

KLASIFIKASI CITRA WAJAH ASLI DAN SINTETIS UNTUK DETEKSI DEEPFAKE MENGGUNAKAN CONVOLUTIONAL NEURAL NETWORK

Maulana Abdul Fattah, Zuly Budiarto

Teknik Informatika, Universitas Stikubank Semarang

Jl. Kendeng No. 5, Semarang 50233, Indonesia

fattahmaulana51@gmail.com

ABSTRAK

Teknologi *deepfake* berbasis *Generative Adversarial Network* (GAN) berkembang pesat dan menghasilkan citra wajah sintetis yang sulit dibedakan dari citra autentik. Kondisi ini mengancam integritas informasi digital secara serius. Sistem deteksi otomatis yang akurat dan efisien sangat dibutuhkan untuk mengatasi ancaman tersebut. Penelitian ini bertujuan mengembangkan sistem klasifikasi biner untuk membedakan citra wajah autentik dan sintetis berbasis arsitektur *Convolutional Neural Network* (CNN) MobileNetV2 dengan pendekatan *transfer learning* dan *fine-tuning*. Dataset terdiri atas 90.000 citra seimbang: 45.000 citra autentik (*Flickr-Faces-HQ*) dan 45.000 citra sintetis (*StyleGAN*). Pra-pemrosesan mencakup standarisasi dimensi 224 x 224 piksel, normalisasi ke rentang [0, 1], serta augmentasi data melalui *horizontal flip* dan rotasi acak. Pelatihan dilakukan dua tahap: *transfer learning* menghasilkan akurasi validasi 80,49%, kemudian *fine-tuning* meningkatkannya menjadi 91,44%. Evaluasi pada 20.000 data uji menghasilkan akurasi 92,01%, *precision* kelas *fake* 0,95, *recall* kelas *real* 0,95, dan *F1-score* 0,92. Hasil ini membuktikan MobileNetV2 efektif sebagai fondasi sistem forensik digital otomatis untuk deteksi *deepfake*.

Kata kunci : *Deepfake, Convolutional Neural Network, MobileNetV2, Transfer Learning, Klasifikasi Citra, Forensik Digital.*

1. PENDAHULUAN

Perkembangan kecerdasan buatan dalam pengolahan citra digital berkembang sangat pesat dalam satu dekade terakhir. Salah satu produknya adalah teknologi *deepfake*, yaitu manipulasi citra wajah manusia berbasis *deep learning* yang mampu menghasilkan wajah sintetis hiperrealistik. Teknologi ini menggunakan arsitektur *Generative Adversarial Networks* (GAN) yang melatih dua jaringan saraf secara bersamaan, yaitu *generator* yang menciptakan citra palsu dan *diskriminator* yang mengevaluasinya, sehingga menghasilkan citra wajah sintetis yang sangat sulit dibedakan dari citra asli [1].

Penyebaran teknologi *deepfake* membawa dampak serius. Di bidang keamanan siber, citra sintetis mampu menyamarkan anomali pada fitur mikro wajah yang menjadi dasar sistem verifikasi biometrik [2]. Di bidang informasi, *deepfake* digunakan untuk memproduksi konten disinformasi secara massal dan otomatis sehingga sistem deteksi berbasis pengamatan manusia terbukti tidak memadai [3]. Keterbatasan ini menciptakan celah serius dalam forensik digital yang belum terjawab secara optimal oleh metode konvensional.

Beberapa penelitian sebelumnya telah berupaya menjawab tantangan ini. CNN terbukti mampu mengklasifikasikan keaslian citra secara akurat melalui deteksi diskontinuitas piksel dan artefak mikroskopis, sekaligus sensitif terhadap anomali tekstur biometrik pada citra sintetis [1], [2]. Integrasi MobileNetV2 bersama MTCNN untuk keperluan waktu nyata terbukti andal dalam mitigasi disinformasi [3].

Penggabungan MobileNetV2 dengan *Support Vector Machine* berhasil menyeimbangkan akurasi klasifikasi dengan efisiensi komputasi [5]. Rancangan blok konvolusi optimal terbukti meningkatkan akurasi deteksi citra sintetis secara konsisten [6]. Perbandingan berbagai arsitektur *deep learning* membuktikan ketahanan model terhadap gangguan piksel pada dataset [7]. Strategi *transfer learning* berhasil mengadaptasi model prelatih terhadap citra buatan *StyleGAN* [8].

Teknik penandaan air pada arsitektur DenseNet121 juga terbukti efektif untuk validasi integritas citra [9]. Namun demikian, mayoritas penelitian tersebut belum mengevaluasi secara sistematis efektivitas dua tahap pelatihan, yaitu *transfer learning* dilanjutkan *fine-tuning*, pada dataset berskala besar dengan komposisi seimbang.

