini berupa citra wajah manusia dengan empat kelas emosi yaitu, netral, senang, marah, dan sedih. Detail jumlah data latih serta data uji yang digunakan dalam percobaan dapat dilihat pada Tabel 1.

Tabel 1 Detail Jumlah Data yang Digunakan

Kelas Emosi	Data Latih	Data Uji
Marah	3993	960
Sedih	4938	1139
Senang	7164	1825
Netral	4982	1216
TOTAL	21077	5140

Pengujian terhadap hasil luaran aplikasi dilakukan terhadap dua dataset pengujian, yaitu Dataset 1 mencakup 68 citra uji dari 17 rekaman konferensi video serta Dataset 2 yang terdiri dari 64 set citra dari 16 orang responden yang melakukan pertemuan secara *live* dengan aplikasi konferensi video. Dalam proses pengambilan data uji ini, terdapat peran dari instruktur yang mengarahkan para peserta konferensi video untuk menunjukkan ekspresi tertentu sehingga diperoleh empat kelas emosi sesuai dengan target peneliti. Namun demikian, kelemahan dari cara pengumpulan data ini adalah karena emosi yang tampil bersifat buatan (*artificial*), maka tidak seluruh partisipan mampu menunjukkan ekspresi wajah sesuai dengan emosi tertentu dengan sempurna. Gambar 5 menampilkan sampel dari data uji yang dikumpulkan tim peneliti.

Gambar 5 Sampel Data Uji yang Digunakan

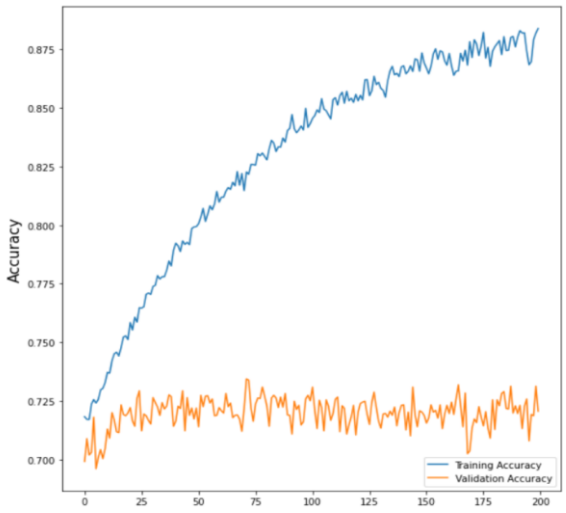
Hasil pengujian dari tahapan pendeteksian lokasi wajah dengan menggunakan metode Viola-Jones untuk Dataset 1 yaitu data pengujian dari rekaman konferensi video mendapatkan akurasi sebesar 78.19%, sedangkan untuk Dataset 2 dengan data uji berupa tangkapan langsung melalui kamera dengan mendapatkan akurasi sebesar 93.15%. Sehingga nilai rata-rata keakuratan sistem untuk mendeteksi lokasi wajah adalah 85.67%.

Selanjutnya dilakukan pembuatan model CNN dengan menggunakan arsitektur VGG-16 pada algoritma CNN dengan 200 epoch. Detail mengenai lapisan (*layer*) yang terdapat pada arsitektur VGG-16 yang digunakan pada tahap pembentukan model CNN dapat dilihat pada Gambar 6.

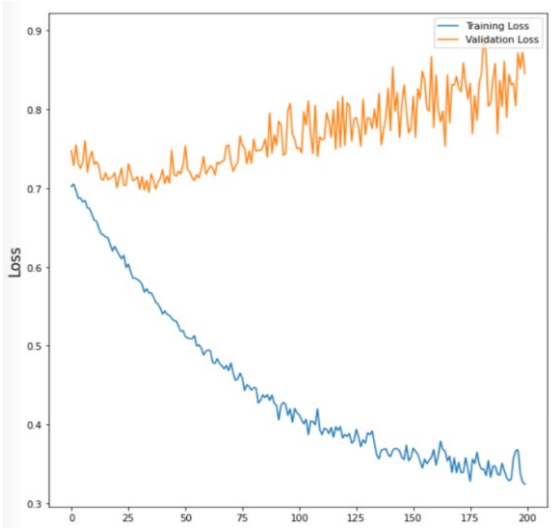
Layer (type)	Output Shape	Param #
batch_normalization_1 (Batch	(None, 48, 48, 1)	4
conv2d_6 (Conv2D)	(None, 48, 48, 36)	360
conv2d_7 (Conv2D)	(None, 48, 48, 36)	11700
max_pooling2d_3 (MaxPooling2	(None, 24, 24, 36)	0
conv2d_8 (Conv2D)	(None, 24, 24, 64)	20800
conv2d_9 (Conv2D)	(None, 24, 24, 64)	36928
max_pooling2d_4 (MaxPooling2	(None, 12, 12, 64)	0
conv2d_10 (Conv2D)	(None, 12, 12, 128)	73856
conv2d_11 (Conv2D)	(None, 12, 12, 128)	147584
max_pooling2d_5 (MaxPooling2	(None, 6, 6, 128)	0
dropout (Dropout)	(None, 6, 6, 128)	0
conv2d_12 (Conv2D)	(None, 6, 6, 256)	295168
conv2d_13 (Conv2D)	(None, 6, 6, 256)	590080
flatten (Flatten)	(None, 9216)	0
dense (Dense)	(None, 64)	589888
dropout_1 (Dropout)	(None, 64)	0
dense_1 (Dense)	(None, 64)	4160
dense_2 (Dense)	(None, 4)	260

