Indonesian Journal of Electronics and Instrumentation Systems (IJEIS)
Vol.x, No.x, April xxxx, pp. 1~5
ISSN (print): 2088-3714, ISSN (online): 2460-7681
DOI: 10.22146/ijeis.xxxx

Klasifikasi Ekspresi Wajah Menggunakan CNN Dalam Keadaan Wild Setting Pada Virtual Meeting

Isnan Firmansyah[1], Diyah Utami K.P[2], Bakhtiar Alldino A.S[3]
[1] Program Studi Elektronika dan Instrumentasi, FMIPA UGM, Yogyakarta, Indonesia
[2, 3] Departemen Ilmu Komputer dan Elektronika, FMIPA UGM, Yogyakarta, Indonesia
e-mail: [1]Isnanfirmansyah2019@mail.ugm.ac.id, [2]diyah.utami.k@mail.ugm.ac.id
[3]b.alldino.as@ugm.ac.id,

Abstrak

Ekspresi wajah merepresentasikan perasaan dalam diri manusia serta dapat menjadi sebuah mediator dalam dunia sosial. Dalam rapat virtual, pemahaman terhadap suasana hati dan emosi peserta sangat dibutuhkan untuk menciptakan interaksi dan kerjasama yang baik. Emosi alami akan muncul ketika peserta berbicara secara spontan dengan keadaaan lingkungan alami tanpa dikondisikan sebelumnya. Oleh karena itu, dibutuhkan sebuah sistem yang mampu mengetahui emosi manusia secara otomatis dalam keadaan spontan (wild setting) pada virtual meeting sehingga ekspresi lebih alami. Teknik machine learning digunakan untuk melakukan klasifikasi ekspresi wajah pada virtual meeting.

Penelitian ini membandingkan performa arsitektur LightCNN dan EfficientNet. Pelatihan dilakukan pada dataset gabungan antara FER-2013, Extended and Augmented Google FER dan CK+. Dataset berjumlah 67.362 citra terbagi menjadi data training 60.184 buah, data validation 3.589 buah, dan data testing 3.589 buah. Input model arsitektur EfficientNet divariasikan menjadi 48x48 dan 224x224 pixels. Optimasi learning rate dilakukan untuk menemukan performa tertinggi dari arsitektur terbaik.

Hasil penelitian menunjukan bahwa arsitektur terbaik adalah EfficientNet dengan input 48x48 pixel. Nilai parameter learning rate paling optimal yaitu 0,0005. performa model dalam klasifikasi ekspresi wajah mencapai akurasi 90,50%, presisi 89,50%, recall 90,69% dan F1-score 90,06%. Implementasi EfficientNet pada video virtual meeting untuk mengklasifikasikan ekspresi wajah mendapatkan performa akurasi sebesar 96,18%.

Kata kunci *- Deteksi Ekspresi Wajah, Pengenalan Ekspresi Wajah (FER), Virtual Meeting, LightCNN, EfficientNet*

Abstract

Facial expressions represent feelings in humans and can be a mediator in the social world. In virtual meetings, understanding the moods and emotions of participants is necessary to create good interaction and cooperation. Natural emotions will arise when participants speak spontaneously in a natural environment without being conditioned beforehand. Therefore, a system is needed that is able to automatically determine human emotions in a spontaneous state (wild setting) at virtual meetings so that expressions are more natural. Machine learning techniques are used to classify facial expressions in virtual meetings.

This research compares the performance of LightCNN and EfficientNet architectures. Training was conducted on a combined dataset between FER-2013, Extended and Augmented Google FER and CK+. The dataset amounted to 67,362 images divided into 60,184 training data, 3,589 validation data, and 3,589 testing data. The input of EfficientNet architecture model is varied into 48x48 and 224x224 pixels. Learning rate optimization is performed to find the highest performance of the best architecture.

The results showed that the best architecture is EfficientNet with 48x48 pixel input. The most optimal learning rate parameter value is 0.0005. model performance in facial expression classification achieved accuracy of 90.50%, precision 89.50%, recall 90.69% and F1-score 90.06%. EfficientNet implementation on virtual meeting videos to classify facial expressions gets an accuracy performance of 96.18%.

Keywords - *Facial Emotion Detection, Facial Emotion Recognition, Facial Expression Recognition (FER), Virtual Meeting, LightCNN, EfficientNet*

Received June 1st,2012; Revised June 25th, 2012; Accepted July 10th, 2012

1. PENDAHULUAN

Ekspresi wajah merupakan salah satu bentuk komunikasi non-verbal yang digunakan oleh manusia. Ekpresi wajah dapat menyampaikan informasi sebesar 55% dalam komunikasi, komunikasi vokal sebesar 38% dan komuniaksi verbal hanya 7% [1]. Ekspresi wajah juga dapat menggambarkan emosi seseorang. Berdasarkan penelitian Paul Ekman emosi dasar seseorang dikelompokan menjadi enam yaitu senang, sedih, marah, terkejut, takut, dan jijik [2]. Emosi dari ekspresi wajah saling berhubungan dengan proses intrapersonal dan hasil social [3].

Ekspresi wajah merepresentasikan perasaan dalam diri manusia dan dapat menjadi sebuah mediator dalam dunia sosial. Dalam rapat virtual atau virtual meeting, pemahaman terhadap emosi dan keadaan suasana hati peserta sangat dibutuhkan untuk menciptakan interaksi dan kerjasama yang baik [4]. Rapat virtual yang seringkali dilakukan secara tatapmuka memiliki keterbatasan dalam interaksi menggunakan bahasa tubuh membuat persepsi dalam mengenal emosi seseorang menjadi terbatas.