Berdasarkan kesenjangan tersebut, penelitian ini bertujuan mengembangkan sistem klasifikasi biner untuk membedakan citra wajah autentik dan sintetis, sekaligus mengukur sejauh mana MobileNetV2 mampu mendeteksi artefak visual pada citra *deepfake* berbasis *StyleGAN*. Arsitektur MobileNetV2 dipilih karena mekanisme *depthwise separable convolutions* yang dimilikinya mampu mempercepat pemrosesan tanpa menurunkan akurasi secara signifikan, sehingga cocok untuk lingkungan komputasi berspesifikasi terbatas [5].

Untuk menjawab tujuan tersebut, penelitian ini membangun sistem klasifikasi menggunakan dataset *Real and Fake Faces* sebanyak 90.000 citra seimbang, terdiri atas 45.000 citra autentik (*Flickr-Faces-HQ*) dan 45.000 citra sintetis (*StyleGAN*). Pelatihan dilakukan dalam dua tahap: *transfer learning* untuk

mengadaptasi bobot pralatih ImageNet, dilanjutkan *fine-tuning* untuk mengoptimalkan seluruh lapisan model terhadap data target. Performa model dievaluasi secara komprehensif menggunakan metrik *accuracy*, *precision*, *recall*, *F1-score*, dan *confusion matrix*.

2. TINJAUAN PUSTAKA

2.1. Pengolahan Citra Digital

Pengolahan citra digital merupakan cabang ilmu komputasi yang berfokus pada analisis, manipulasi, dan interpretasi citra menggunakan algoritma berbasis komputer. Dalam representasi digital, citra dimodelkan sebagai matriks piksel yang menyimpan informasi intensitas warna dan kecerahan. Proses pengolahan citra bertujuan untuk meningkatkan kualitas visual, mengekstraksi informasi esensial, serta memungkinkan sistem komputer memahami pola visual yang terkandung di dalamnya [10].

Alur pengolahan citra umumnya mencakup empat tahapan utama: akuisisi, pra-pemrosesan, ekstraksi fitur, dan klasifikasi. Tahap pra-pemrosesan meliputi normalisasi, penyesuaian dimensi, dan augmentasi yang bertujuan meningkatkan kualitas data masukan. Integritas tahapan ekstraksi fitur visual sangat krusial dalam membangun sistem forensik media digital yang andal, terutama dalam skenario di mana manipulasi citra dilakukan pada tingkat piksel yang sangat halus [11].

Perkembangan teknologi kecerdasan buatan, khususnya *deep learning*, telah merevolusi pendekatan pengolahan citra digital dengan memungkinkan sistem mempelajari representasi fitur secara otomatis dan hierarkis dari data mentah. Integrasi arsitektur *deep learning* yang canggih, seperti kombinasi *Vision Transformer* dengan CNN, dapat meningkatkan kemampuan sistem dalam menganalisis artefak manipulasi citra yang kompleks, termasuk pada konten *deepfake* video [12]. Hal ini menjadikan pendekatan *deep learning* sebagai standar modern dalam berbagai aplikasi analisis citra forensik.

2.2. Deepfake dan Generative Adversarial Network

Deepfake merupakan teknologi manipulasi media digital berbasis kecerdasan buatan yang mampu menghasilkan konten visual atau audio sintetis yang merepresentasikan objek nyata secara meyakinkan. Sebagian besar sistem *deepfake* modern mengadopsi arsitektur *Generative Adversarial Network* (GAN) yang terdiri atas dua komponen saling berkompetisi: *generator* yang bertugas memproduksi citra sintetis, dan *discriminator* yang bertugas membedakan citra asli dari citra buatan. Melalui mekanisme pelatihan adversarial yang berulang, generator secara bertahap menghasilkan citra dengan tingkat realisme yang semakin meningkat [10].

Algoritma StyleGAN, yang digunakan sebagai sumber citra sintetis pada penelitian ini, merupakan varian GAN mutakhir yang mampu menghasilkan citra wajah dengan kualitas sangat tinggi melalui kendali

gaya (*style control*) pada berbagai tingkat resolusi. Karakteristik ini menjadikan citra hasil StyleGAN sangat sulit dibedakan secara visual dari citra wajah autentik, sehingga menciptakan tantangan signifikan bagi sistem deteksi otomatis [13].

2.3. Convolutional Neural Network (CNN)

Convolutional Neural Network (CNN) merupakan arsitektur jaringan saraf tiruan yang dirancang secara khusus untuk mengolah data berbentuk kisi, termasuk citra dua dimensi. Struktur dasar CNN meliputi *convolution layer*, *activation layer*, *pooling layer*, dan *fully connected layer*. *Convolution layer* mengekstraksi fitur visual melalui operasi konvolusi antara citra masukan dan filter atau *kernel* yang dapat dipelajari, sedangkan *pooling layer* mereduksi dimensi *feature map* guna menurunkan kompleksitas komputasi sambil mempertahankan informasi esensial [16].