Total params: 1,770,788
Trainable params: 1,770,786

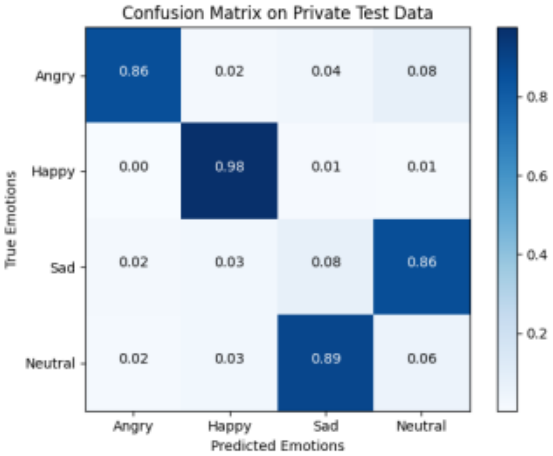
Gambar 6 Arsitektur Layer VGG-16 yang Digunakan



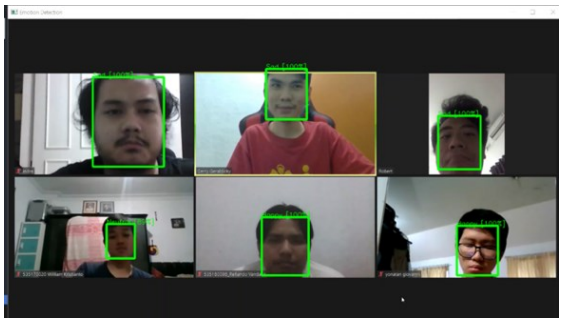
Gambar 7 Grafik Accuracy Pada Model VGG-16 (200 epoch)



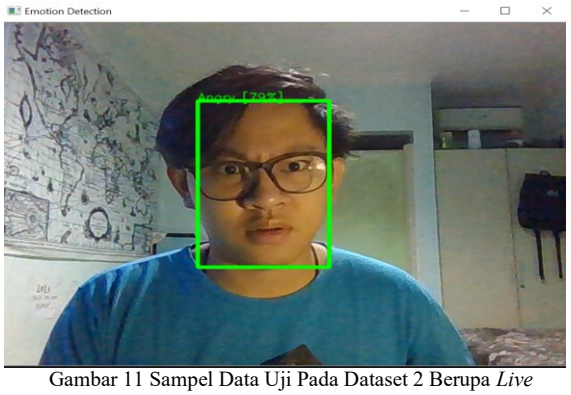
Gambar 8 Grafik Loss Pada Model VGG-16 (200 epoch)



Gambar 9 Confusion Matrix Pada Model VGG-16 (200 epoch)



Gambar 10 Sampel Data Uji Pada Dataset 1 Hasil Rekaman Konferensi Video



Gambar 11 Sampel Data Uji Pada Dataset 2 Berupa *Live Detection* dari Konferensi Video

Pengujian berikutnya dilakukan terhadap data *frame* video yang tidak dilatihkan sebelumnya. Pengujian dilakukan terhadap dua Dataset, yaitu Dataset 1 dengan data berasal dari rekaman konferensi video dan Dataset 2 yaitu data berasal dari *frame* video yang diperoleh secara *live detection*. Sampel data uji pada Dataset 1 tertera pada Gambar 10 dan sampel data uji pada Dataset 2 terlihat pada Gambar 11. Salah satu perbedaan signifikan dari kedua Dataset adalah pada Dataset 1 jumlah wajah yang terdapat pada *frame* video lebih dari satu orang, sementara pada Dataset 2 jumlah wajah yang terdapat pada *frame* hanya satu orang.