Penelitian Shao & Qian menunjukan bahwa Light-CNN dapat mendeteksi tujuh emosi wajah [5]. Model dilatih dengan dataset yang memiliki citra wajah yang diambil secara alami yaitu FER-2013. Permasalahan overfitting dapat diatasi dengan metode yang diusullkan. Keberhasilan klasifikasi hanya menggunakan parameter tidak lebih dari dua juta. Hal tersebut enam kali lebih sedikit daripada parameter yang digunakan ketika menggunakan arsitektur pretrained CNN. Arsitektur Light-CNN dinilai cukup baik untuk menyelesaikan permasalahan klasifikasi ekspresi wajah, Namun akurasi yang hasilkan hanya sekitar 68%.

EfficientNet merupakan arsitektur lain yang mampu mengatasi permasalahan klasifikasi. Penelitian ini menggunakan arsitektur yang efisien karena menggunakan metode compound scaling method sehingga dapat ditingkatkan sesuai dengan kebutuhan. Hal ini memungkinkan untuk mengubah width, depth, and resolution dari sebuah Convolutioinal Network. Parameter yang digunakan 21 kali lebih sedikit daripada arsitektur lain dengan akurasi yang hampir sama [6]. EfficientNet mampu mendeteksi 10 kategori dengan baik pada dataset CIFAR-10 yaitu 98.1%. Namun penelitian tersebut tidak diterapkan pada klasifikasi ekspresi wajah.

Keterbatasan pada virtual meeting untuk memahami ekspresi peserta salah satunya dikarenakan interaksi hanya terjadi tatap muka secara daring. Hal tersebut membuat komunikasi dan kerjasama antar anggota kelompok menjadi terhambat. Dibutuhkan sebuah sistem yang mampu mengetahui emosi manusia secara otomatis sehingga terbentuk interaksi sosial yang baik meskipun secara daring. Pendeteksian emosi pada virtual meeting dengan keadaan wild setting perlu dilakukan. Keadaan wild setting adalah keadaan saat peserta berbicara secara spontan dengan keadaaan ylingkungan yang alami tanpa dikondisikan sebelumnya seperti pose wajah dan pencahayaan.

Oleh karena itu, dalam penelitian ini akan dilakukan perbandingan metode Light-CNN dan EfficientNet untuk mengetahui performa terbaik dalam klasifikasi ekspresi wajah pada virtual meeting. Dua arsitektur tersebut dipilih karena memiliki jumlah parameter yang rendah sehingga memiliki waktu komputasi yang lebih cepat. Proses training model akan menggunakan dataset gabungan FER-2013 dengan dataset lain. Dataset FER-2013 yang memiliki citra wajah dengan posisi wajah acak dan data diambil secara alami pada kondisi lingkungan yang alami atau ‘wild setting’. Performa arsitektur akan dibandingkan berdasarkan akurasi, recall, F1-Score, dan jumlah parameter sebagai evaluasi kinerja masing-masing model. Selanjutnya, diharapkan dapat diperoleh arsitektur dengan performa terbaik dalam klasifikasi ekspresi wajah manusia.

Berdasarkan pemaparan sebelumnya, dapat dirumuskan bahwa rumusan masalah pada penelitian ini adalah deteksi ekspresi wajah dibutuhkan dalam interaksi manusia pada virtual meeting dalam keadaan ‘*wild setting’* sehingga terdeteksi ekspresi yang lebih alami. Tujuan dari penelitian ini adalah melakukan pengenalan ekspresi wajah menggunakan arsitektur CNN pada virtual meeting dalam keadaan lingkungan yang alami.

2. METODE PENELITIAN

2.1 Facial Expression Recognition (FER)

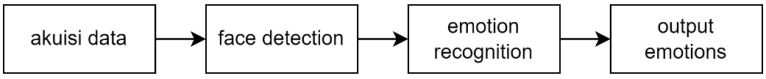
Facial Expression Recognition (FER) atau pengenalan ekspresi wajah merupakan proses untuk mengetahui emosi atau perasaan yang muncul dari wajah seseorang. Secara universal emosi dasar manusia dibedakan menjadi enam emosi dasar yaitu senang, sedih, marah, terkejut, takut, dan jijik. Selain enam emosi tersebut maka dikelompokan sebagai emosi netral.

FER merupakan satu bagian penting dalam human-computer interaction yang membuat komputer memahami ekspresi wajah berdasarkan pikiran manusia [7]. Terdapat banyak dataset citra wajah yang berisi aktor manusia yang menujukan emosi dasar dengan kondisi lingkungan yang terkendali. Selain, itu terdapat jenis dataset citra wajah yang diambil secara spontan dalam suasana yang tidak terkendali.

Metode machine learning tradisional seperti support vektor machine, dictionary learning, dan bayessian classifier berhasil mengklasifikan ekspresi yang berpose dalam lingkungan yang terkendali. Namun, metode tersebut tidak fleksibel ketika mengklasifikan citra yang diambil dengan cara yang tidak terkendali atau secara spontan (in wild setting). Selain itu, pendekatan yang dilakukan hanya bergantung pada dataset sebagai data training.

Dalam penelitian Mollahosseini dkk, [8] membedakan konsep tentang perilaku wajah menjadi dua yaitu message-based dan sign-based. Algoritma klasifikasi ekspresi wajah dapat dikelompokan berdasarkan pendekatan menggunakan sign-based atau message-based. Ketika menggunakan sign-based maka model akan dilatih untuk mendeteksi AU kemudian dikonversi menggunakan EMFACS. Berbeda ketiga menggunakan pendekatan message-based, model akan dilatih dengan dataset yang sudah dianotasi. Pada penelitian ini, akan menggunakan message-based untuk membuat model neural network.