Kemampuan CNN dalam mengidentifikasi hierarki fitur visual, mulai dari tepi dan tekstur pada lapisan awal hingga pola semantik kompleks pada lapisan dalam, menjadikannya metode yang dominan dalam klasifikasi citra. Afandi dan Purnama [17] membuktikan bahwa arsitektur berbasis CNN, khususnya yang dikombinasikan dengan modul sekuensial seperti *Gated Recurrent Unit* (GRU), mampu mengeksplorasi fitur temporal dan spasial secara simultan dalam skenario deteksi *deepfake* video. Dalam konteks deteksi *deepfake* berbasis citra diam, CNN mampu mengidentifikasi artefak visual khas yang sering muncul pada citra hasil manipulasi, seperti inkonsistensi tekstur wajah, distorsi piksel lokal, dan ketidakseimbangan pencahayaan yang diinduksi oleh algoritma generatif [14].

2.4. Arsitektur MobileNetV2

MobileNetV2 merupakan arsitektur CNN yang dioptimalkan untuk efisiensi komputasi tinggi dengan kapasitas memori minimal, sehingga sangat sesuai untuk deployment pada perangkat dengan sumber daya terbatas. Inovasi utama MobileNetV2 terletak pada dua komponen desain: *depthwise separable convolution* dan *inverted residual block*. Teknik *depthwise separable convolution* memisahkan operasi konvolusi spasial dari konvolusi kanalnya, sehingga secara dramatis mengurangi jumlah operasi matematis dibandingkan konvolusi standar [18].

Blok *inverted residual* mengoperasikan transformasi pada ruang fitur berukuran lebih besar melalui lapisan ekspansi, kemudian mengompresnya kembali melalui lapisan proyeksi, dengan koneksi residual yang menghubungkan lapisan input berukuran sempit secara langsung ke output. Mekanisme ini memungkinkan jaringan mempertahankan informasi penting selama propagasi, sehingga kinerja klasifikasi tetap terjaga meskipun jumlah parameter relatif kecil [18].

2.5. Transfer Learning

Transfer learning merupakan paradigma pembelajaran mesin di mana representasi pengetahuan yang telah diperoleh dari pelatihan pada dataset sumber skala besar dimanfaatkan kembali untuk menyelesaikan tugas baru pada domain yang berbeda namun berkaitan. Pendekatan ini sangat efektif ketika data pelatihan pada domain target bersifat terbatas, karena model dapat memanfaatkan fitur visual dasar seperti deteksi tepi, tekstur, dan pola bentuk yang telah dipelajari dari dataset besar seperti ImageNet [19].

Terdapat dua strategi utama dalam penerapan *transfer learning*: (1) *feature extraction*, yaitu membekukan seluruh parameter model pra-terlatih dan hanya melatih lapisan klasifikasi yang baru ditambahkan; dan (2) *fine-tuning*, yaitu mengaktifkan kembali sebagian lapisan dari model pra-terlatih untuk dilatih ulang dengan laju pembelajaran yang lebih kecil, memungkinkan adaptasi bobot terhadap karakteristik spesifik dataset target [15].

2.6. Penelitian Terdahulu

Beberapa penelitian terdahulu telah mengkaji deteksi *deepfake* menggunakan berbagai pendekatan *deep learning*. CNN terbukti mampu mengklasifikasikan keaslian citra secara akurat melalui deteksi diskontinuitas piksel dan artefak mikroskopis pada citra wajah sintetis [1].

Penelitian lain mengonfirmasi bahwa metode berbasis CNN juga sensitif terhadap anomali tekstur biometrik yang muncul pada citra hasil generasi GAN [2].

Pada pendekatan berbasis arsitektur ringan, integrasi MobileNetV2 bersama MTCNN terbukti andal untuk deteksi *deepfake* secara waktu nyata dalam skenario mitigasi disinformasi [3]. Penggabungan MobileNetV2 dengan *Support Vector Machine* juga berhasil menyeimbangkan akurasi klasifikasi dengan efisiensi komputasi pada lingkungan berspesifikasi terbatas [5].

Rancangan blok konvolusi optimal pada arsitektur CNN terbukti meningkatkan akurasi deteksi citra sintetis secara konsisten [6]. Perbandingan berbagai arsitektur *deep learning* membuktikan ketahanan model terhadap gangguan piksel pada dataset citra wajah [7]. Strategi *transfer learning* berhasil mengadaptasi model pralatih terhadap citra buatan *StyleGAN* dengan akurasi tinggi [8]. Teknik penandaan air pada arsitektur DenseNet121 juga terbukti efektif untuk validasi integritas citra digital [9].