Tabel 2 Hasil Pengujian Pada Dataset 1 Berdasarkan Kelas Emosi

Kelas	*Sensitivity*	*Specificity*	*Precision*	*Negative Prediction Value*	*Accuracy*	*F1-Score*
Marah	0.097	0.979	0.611	0.764	0.758	0.168
Senang	0.724	0.531	0.349	0.848	0.581	0.471
Netral	0.339	0.867	0.458	0.799	0.736	0.390
Sedih	0.445	0.824	0.450	0.822	0.732	0.447

Tabel 3 Hasil Pengujian Pada Dataset 2 Berdasarkan Kelas Emosi

Kelas	*Sensitivity*	*Specificity*	*Precision*	*Negative Prediction Value*	*Accuracy*	*F1-Score*
Marah	0.889	0.500	0.889	0.500	0.818	0.889
Senang	0.318	0.625	0.700	0.250	0.400	0.438
Netral	0.500	0.143	0.077	0.667	0.188	0.133
Sedih	1	0.500	0.750	1	0.800	0.857

Untuk skenario pengujian pertama menggunakan Dataset 1, dilakukan perhitungan akurasi terhadap keberhasilan pendeteksian area wajah dari *frame* video. Perhitungan akurasi pendeteksian area wajah dilakukan dengan membandingkan jumlah kotak area wajah yang terdeteksi benar dengan seluruh jumlah kotak area yang terdeteksi oleh program. Untuk Dataset 1, kotak area wajah yang terdeteksi dengan benar berjumlah 182, sedangkan jumlah keseluruhan kotak yang terdeteksi berjumlah 459, maka nilai keberhasilan deteksi yang diperoleh adalah 39.65%. Sedangkan untuk Dataset 2, banyaknya area wajah pada dataset pengujian yang benar adalah 68, sementara banyaknya area yang terdeteksi adalah 49, maka nilai keberhasilan deteksi yang diperoleh sebesar 72.06%.

Pada eksperimen berikutnya, pengujian juga dilakukan untuk mengetahui akurasi proses identifikasi jenis emosi pada kedua Dataset. Jenis emosi pada wajah dibatasi pada 4 kelas, yaitu netral, senang, sedih, dan marah. Dalam melakukan analisis kinerja sistem yang dikembangkan, digunakan enam nilai analisis yang meliputi *sensitivity*, *specificity*, *precision*, *negative predictive value*, *accuracy* serta *F1-Score*. Tabel 2 menunjukkan nilai analisis terhadap Dataset 1, sedangkan Tabel 3 menunjukkan hasil pengujian dengan enam nilai analisis yang sama terhadap Dataset 2.

Berdasarkan Tabel 2, terlihat bahwa untuk Dataset 1, diperoleh nilai akurasi yang cukup tinggi untuk seluruh jenis kelas emosi. Secara rata-rata dari keseluruhan kelas, nilai akurasi dari sistem adalah 70.17% dengan nilai *sensitivity* 0.402, nilai *specificity* 0.800, nilai *precision* sebesar 0.467, nilai *negative predictive value* sebesar 0.808, dan nilai *F1-Score* rata-rata sebesar 0.369. Selanjutnya, Tabel 3 menampilkan hasil pengujian terhadap Dataset 2. Berdasarkan Tabel 2, diperoleh nilai rata-rata akurasi dari sistem sebesar 55.14% dengan nilai *sensitivity* 0.677, nilai *specificity* 0.442, nilai *precision* sebesar 0.604, nilai *negative predictive value* sebesar 0.604, dan nilai *F1-Score* rata-rata sebesar 0.579.

Dari beberapa percobaan yang telah dilakukan, terlihat bahwa hasil pengujian terhadap Dataset 2 yang mengandung data *live detection* memberikan hasil yang lebih rendah dibandingkan dengan hasil yang diberikan oleh Dataset 1 yang merupakan data hasil rekaman video konferensi. Perbedaan yang cukup signifikan pada hasil akurasi pengujian dari kedua dataset disebabkan oleh permasalahan data yang dikumpulkan dimana peserta pada *live detection* tidak mengeluarkan ekspresi yang kuat dan sungguh-sungguh, sehingga mengakibatkan raut wajah dari tiap peserta tidak terlihat jelas emosinya, maupun akibat pencahayaan yang kurang baik. Secara umum, model CNN dengan arsitektur VGG-16 cukup mampu mengenali emosi pada citra wajah dengan baik yaitu pada ekspresi senang, sedih, dan juga netral. Akan tetapi masih diperlukan perbaikan sistem untuk meningkatkan hasil identifikasi emosi yang tepat pada wajah manusia. Beberapa penyesuaian untuk perbaikan nilai akurasi seperti menambahkan teknik *transfer learning* atau penggunaan arsitektur lainnya perlu diujicobakan.