Secara umum ekspresi wajah manusia dapat diklasifikan menngunakan conventional-learning dan deep-learning [9]. Conventional-learning merupakan metode yang melakukan ektraksi hand-crafted feature untuk mengambil fitur dari sebuah citra. hand-crafted feature adalah fitur yang didapat dari hasil algoritma dengan menggunakan informasi pada gambar itu sendiri. Setelah itu, fitur yang sudah terekstraksi dimasukan ke classifier. Deep-learning adalah metode yang otomatis melakukan ekstraksi fitur dan melakukan klasifikasi secara end-to-end. Proses FER menggunakan deep-learning disajikan pada gambar 1.

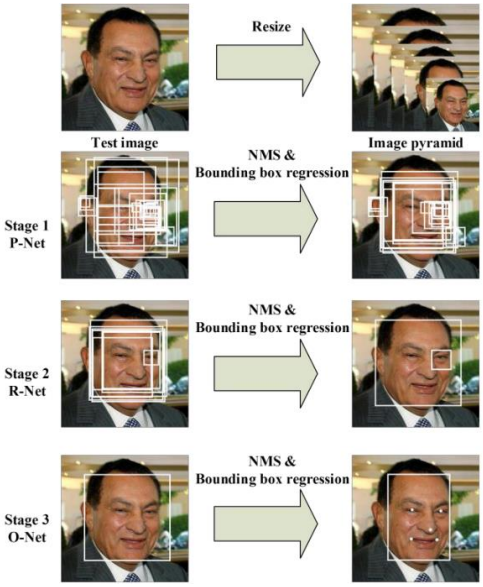


Gambar 1 Kerangka kerja FER

2.2 Face Detection

Face Detection atau dekteksi wajah merupakan proses pengolahan citra untuk mendeteksi keberadaan wajah manusia dalam sebuah citra. *Face detection* dapat dilakukan dengan metode seperti knowledge-based, feature-based, template matching, dan appearance-based [9]. Knowledge-base adalah cara deteksi wajah manusia berdasarkan sebuah aturan-aturan khusus yang dibuat. Selanjutnya, feature-based merupakan cara mencari bagian kecil dari wajah seperti mata dan hidung sehingga dapat dideteksi sebagai wajah manusia. Metode feature-based rentan terhadap perubahan cahaya dan noise. Selain itu, template matching merupakan teknik yang mendeteksi wajah dengan cara membandingkan sebuah citra masukan dengan citra lain. Citra pembanding dapat berupa citra wajah standar atau dapat citra wajah tertentu. Berbeda dengan appearance-based, metode ini menerapkan analisis secara statistik maupun menerapkan machine learning untuk mengidentifikasi wajah manusia. Metode ini yang sering diterapkan untuk klasifikasi emosi wajah. Multi-task Cascaded Convolutional Networks (MTCNN) adalah metode untuk mendeteksi wajah dan posisi landmark pada wajah. Pada penelitian K. Zhang dkk, [10] menjelaskan bahwa MTCNN menerapkan tiga tahap multi-task

deep convolutional networks. Tiga tahap tersebut terdiri dari Proposal-Network(P-Net), Refinement Network (R-Net), dan Output Network (O-Net).



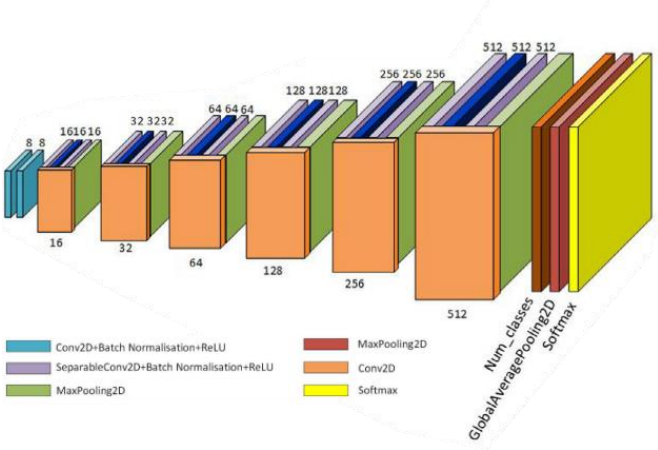
Gambar 2 Kerangka kerja MTCNN

Arsitektur MTCNN dilatih untuk dapat menklasifikasikan wajah atau bukan wajah, melaukan bounding box regression, dan menemukan lokasi landmark pada wajah. Masing-masing arsitektur untuk setiap langkah ditunjukan pada gambar 2. Keluaran dari metode ini adalah kotak pembatas atau bounding-box dan landmark yang terdeteksi. Landmark pada wajah meliputi mata kiri, mata kanan, hidung, mulut kiri, dan mulut kanan. Metode tersebut mendapatkan nilai recall yang tinggi ketika diuji pada dataset WIDER FACE sebesar 85.5%. Selain itu, MTCNN mendapatkan nilai mean error yang rendah pada AFLW yaitu 6.9%.

2.3 Light CNN

Light-CNN dikenalkan oleh Shao memiliki arsitektur yang sederhana [5]. Arsitektur yang terinspirasi dari Xception and ResNet dengan menambahkan residual convolution. Sebelum memasuki arsitektur, citra masukan akan diubah menjadi 64x64. Pada layer convolution pertama dan kedua menggunakan delapan filter berukuran 3x3, stride 1x1 dan dilanjutkan batch normalization dengan fungsi aktifasi ReLu. Setelah itu akan memasuki modul dengan residual convolution. Pada dua layer terakhir terdapat global average pooling (GPA) dan softmax. GPA digunakan untuk menghindari terjadinya overfiting. Light-CNN membutuhkan input dengan resolusi 224x224 satu channel. Total digunakan sebanyak 28 layer dimulai dari layer input sampai dengan GPA dalam arsitektur Light-CNN.

Light-CNN memiliki enam modul depthwise separable residual convolution sebagai komponen utama. Setiap modul memiliki tiga layer separable convolution dan diikuti satu layer convolution. Setiap modul menggunakan fungsi aktifasi ReLu. Arsitekur disajikan secara sederhana pada gambar 3.