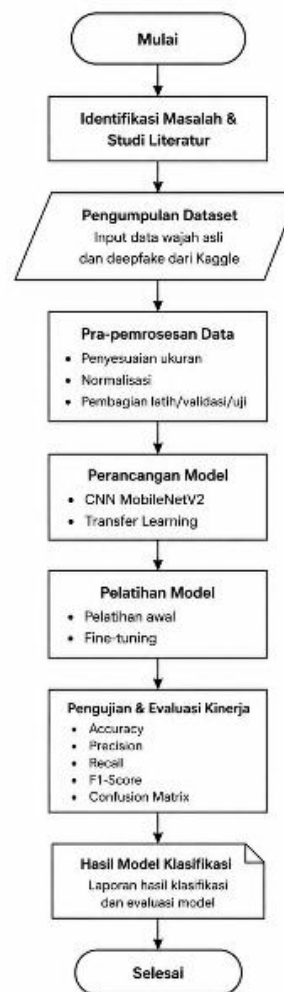
Berdasarkan kajian tersebut, mayoritas penelitian terdahulu hanya menerapkan satu tahap pelatihan tanpa mengombinasikan *transfer learning* dan *fine-tuning* secara sistematis. Evaluasi pada dataset berskala besar dengan komposisi kelas yang seimbang juga masih jarang dilakukan. Penelitian ini mengisi kesenjangan tersebut dengan menerapkan strategi dua tahap pelatihan secara berurutan pada dataset 90.000

citra dengan komposisi seimbang antara kelas autentik dan sintetis.

3. METODE PENELITIAN

3.1. Desain dan Alur Penelitian

Penelitian ini mengadopsi desain eksperimen komputasional kuantitatif yang bertujuan untuk membangun, melatih, dan mengevaluasi model klasifikasi citra wajah berbasis *deep learning*. Alur penelitian yang diterapkan mencakup tahapan-tahapan berikut secara berurutan: identifikasi permasalahan, kajian literatur, pengumpulan dataset, prapemrosesan data, perancangan arsitektur model, pelatihan model, evaluasi kinerja model, serta analisis dan pembahasan hasil. Secara visual, tahapan-tahapan sistematis tersebut direpresentasikan pada Gambar 1.



Gambar 1. Alur Penelitian Klasifikasi Citra *Deepfake*

3.2. Dataset Penelitian

Penelitian ini menggunakan dataset *140K Real and Fake Faces* yang tersedia secara publik melalui platform Kaggle. Dataset ini terdiri atas 90.000 citra wajah yang terbagi secara proporsional ke dalam dua kategori kelas, yaitu 45.000 citra wajah autentik yang bersumber dari dataset *Flickr-Faces-HQ* (FFHQ) dan 45.000 citra wajah sintetis yang dihasilkan oleh

algoritma *StyleGAN*. Distribusi kelas yang seimbang (*balanced dataset*) dengan rasio 1:1 antara kelas *real* dan kelas *fake* diterapkan untuk meminimalkan bias model terhadap kelas mayoritas dalam proses pelatihan.

Dataset selanjutnya dipartisi ke dalam tiga subset dengan fungsi yang berbeda, sebagaimana disajikan pada Tabel 1.

Tabel 1. Distribusi Partisi Dataset Penelitian

Subset Data	Jumlah Citra	Komposisi (Real : Fake)
Training Set	60.000	30.000 : 30.000
Validation Set	10.000	5.000 : 5.000
Testing Set	20.000	10.000 : 10.000
Total	90.000	45.000 : 45.000

3.3. Prapemrosesan Data

Tahap prapemrosesan data merupakan proses kritis dalam mempersiapkan dataset agar sesuai dengan spesifikasi arsitektur *MobileNetV2* dan mengoptimalkan kualitas data masukan. Implementasi prapemrosesan dilakukan menggunakan *framework TensorFlow* dengan memanfaatkan *TensorFlow Dataset API*.

- Standardisasi Dimensi Citra: Seluruh citra dalam dataset diubah ukurannya menjadi 224×224 piksel dengan tiga kanal warna (RGB), yang merupakan dimensi *input* standar arsitektur *MobileNetV2*. Proses standardisasi ini dieksekusi secara otomatis melalui fungsi `image_dataset_from_directory` pada *TensorFlow*.
- Normalisasi Nilai Piksel: Nilai intensitas piksel yang semula berada dalam rentang $[0, 255]$ dinormalisasi ke rentang $[0, 1]$ menggunakan lapisan *Rescaling* pada *TensorFlow*. Operasi normalisasi ini dapat diformulasikan sebagai:

$$x_{norm} = \frac{x}{255}$$

- di mana x merupakan nilai piksel orisinal dan x_{norm} merupakan nilai piksel yang telah ternormalisasi. Normalisasi ini bertujuan untuk meningkatkan stabilitas proses pelatihan dan mempercepat konvergensi optimasi parameter jaringan.
- Augmentasi Data: Augmentasi data diterapkan secara eksklusif pada *training set* untuk meningkatkan invariansi model dan meminimalkan risiko *overfitting*. Teknik augmentasi yang diimplementasikan meliputi (a) *Horizontal Flip*: pembalikan citra secara horizontal untuk mensimulasikan variasi orientasi wajah, yang divalidasi tidak mengubah karakteristik biologis wajah; dan (b) *Random Rotation*: rotasi citra dengan sudut kecil secara acak untuk mengakomodasi variasi posisi kepala yang wajar. Teknik augmentasi seperti penyesuaian kecerahan dan *zoom* secara sengaja dihindari, mengingat artefak kecerahan dan detail tepi wajah merupakan indikator penting yang digunakan model dalam membedakan citra autentik dari citra sintetis.