4. KESIMPULAN

Setelah melakukan proses pengujian, kesimpulan yang dapat diambil dari sistem pengidentifikasi emosi pengguna konferensi video menggunakan Convolutional Neural Network (CNN) adalah sebagai berikut:

1. Sistem mampu melakukan pendeteksian area wajah manusia secara otomatis dengan metode Viola-Jones dengan akurasi tertinggi sebesar 93.15%.
2. Model CNN dengan menggunakan arsitektur VGG-16 yang dilatihkan dengan 200 epoch memiliki performa akurasi yang cukup baik yaitu sebesar 88.39% untuk data latih dan validasi akurasi sebesar 70.17%.
3. Kemampuan sistem untuk melakukan pengenalan emosi sangat beragam tergantung pada jenis emosi yang akan dikenali, kejelasan emosi yang tergambarkan melalui wajah, serta suasana lingkungan sekitar.

Untuk penelitian lanjutan dapat dilakukan perbaikan yang mencakup:

1. Pengambilan data yang lebih baik dari sisi ekspresi wajah yang tampil, pencahayaan yang cukup, serta keseimbangan jumlah data.
2. Pengintegrasian aplikasi yang dibuat dengan aplikasi konferensi video seperti *Zoom*, *Microsoft Teams*, *Google Meet*, dan lain-lain sehingga pengguna dapat secara langsung memperoleh hasil dalam satu aplikasi aktif.
3. Ujicoba sistem dengan arsitektur dan parameter CNN lainnya untuk mendapatkan hasil yang lebih optimal.

DAFTAR PUSTAKA

ANDIKA, L.A., PRATIWI, H., HANDAJANI, S.S. 2019. Klasifikasi penyakit pneumonia menggunakan metode convolutional neural network dengan optimasi adaptive momentum. Indonesian Journal of Statistics and Its Applications, 3(3), pp. 331-340.

BATOTENG, F.G., PASIAK, T.F. dan TICOALU, S.H.R. 2015. Gambaran musculi facialis pada ekspresi wajah dan emosi dengan menggunakan facial action coding system pada calon presiden Jokowi. Jurnal e-Biomedik, 3(1), pp.280-284.

DU, S., MARTINEZ, A.M. 2015. Compound facial expressions of emotion: from basic research to clinical applications. Dialogues in Clinical Neuroscience, 17(4), pp.443-455.

GIANNOPOULUS, P., PERIKOS, I, HATZILYGEROUDIS I. 2018. Deep learning approaches for facial emotion recognition: a case study on FER-2013. Advances in Hybridization of Intelligent Methods, pp.1-16.

KIM, B.K., ROH, J., DONG, S.Y., LEE, S.Y. 2016. Hierarchical committee of deep convolutional neural networks for robust facial expression recognition. Journal on Multimodal User Interfaces, 10(2), pp.173-189.

KUSUMA, G.P., JONATHAN, LIM, A.P. 2020. Emotion recognition on FER-2013 face images using fine-tuned VGG-16. Advances in Science, Technology and Engineering Systems Journal, 5(6), pp.315-322.

LEE, M., LEE Y.K., LIM M-T., KANG, T-K. 2020. Emotion recognition using convolutional neural network with selected statistical photoplethysmogram features, Applied Science, 10(10), 3501.

LIEBOLD, B., RICHTER, R., TEICHMANN, M., HAMKER, F., and OHLER, P. 2015. Human capacities for emotion recognition and their implications for computer vision. I-com, 14(2), pp.126-137.

LINA, TAKAHASHI, T., IDE, I. dan MURASE, H. 2009. Incremental unsupervised-learning of appearance manifold with view-dependent covariance matrix for face recognition from video sequences. IEICE Trans. on Information and Systems, E92-D(4), pp.642-652.

MELINTE, D.O. and VLADAREANU, L. 2020. Facial expressions recognition for human-robot interaction using deep convolutional neural networks with rectified adam optimizer. Sensors, 20(8), 2393.

OH, G., RYU, J., JEONG, E., YANG, J.H., HWANG, S., LEE, S., and LIM, S. 2021. Sensors, 21(6), 2166.

OUYANG, X., KAWAII, S., GOH, E.G.H., SHEN, S., DING, W., MING, H., HUANG, D-Y. 2017. Audio-visual emotion recognition using deep transfer learning and multiple temporal models. ACM Conf. on Multimodal Interaction.

VIOLA, P. and JONES, M. 2001. Rapid Object Detection Using a Boosted Cascade of Simple Features. IEEE Conf. on Computer Vision and Pattern Recognition, 990517.

WANG, F., WU, S., ZHANG, W., XU, Z., ZHANG Y., WU C., and COLEMAN, S. 2020. Emotion recognition with convolutional neural network and EEG-based EFDMs. Neuropsychologia, 146(10), 107506.

YU, J. and ZHANG, W. 2021. Face mask wearing detection algorithm based on improved YOLO-v4. Sensors, 21(9), 3263.