Gambar 3 Arsitektur Light-CNN

Metode penelitian meliputi analisa permasalahan, arsitektur atau rancangan metode yang digunakan untuk menyelesaikan masalah**.** Analisa permasalahan mendeskripsikan permasalahan yang ada dan diselesaikan dalam penelitian ini. Rancangan menggambarkan cara penyelesaian masalah dan sebaiknya disajikan dalam bentuk diagram dengan penjelasan yang lengkap. Misalnya diagram pemrosesan data, dari data mentah sampai selesai, diagram rancangan perangkat keras.

2.4 EfficientNet

Arsitektur EfficientNet adalah arsitektur CNN yang diusulakan oleh [6] pada tahun 2019. EfficientNet merupakan arsitektur yang menerapkan compound scaling method untuk meningkatkan akurasi dan ukuran model. Compound scaling method memungkinkan untuk mengatur depth, width, resolution dari sebuah arsitektur CNN. EfficientNet menggunakan blok utama MBConv yang terdiri dari proses optimasi squeeze-and-excitation. Citra input akan melalui sebuah layer Convolutional setelah itu akan melalui serangkaian blok MBConv sampai dengan melewati satu convolutional layer, pooling, dan Fully Connected layer pada langkah terakhir. Rangkaian MBConv membantu mengurangi jumlah operasi yang dibutuhkan sehingga secara keseluruhan menurunkan ukurang dari model. Arsitektur pada gambar 4 menjadi baseline arsitektur EfficientNet yang disebut EfficientNet-B0. Selanjutnya dapat dikembangkan mulai dari EfficientNet-B1 – B7 dengan merubah parameter scaling.

Stage i	Operator $\hat{\mathcal{F}}_i$	Resolution $\hat{H}_i \times \hat{W}_i$	#Channels $\hat{C}_i$	#Layers $\hat{L}_i$
1	Conv3x3	224×224	32	1
2	MBConv1, k3x3	112×112	16	1
3	MBConv6, k3x3	112×112	24	2
4	MBConv6, k5x5	56×56	40	2
5	MBConv6, k3x3	28×28	80	3
6	MBConv6, k5x5	14×14	112	3
7	MBConv6, k5x5	14×14	192	4
8	MBConv6, k3x3	7×7	320	1
9	Conv1x1 & Pooling & FC	7×7	1280	1

Gambar 4 Arsitektur EfficientNet

2.5 Akuisisi data dan Preprocessing

Dataset yang digunakan pada penelitian ini adalah dataset gabungan dari FER-2013, Extended and Augmented Google FER, dan CK+. Hanya tujuh kelas utama yang akan digunakan yaitu marah, jijik, takut, senang, sedih, terkejut, dan netral. Kelas contempt pada

Extended and Augmented Google FER dan CK+ tidak gunakan. Total citra yang digunakan yaitu 67.362 buah berukuran 48x48 pixels dengan mode warna grayscale. Penggabungan data bertujuan untuk mendapatkan dataset yang lebih bervariasi sehingga mendekati keadaan sebenarnya.

Dataset gabungan sudah dianotasi kedalam tujuh emosi manusia yaitu senang, sedih, marah, takut, terkejut, jijik, dan netral. Masing-masing kelas memiliki jumlah yang berbeda. Jumlah citra untuk setiap emosi pada setiap dataset disajikan dalam tabel 4.1. Pembagian data yaitu data training sebanyak 60.184, data validation 3.589, dan data testing 3.589.

Data training merupakan data gabungan dari data kategori Training data dataset FER-2013, data kategori train Augmented and Extended Google FER tanpa kelas contempt, dan data kategori train CK+ kecuali emosi contempt. Data validation terdiri dari data kategori PublicTest dataset FER-2013. Data testing diambil dari dataset FER-2013 kategori PrivateTest. Data testing dipilih dari dataset FER-2013 dikarenakan data testing memiliki banyak variasi keadaan seperti usia, gender, pose wajah, oklusi, dan pencahayaan. Persebaran emosi pada dataset gabungan lebih lengkap ditunjukan pada Tabel 1.

Tabel 1 Persebaran emosi pada dataset gabungan

	Angry	Disgust	Fear	Happy	Neutral	Sad	Surprise
Training	8092	2893	8267	14613	9923	9775	6621
Validation	467	56	496	895	607	653	415
Test	491	55	528	879	626	594	416
Total	9050	3004	9291	16387	11156	11022	7452

Selanjutnya, citra wajah yang sudah diproses akan melewati tahan preprocessing berikutnya. Resolusi dan channel sebuah citra akan dirubah sesuai arsitektur yang akan diuji. Pada arsitektur Light-CNN resolusi citra akan diubah menjadi 224x224 dan mengunakan satu channel atau *grayscale*. Arsitektur EfficientNet menggunakan resolusi 48x48 dengan mode warna RGB.

2.6 Data Augmentasi

Data Augmentasi bertujuan untuk membuat data lebih bervarisai dan lebih banyak. Penerapan data augmetasi akan menggunakan ImageDataGenerator yang disediakan oleh *library* Keras API. Teknik augmentasi yang akan dilakukan yaitu *horizontal flips, rotation, shear, dan zooming*. *Zooming* merupakan teknik yang digunakan untuk membuat variasi dengan memperbesar atau memperkecil citra asli. *Shear* adalah teknik augmentasi untuk memberikan distorsi citra terhadap sumbu x dan y.