- Konfigurasi Pipeline Dataset: Dataset dikonfigurasi dalam *batch* berukuran 32 citra menggunakan mekanisme *parallel data loading* dan *prefetching* untuk mengoptimalkan pemanfaatan sumber daya komputasi GPU selama pelatihan. Teknik *mixed precision training* yang mengombinasikan tipe data *float16* dan *float32* juga diterapkan untuk meningkatkan efisiensi memori dan mempercepat komputasi.

3.4. Arsitektur Model

Model klasifikasi yang dikembangkan mengintegrasikan arsitektur *MobileNetV2* pra-train (*pretrained weights* dari ImageNet) sebagai *feature extractor* yang dipadukan dengan lapisan klasifikasi kustom. Secara struktural, aliran data pada konfigurasi model ini dimulai dengan melewati *Data Augmentation Layer* yang bertugas menerapkan transformasi acak pada citra masukan. Setelah itu, tensor citra diteruskan ke dalam *MobileNetV2 Base Model* dengan konfigurasi `include_top=False` yang bertindak sebagai pengekstraksi fitur hierarkis, di mana parameter pada bagian ini dibekukan pada tahap awal pelatihan. Peta fitur yang dihasilkan kemudian direduksi dimensinya menggunakan *Global Average Pooling Layer* melalui operasi rata-rata spasial global. Representasi fitur yang telah dipadatkan tersebut selanjutnya diproses oleh sebuah *Dense Layer* yang terdiri atas 128 unit dengan fungsi aktivasi ReLU. Untuk meminimalkan risiko *overfitting*, jaringan ini disematkan *Dropout Layer* dengan probabilitas penonaktifan unit acak sebesar 0,3. Sebagai tahap akhir konvolusi, proses penentuan kelas diselesaikan oleh *Output Layer* berkapasitas 1 unit dengan fungsi aktivasi sigmoid guna menghasilkan nilai probabilitas keanggotaan citra dalam rentang $(0, 1)$.

3.5. Prosedur Pelatihan

Pelatihan model dilaksanakan melalui dua tahapan terstruktur. Pada Tahap 1 (*Transfer Learning*), seluruh lapisan *MobileNetV2* dibekukan, sehingga hanya lapisan klasifikasi kustom yang diperbarui bobot parameternya. Tahap ini dilaksanakan selama 10 *epoch* dengan tujuan melatih representasi fitur tingkat tinggi yang spesifik terhadap dataset penelitian. Pada Tahap 2 (*Fine Tuning*), pembekuan lapisan dicabut secara selektif pada lapisan-lapisan bagian atas jaringan, memungkinkan penyesuaian bertahap fitur yang telah dipelajari terhadap karakteristik dataset. Tahap ini menggunakan laju pembelajaran yang lebih kecil untuk mencegah perubahan bobot yang terlalu drastis, yang berpotensi menghilangkan representasi fitur yang telah terbentuk.

3.6. Metrik Evaluasi

Kinerja model dievaluasi menggunakan serangkaian metrik yang komprehensif. Akurasi (*Accuracy*) mengukur proporsi prediksi yang benar terhadap keseluruhan data uji, dirumuskan sebagai:

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN}$$

Presisi (*Precision*) mengukur ketepatan prediksi positif model:

$$Precision = \frac{TP}{TP + FP}$$

Recall mengukur kemampuan model dalam mendeteksi seluruh instansi kelas positif:

$$Recall = \frac{TP}{TP + FN}$$

F1-score merupakan rata-rata harmonik dari presisi dan *recall*:

$$F1 = 2 \times \frac{Precision \times Recall}{Precision + Recall}$$

Analisis *confusion matrix* juga dilakukan untuk mengevaluasi distribusi kesalahan klasifikasi pada masing-masing kelas. Dalam formulasi tersebut, *TP* (*True Positive*) merepresentasikan jumlah citra *fake* yang diprediksi benar, *TN* (*True Negative*) merepresentasikan jumlah citra *real* yang diprediksi benar, *FP* (*False Positive*) merupakan jumlah citra *real* yang keliru diprediksi sebagai *fake*, dan *FN* (*False Negative*) merupakan jumlah citra *fake* yang keliru diprediksi sebagai *real*.

4. HASIL DAN PEMBAHASAN

4.1. Karakteristik dan Pra-Pemrosesan Dataset

Eksperimen komputasional pada penelitian ini memanfaatkan 90.000 citra wajah beresolusi tinggi yang terbagi secara proporsional (rasio 1:1) ke dalam kelas *real* dan *fake*. Data asli diekstraksi dari korpus *Flickr-Faces-HQ* (FFHQ), sementara data sintesis diproduksi melalui arsitektur *Generative Adversarial Network* (GAN) varian *StyleGAN*. Pendekatan distribusi seimbang ini diterapkan secara ketat guna mengeliminasi potensi bias klasifikasi terhadap kelas mayoritas selama fase pelatihan.

Tabel 2. Sampel visual dataset wajah autentik dan sintesis.



Guna mengoptimalkan kinerja arsitektur MobileNetV2, seluruh dataset diproses melalui *pipeline* TensorFlow dengan menyeragamkan dimensi spasial menjadi 224×224 piksel. Selain itu, nilai intensitas piksel direduksi ke dalam rentang 0 hingga

1 melalui teknik normalisasi untuk menjamin stabilitas konvergensi model. Guna menekan risiko *overfitting*, variasi data pelatihan diperluas melalui augmentasi rotasi acak dan rotasi horizontal (*horizontal flip*). Konfigurasi *batch size* ditetapkan sebanyak 32 citra, dan dataset dipartisi menjadi 60.000 citra latih, 10.000 citra validasi, serta 20.000 citra uji.

4.2. Implementasi dan Pelatihan Model

Model pendeteksi *deepfake* dibangun dengan memanfaatkan ekstraktor fitur MobileNetV2 berbasis bobot pralatih ImageNet, yang dimodifikasi dengan penambahan *Global Average Pooling*, lapisan *Dense* (128 neuron, aktivasi ReLU), lapisan *Dropout* (0,3), dan lapisan *Output* beraktivasi Sigmoid. Efisiensi komputasi ditingkatkan melalui penerapan *mixed precision training* yang mengombinasikan tipe data *float16* dan *float32*.

Fase pembelajaran mesin dieksekusi dalam dua tahapan komprehensif. Pada fase pertama (*Transfer Learning*), parameter jaringan dasar dibekukan, dan pelatihan dikhususkan pada lapisan klasifikasi selama 10 *epoch*. Hasilnya, model mencapai akurasi validasi puncak sebesar 80,49%.

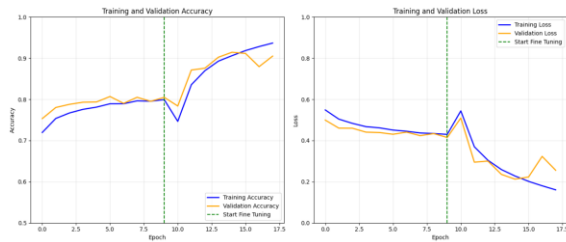
Tabel 3. Rekapitulasi Akurasi dan Loss pada Fase Transfer Learning

Epoch	Training Accuracy	Validation Accuracy
1	0.6942	0.7533
2	0.7511	0.7802
3	0.7674	0.7879
4	0.7784	0.7931
5	0.7832	0.7937
6	0.7928	0.8070
7	0.7896	0.7903
8	0.7997	0.8051
9	0.7971	0.7952
10	0.8033	0.8049

Fase kedua berfokus pada *fine-tuning*, di mana sebagian lapisan diekstraksi ulang dengan *learning rate* yang direduksi untuk penyesuaian bobot yang lebih presisi. Tahapan ini sukses mengeskalasi performa pengenalan pola, ditandai dengan lonjakan akurasi validasi yang signifikan mencapai 91,44%.

Tabel 4. Rekapitulasi Akurasi dan Loss pada Fase Fine Tuning

Epoch	Training Accuracy	Validation Accuracy
1	0.6883	0.7839
2	0.8262	0.8712
3	0.8635	0.8757
4	0.8892	0.9023
5	0.9043	0.9144
6	0.9173	0.9116
7	0.9274	0.8794
8	0.9353	0.9050



Gambar 2. Kurva Training & Validation (Accuracy dan Loss)

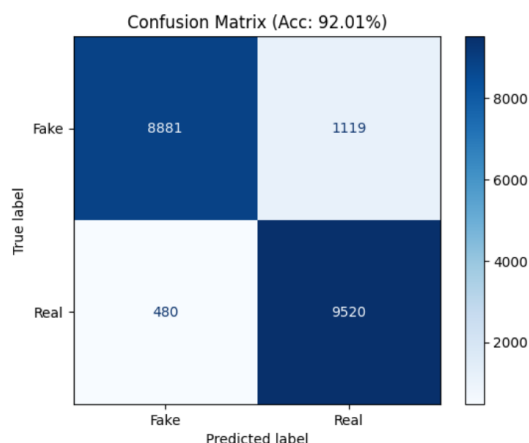
4.3. Evaluasi Kinerja dan Analisis Misklasifikasi

Kemampuan generalisasi arsitektur divalidasi menggunakan 20.000 citra uji yang sepenuhnya terisolasi dari fase pelatihan. Secara keseluruhan, klasifikator MobileNetV2 berhasil mencatatkan metrik akurasi komprehensif sebesar 92,01%.