2.7 Pelatihan sistem

Sistem akan dilatih dengan menggunakan citra wajah pada data *training* kemudian dilakukan validasi menggunakan data *validation*. Sistem akan dilatih menggunakan dua arsitektur yaitu Light-CNN dan EfficinetNet. Rancangan pelatihan sistem ditunjukan pada gambar x. Sistem akan dilatih setelah dilakukan preprocesing dan data augmentasi. Pelatihan model akan dilakukan secara bergantian kemudian dilakukan evaluasi untuk setiap asitektur. Resolusi dan channel sebuah citra akan dirubah sesuai arsitektur yang akan diuji. Pada arsitektur Light-CNN resolusi citra akan diubah menjadi 224x224 dan mengunakan satu channel atau *grayscale*. Arsitektur EfficientNet menggunakan resolusi 48x48 dengan mode warna RGB.

Proses training membutuhkan parameter yang tepat untuk mendapatkan akurasi yang tinggi. Oleh karena itu, diperlukan *tuning hyperparameter* untuk menemukan parameter yang paling optimal pada model klasifikasi ekspresi wajah. Parameter yang terbaik kemudian diambil untuk dipilih parameter yang lain. Selain itu, pada penelitian ini juga membandingkan dua

arsitektur yaitu LightCNN dan EfficientNet. Adapun variasi parameter training yaitu learning rate 0.1, 0.01, 0.001, dan 0.0005.

2.8 Pengujian Sistem

Pengujian sistem akan dilakukan untuk mengetahui performa model dalam mengklasifikan ekspresi wajah. Model akan diuji menggunakan data uji yang belum pernah diikutkan dalam proses training. Selain itu, model akan melakukan klasifikasi ekspresi wajah pada sebuah data citra dan data video virtual meeting [11]. Performa model kemudian akan ditinjau menggunakan confusion matrik. Evaluasi performa model akan dilakukan dengan menghitung nilai akurasi, presisi, recall, F1-score yang dihasilkan untuk setiap arsitektur yang diimplementasikan. Hasil pengujian model akan disajikan dalam bentuk confusion matrix. Dari confusion matrix tersebut kemudian bisa didapatkan nilai True Positive (TP), True Negative (TN), False Positive (FP), danFalse Negative (FN), sehingga dapat dihitung nilai akurasi, presisi, recall, dan f1-score dengan persamaan (1), (2), (3), dan (4)

$$akurasi = \frac{total\ prediksi\ benar}{total\ prediksi} \tag{1}$$

$$recall = \frac{TP}{TP + FN} \tag{2}$$

$$precision = \frac{TP}{TP + FP} \tag{3}$$

$$F1\ Score = 2 \times \frac{precision \times recall}{precision + recall} \tag{4}$$

3. HASIL DAN PEMBAHASAN

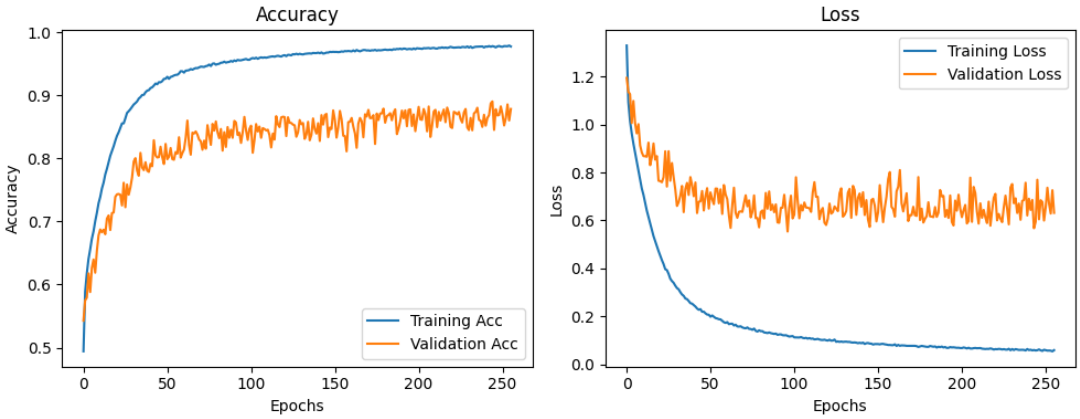
3.1 Hasil Variasi Arsitektur

Sistem klasifikasi membutuhkan arsitektur yang tepat untuk dapat melakukan pengenalan ekspresi wajah. Pada penelitian ini akan mevariasikan dua arsitektur yaitu LightCNN dan EfficientNet. Citra sebagai masukan sudah disesuaikan pada tahap preprocessing. Pemilihan arsitektur mempengaruhi akurasi model dalam mengklasifikan emosi atau ekspresi wajah. Ketiga model akan melakukan klasifikasi caitra ke dalam tujuh jenis emosi yaitu marah, jijik, takut, senang, sedih, terkejut, dan netral. Pada pengujian arsitektur, parameter learning rate akan ditetapkan menggunakan 0.001 dan optimizer Adam. Parameter lain yang digunakan saat training dapat dilihat pada tabel 6.1. Sedangkan arsitektur EfficientNet menggunakan weight dari pretrained model yaitu ‘imagenet’.

Tabel 2 Parameter yang digunaan saat training

Model	Parameters	Keterangan
LightCNN	Image Size	224x224 – *grayscale*
	Optimizer	Adam
	Learning Rate	0.001
EfficientNetB0	Image Size	48x48 – RGB
	Optimizer	Adam
	Learning Rate	0.001
	Weight Initializer	Imagenet
EfficientNetB0	Image Size	224x224 – RGB
	Optimizer	Adam
	Learning Rate	0.001
	Weight Initializer	Imagenet

Model EfficientNet memiliki beberapa versi mulai dari B0-B7. Pada pengujian ini menggunakan EfficientNet versi B0. EfficientNetB0 memiliki jumlah parameter sebanyak 4,058,538. Hasil dari arsitektur berukuran 47.2 MB. Model dilatih dengan input gambar 48x48 menggunakan fungsi optimezer Adam, fungsi loss c*ategorical crossentropy loss*, dan learning rate 0.001. Model dilatih dengan epoch 256, menghasilkan akurasi pelatihan 97.82 % dan loss 0.0586. Selain itu, didapatkan *validation accuracy* dan *validation loss* berturut-turut sebesar 89.05 % dan 0.5946. Grafik akurasi dan loss pada tahap training disajikan pada gambar 5.



Gambar 5 Grafik Akurasi dan Loss model EfficientNet

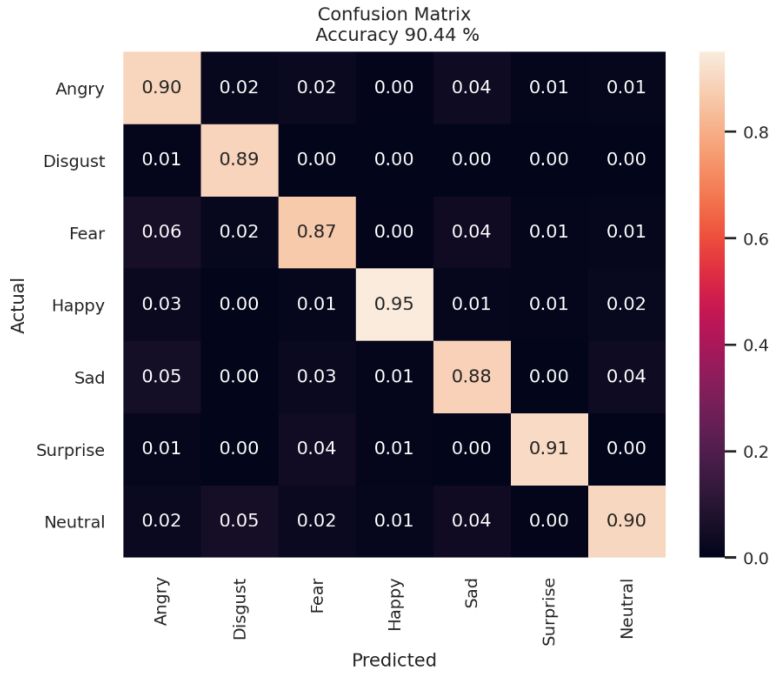
Model EfficientNet kemudian diuji menggunakan data *testing* dapat mencapai akurasi sebesar 90.44%, Presisi sebesar 90.21% , recall mencapai 89.89%, dan F1-score sebesar 90.01%. Akurasi, presisi, recall, dan F1-score dari model tidak memiliki perbedaan yang besar. Hasil evaluasi model EfficientNet terhadap data uji selengkapnya disajikan dalam gambar 6.

Classification Report:

	precision	recall	f1-score	support
Angry	0.83	0.90	0.86	491
Disgust	0.91	0.89	0.90	55
Fear	0.87	0.87	0.87	528
Happy	0.96	0.95	0.96	879
Sad	0.87	0.88	0.88	594
Surprise	0.96	0.91	0.93	416
Neutral	0.91	0.90	0.91	626
accuracy			0.90	3589
macro avg	0.90	0.90	0.90	3589
weighted avg	0.91	0.90	0.90	3589

Gambar 6 Classification Report dari model EfficientNet

Hasil prediksi menujukan bahwa akurasi tertinggi yaitu 95% ketika model melakukan prediksi pada ekspresi senang. Sedangkan, akurasi sebesar 88% pada ekspresi sedih merupakan akurasi terendah untuk model EfficientNet. Gambar 7 menujukan hasil visualisasi confussion matrix untuk setiap kelas.



Gambar 7 Confussion Matrix dari model EfficientNet

Arsitektur EfficientNet dengan input 48x48 memberikan akurasi terbaik dibandingkan dengan LightCNN. Akurasi EfficientNet saat pelatihan mencapai 97.82% sedangkan LightCNN sebesar 82.63%. Pada saat pengujian akurasi EfficientNet sebesar 90.44% berbeda dengan LightCNN yang mendapatkan akurasi sebesar 77.85%. Akurasi model EfficientNet lebih unggul ketika pelatihan maupun saat pengujian. Perbandingan performa pada saat pelatihan dapat dilihat pada tabel 3 sedangakan hasil evaluasi model saat pengujian dapat dilihat pada tabel 4.

Tabel 3 Perbandingan performa Model pada data training

Metrik	**LightCNN**	**EfficientNetB0 48x48**	**EfficientNetB0 224x224**
Akurasi	82.63	97.82	94.25
Validation Accuracy	76.62	89.05	88.02
Loss	0.4802	0.0586	0.1622
Validation Loss	0.7261	0.5946	0.4868

Tabel 4 Perbandingan performa Model pada data testing

Metrik	**LightCNN**	**EfficientNet 48x48**	**EfficientNet 224x224**
Akurasi	77.85	90.44	89.63
Presisi	76.48	90.21	89.18
Recall	75.11	89.89	89.12
F1-score	75.74	90.01	89.11

Bedasarkan pelatihan dan pengujian yang sudah dilakukan, arsitektur terbaik untuk klasifiaksi ekspresi adalah EfficientNetB0 dengan input 48x48. Oleh karena itu model ini akan digunakan sebagai model pengklasifikasi ekspresi wajah pada *virtual meeting*.

3.2 Proses Tuning Hyperparameter

Proses tuning bertujuan untuk mencari parameter yang paling sesuai untuk mendapatkan akurasi tertinggi. Model EfficientNetB0 dengan input 48x48 menjadi model yang

terbaik setelah dibandingkan dengan model lainnya. Model kemudian akan dilakukan pengujian dengan 4 variasi *learning rate* yaitu 0.1, 0.01, 0.001, dan 0,0005. Adapun perbandingan hasil pelatihan dapat dilihat pada tabel 5.