Tabel 5. Hasil Evaluasi Model Berdasarkan Classification Report

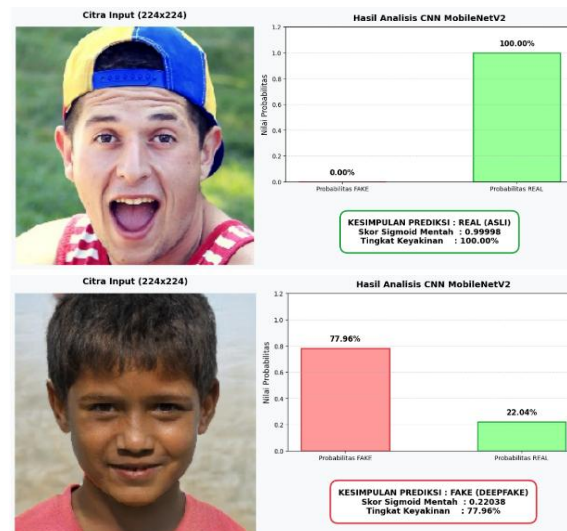
Kelas	Precision	Recall	F1-Score	Support
Fake	0.95	0.89	0.92	10.000
Real	0.89	0.95	0.92	10.000
Accuracy			0.92	20.000

Tinjauan lebih mendalam terhadap *classification report* menunjukkan keseimbangan kinerja pada metrik *F1-Score* (0,92 untuk kedua kelas). Namun, kelas *fake* menghasilkan *precision* yang sangat tinggi (0,95) diiringi *recall* yang lebih rendah (0,89) dibandingkan kelas *real*. Fenomena ini tervalidasi melalui matriks kebingungan (*confusion matrix*), di mana model menghasilkan 1.119 *False Negatives* (citra *fake* yang dianggap *real*) berbanding 480 *False Positives*.



Gambar 3. Confusion Matrix Hasil Pengujian Model

Kemampuan model dalam memprediksi data baru juga divalidasi melalui pengujian inferensi interaktif. Model mampu memberikan probabilitas keluaran (*sigmoid*) yang sangat meyakinkan untuk citra uji yang terklasifikasi dengan benar, baik pada citra asli maupun citra sintetis, seperti yang ditunjukkan pada Gambar 4.



Gambar 4. Hasil pengujian inferensi pada citra asli (atas) dan citra sintetis (bawah)

Tingginya angka *False Negative* secara teknis diakibatkan oleh kemampuan superior *StyleGAN* dalam mereproduksi makro-spasial wajah secara natural. Ketika gambar mengalami penyusutan resolusi (menjadi 224×224 piksel) selama pra-pemrosesan, artefak mikroskopis frekuensi tinggi yang biasanya menjadi penanda *deepfake* ikut tereduksi, sehingga arsitektur konvolusional menafsirkan kehalusan tekstur tersebut sebagai data autentik.

Sebaliknya, dari hasil evaluasi visual yang dilakukan, ditemukan bahwa kegagalan model dalam mengenali wajah manusia asli (*False Positives*) mayoritas dipicu oleh lingkungan visual yang ekstrem. Pencahayaan yang menghasilkan bayangan kontras atau lensa yang kurang fokus (*blur*) menyembunyikan pori-pori kulit alami subjek, yang merupakan fitur penentu utama bagi model dalam memvalidasi keaslian citra.

5. KESIMPULAN DAN SARAN

Penelitian ini berhasil mengembangkan dan memvalidasi sistem klasifikasi citra wajah autentik dan sintetis berbasis arsitektur MobileNetV2 melalui pendekatan *transfer learning* dan *fine-tuning*. Strategi pelatihan dua tahap ini terbukti efektif mengoptimalkan kinerja model, ditandai dengan peningkatan akurasi validasi dari 80,49% pada tahap awal menjadi 91,44% pasca-*fine-tuning*. Pada evaluasi akhir terhadap 20.000 citra uji, model mencatatkan akurasi keseluruhan sebesar 92,01% dengan *F1-score* 0,92 yang seimbang pada kedua kelas (*precision fake* 0,95; *recall real* 0,95), mengindikasikan ketiadaan bias klasifikasi yang signifikan. Meskipun berkinerja tinggi, analisis *confusion matrix* mengungkap kerentanan model terhadap *False Negative* (1.119 citra sintetis terklasifikasi autentik) dibandingkan *False Positive* (480 citra), yang sangat berkorelasi dengan tingginya presisi algoritma *StyleGAN* serta hilangnya informasi frekuensi tinggi akibat reduksi resolusi

(downsampling). Keterbatasan lainnya terletak pada generalisasi lintas domain akibat pelatihan yang eksklusif pada citra StyleGAN. Oleh karena itu, penelitian lanjutan direkomendasikan untuk memperluas cakupan dataset dengan mesin generatif lain (seperti *Stable Diffusion* dan *Midjourney*), mengeksplorasi arsitektur mutakhir seperti *EfficientNetV2* atau hibrida *Vision Transformer* (ViT), mengintegrasikan analisis domain frekuensi (transformasi Fourier atau wavelet) guna mempertahankan artefak mikroskopis, serta menerapkan pendekatan *Explainable AI* (Grad-CAM) untuk meningkatkan transparansi keputusan model.