Tabel 5 Perbandingan Hasil Pelatihan terhadap variasi learning rate

Learning-rate	Akurasi Training	Akurasi Validation	Loss Training	Loss Validation	Akurasi Testing
0.1	71.04	67.32	0.7762	0.9296	67.12
0.01	83.90	75.90	0.4406	0.6972	76.01
0.001	97.82	89.05	0.0586	0.5946	90.44
0.0005	98.18	89.91	0.0495	0.5997	90.50

Pengujian learning rate pertama yaitu menetapkan nilai learning rate menjadi 0.1. Selanjutnya, model akan dilatih ulang dengan nilai learning rate lainnya. Pengujian dengan learning rate 0.1 menghasilkan nilai akurasi 67.12%. pada pengujian selanjutnya naik menjadi 76.11%. Model mencapai akurasi 90.44% naik sebesar 14,43% setelah menggunakan learning rate 0.001. Pada pengujian selanjutnya yaitu dengan nilai learning rate 0.0005 mendapat akurasi 90.50%. Perubahan nilai akurasi *training* berbanding lurus dengan validation accuracy dan akurasi *testing*. Loss training cenderung semakin kecil ketika menggunakan learning rate yang kecil.

Perubahan akurasi cukup signifikan ketika digunakan learning rate sebesar 0.001. namun pengujian menggunakan learning rate 0.0005 mendapatkan hasil *validation accuracy* dan akurasi *testing* paling tinggi yaitu sebesar 89.91%. Adapun hasil performa variasi learning rate disajikan pada tabel 6.6. nilai akurasi, presisi dan recall tertingi didapatkan ketika menggunakan learning rate 0.005. F1-score tertinggi tetap diperoleh dengan learning rate yang kecil yaitu 0.0005. Adapun perbandingan hasil performa model dapat dilihat pada tabel 6.

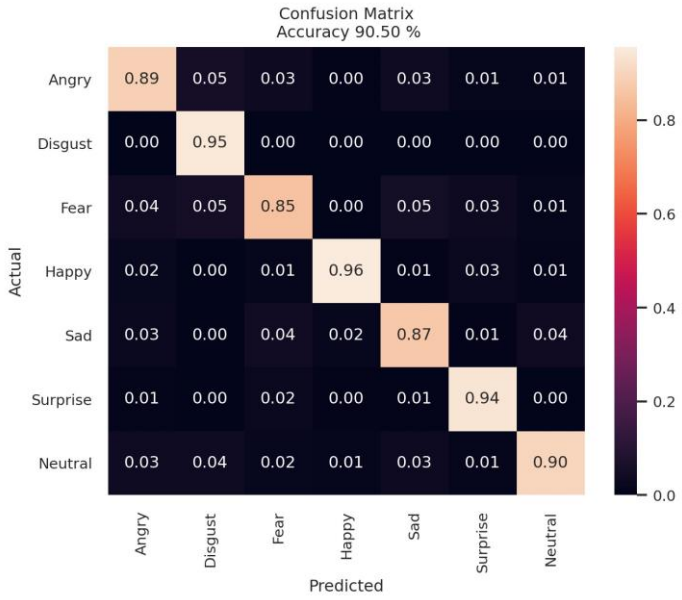
Tabel 6 Hasil Performa model terhadap variasi learning rate

Learning Rate	Akurasi	Presisi	Recalll	F1-Score
0.1	67.12	67.25	63.47	64.64
0.01	77.18	77.96	76.83	77.10
0.001	90.44	90.21	89.89	90.01
0.0005	90.50	89.50	90.69	90.06

Kesimpulan dari pengujian learning rate yang sudah dilakukan yaitu nilai learning rate paling optimal untuk klasifikasi menggunakan EfficientNet adalah 0.0005. Nilai learning rate 0.0005 pada model EfficientNetB0 48x48 terbukti menghasilkan nilai terbaik dengan akurasi sebesar 90.50% dan F1-score 90.06%.

3.3 Hasil klasifikasi ekspresi wajah

Sistem pengenalan ekspresi wajah bekerja untuk meklasifikasikan ekspresi wajah manusia kedalam tujuh emosi dasar. Tujuh emosi tersebut adalah marah, jijik, takut, senang, sedih, terkejut, dan netral. Pengujian dilakukan menggunakan data uji atau PrivateTest yang termasuk dalam dataset FER-2013. Data *testing* berjumlah 3.589 dalam bentuk citra berukuran 48x48 pixel telah dianotasi kedalam tujuh emosi. Setelah diuji, sistem klasifikasi ekspresi wajah berhasil mendapatkan akurasi 90.50%, presisi 89.50%, recall 90.69% dan F1-score mencapai 90.06%. Model mampu melakukan klasifikasi dengan baik yaitu memiliki akurasi diatas 80% [12].



Gambar 8 Confusion Matrik Model Terbaik

Adapun hasil visualisasi *confusion matrix* dalam heatmap disajikan pada gambar 8. Ekspresi senang memiliki akurasi tertinggi mencapai 96% daripada ekspresi lainnya. Sedangkan eksrpesi takut memiliki akurasi terendah yaitu sebesar 85%. Empat dari tujuh emosi dapat diklasifikasikan dengan akurasi diatas 90%. Contoh gambar hasil klasifikasi disajikan dalam gambar 9.

Gambar 9 Hasil Klasifikasi pada citra

3.4 Hasil klasifikasi pada video virtual meeting

Implementasi sistem klasifikasi ekspresi wajah pada virtual meeting menggunakan model terbaik yaitu EfficientNetB0 dengan resoulusi *input* 48x48 dan mode warna RGB. Sistem diterapkan pada video yang diambil dari kanal youtube wolf wacth. Video tersebut berisi tentang reuni aktor MTV Teen Wolf menggunakan menggunakan *virtual meeting*. Peserta virtual meeting terdiri 12 orang terdiri dari delapan laki-laki dan empat perempuan. Semua peserta tidak menggunakan *virtual background*. Pada menit 27:25 sampai 27:35 video dipotong, kemudian dirubah menjadi 240 frame. Setelah itu, setiap frame ekspresi hasil prediksi model dianalisis dengan kejadian sebenarnya. Potongan video tersebut menampilkan keadaan ketika seseorang melakukan tanya jawab kepada peserta lainnya.

Implementasi sistem pada sebuah video telah berhasil melakukan 3.596 prediksi. Hasil dari presiksi yaitu 3.113 *true positive*, 483 *false positive*, dan 483 *false negative*. Pada video tersebut didominasi oleh ekspresi senang dan netral. Ekspresi senang terdiri dari 1.933 *true positive*, 170 *false positive*, dan 70 *false negative*. Netral terdiri dari 1.170 *true positive*, 24 *false positive*, dan 385 *false negative*. Akurasi sistem pada video merupakan rata-rata akurasi dari setiap ekspresi. Dengan demikian dapat dihitung akurasi sistem klasifikasi ekspresi wajah pada video virtual meeting sebesar 96.18%.

4. KESIMPULAN

Berdasarkan hasil penelitian klasifikasi ekpresi wajah menggunakan convolutional neural network dalam keadaan *wild setting* pada *virtual meeting* yang sudah dilakukan, dapat disimpulkan bahwa Sistem dapat mengklasifikan ekspresi wajah pada keadaan wild setting dengan arsitektur EfficientNetB0. Arsitektur EfficientNet adalah arsitektur terbaik dalam melakukan klasifikasi ekspresi wajah pada keadaan wild setting daripada LightCNN dengan akurasi mencapai 90.44%, presisi sebesar 90.21%, recall sebesar 89.89% dan F1-score mencapai 90.01%. Nilai parameter input paling optimal pada arsitektur EfficientNetB0 adalah 48x48 pixel dan nilai parameter learning rate yaitu 0.0005. Pengujian dilakukan pada 3.589 citra dengan akurasi 90.50%, presisi 89.50%, recall 90.69%, dan F1-score 90.06%. Implementasi klasifikasi ekspresi wajah pada video virtual meeting menggunakan metode deteksi wajah MTCNN mendapatkan akurasi mencapai 96.18%.

5. SARAN

Saran yang dapat diterapkan untuk pengembangan penelitian selanjutnya adalah Mengembangkan sistem dengan deteksi wajah yang memiliki frame per second tinggi., Meningkatkan akurasi dengan arsitektur yang mendukung ekspresi wajah secara realtime pada perangkat embeded, Melakukan pelatihan menggunakan data yang diambil secara langsung.

DAFTAR PUSTAKA

[1] A. Mehrabian, "Communication Without Words," *Psychol. Today*, vol. 2, hal. 52–55, 1968.

[2] P. Ekman dan W. V. Friesen, "Constants across cultures in the face and emotion.," *J. Pers. Soc. Psychol.*, vol. 17, no. 2, hal. 124–129, 1971, doi: 10.1037/h0030377.

[3] D. Keltner, A. M. Kring, dan G. A. Bonanno, "Fleeting signs of the course of life: Facial expression and personal adjustment," *Curr. Dir. Psychol. Sci.*, vol. 8, no. 1, hal. 18–22, 1999, doi: 10.1111/1467-8721.00005.

[4] N. Joshi, N. Beecken, H. Bah, F. Steinicke, dan J. Degner, "Advanced Emotion Analytics of Virtual Group Meetings involving Intelligent Virtual Agents," *Proc. - 2022 IEEE Conf. Virtual Real. 3D User Interfaces Abstr. Work. VRW 2022*, hal. 344–350, 2022, doi: 10.1109/VRW55335.2022.00077.

[5] J. Shao dan Y. Qian, "Three convolutional neural network models for facial expression recognition in the wild," *Neurocomputing*, vol. 355, hal. 82–92, 2019, doi: 10.1016/j.neucom.2019.05.005.

[6] M. Tan dan Q. V Le, "EfficientNet : Rethinking Model Scaling for Convolutional Neural Networks," 2019.

[7] Z. Song, "Facial Expression Emotion Recognition Model Integrating Philosophy and Machine Learning Theory," *Front. Psychol.*, vol. 12, no. September, 2021, doi: 10.3389/fpsyg.2021.759485.

[8] A. Mollahosseini, D. Chan, dan M. H. Mahoor, "Going deeper in facial expression recognition using deep neural networks," *2016 IEEE Winter Conf. Appl. Comput. Vision, WACV 2016*, hal. 1–10, 2016, doi: 10.1109/WACV.2016.7477450.

[9] M. Sajjad *dkk.*, "A comprehensive survey on deep facial expression recognition: challenges, applications, and future guidelines," *Alexandria Eng. J.*, vol. 68, hal. 817–840, 2023, doi: 10.1016/j.aej.2023.01.017.

[10] K. Zhang, Z. Zhang, Z. Li, dan Y. Qiao, "Joint Face Detection and Alignment Using Multitask Cascaded Convolutional Networks," *IEEE Signal Process. Lett.*, vol. 23, no. 10, hal. 1499–1503, 2016, doi: 10.1109/LSP.2016.2603342.

[11] W. Watch, "Teen Wolf 9-Years Later | MTV Reunion." 2020, [Daring]. Tersedia pada: https://www.youtube.com/watch?v=q1V1Zu7XYaI.

[12] P. Utami, R. Hartanto, dan I. Soesanti, "The EfficientNet Performance for Facial Expressions Recognition," *2022 5th Int. Semin. Res. Inf. Technol. Intell. Syst.*, no. 1, hal. 756–762, 2023, doi: 10.1109/isriti56927.2022.10053007.