DAFTAR PUSTAKA

- [1] S. A. Gulo, A. A. Pertiwi, S. Putri, S. Nasution, and H. Syahputra, "Deteksi deepfake dalam citra menggunakan convolutional neural network (CNN)," *Jurnal Mahasiswa Teknik Informatika*, vol. 9, no. 5, 2025.
- [2] J. Mu, M. Adrezo, and A. N. Haikal, "Identifikasi wajah asli dan buatan deepfake menggunakan metode convolutional neural network," *Teknika*, vol. 13, no. 1, pp. 45–50, Jan. 2024, doi: 10.34148/teknika.v13i1.705.
- [3] H. A. Salim, A. Muliawan, and D. A. Fauziah, "Deteksi deepfake dalam mitigasi risiko disinformasi digital menggunakan metode multi-task cascaded convolutional network dan MobileNetV2 di Indonesia," *semanTIK*, vol. 11, no. 2, pp. 194–200, Jul.–Dec. 2025.
- [4] R. F. Taufiqurrahman, F. T. Anggraeny, and A. N. Sihananto, "Penerapan arsitektur CNN-dilated untuk deteksi manipulasi citra deepfake," *ILKOMNIKA: Journal of Computer Science and Applied Informatics E*, vol. 7, no. 3, pp. 495–506, 2025, doi: 10.28926/ilkomnika.v7i3.821.
- [5] B. D. Satoto, A. Yasid, A. Rahmadhani, B. K. Khotimah, B. Irmawati, and S. B. Hamid, "Deepfake detection using MobileNetV2 integrated with a support vector machine classifier," *EPJ Web Conf.*, vol. 344, p. 01019, 2025, doi: 10.1051/epjconf/202534401019.
- [6] A. Raza, K. Munir, and M. Almutairi, "A novel deep learning approach for deepfake image detection," *Applied Sciences*, vol. 12, no. 19, p. 9820, 2022, doi: 10.3390/app12199820.
- [7] A. F. Akbar, P. D. W. Ayu, and D. P. Hostiadi, "Performance analysis of deep learning architectures in classifying fake and real images," *JUITA: Jurnal Informatika*, vol. 13, no. 2, pp. 167–176, Jul. 2025.
- [8] J. Sharma, S. Sharma, V. Kumar, H. S. Hussein, and H. Alshazly, "Deepfakes classification of faces using convolutional neural networks," *Traitement du Signal*, vol. 39, no. 3, pp. 1027–1037, Jun. 2022, doi: 10.18280/ts.390330.
- [9] F. A. Jiwani and B. S. W. Poetro, "Sistem deteksi gambar deepfake menggunakan CNN DenseNet-121 dengan watermarking least significant bit (LSB)," *Jurnal Rekayasa Sistem Informasi dan Teknologi*, vol. 2, 2025.
- [10] R. Tolosana, R. Vera-Rodriguez, J. Fierrez, A. Morales, and J. Ortega-Garcia, "DeepFakes and beyond: A survey of face manipulation and fake detection," *Information Fusion*, vol. 64, pp. 131–148, 2020, doi: 10.1016/j.inffus.2020.06.014.
- [11] L. Verdoliva, "Media forensics and DeepFakes: An overview," *IEEE Journal of Selected Topics in Signal Processing*, vol. 14, no. 5, pp. 910–932, Aug. 2020, doi: 10.1109/JSTSP.2020.3002101.
- [12] D. Coccomini, N. Messina, C. Gennaro, and F. Falchi, "Combining EfficientNet and vision transformers for video deepfake detection," in *Image Analysis and Processing – ICIAP 2021*, 2022, pp. 219–229, doi: 10.1007/978-3-031-06433-3_19.
- [13] Y. Mirsky and W. Lee, "The creation and detection of deepfakes," *ACM Computing Surveys*, vol. 54, no. 1, pp. 1–41, Jan. 2021, doi: 10.1145/3425780.
- [14] I. U. W. Mulyono, E. H. Rachmawanto, C. A. Sari, and M. K. Sarker, "A high accuracy of deep learning based CNN architecture: Classic, VGGNet, and ResNet50 for Covid-19 image classification," *TELKOMNIKA (Telecommunication Computing Electronics and Control)*, vol. 22, no. 5, pp. 1187–1195, Oct. 2024, doi: 10.12928/TELKOMNIKA.v22i5.26017.
- [15] J.-Y. Ong, L.-Y. Ong, and M.-C. Leow, "Addressing overfitting in comparative study for deep learning-based classification," *TELKOMNIKA (Telecommunication Computing Electronics and Control)*, vol. 23, no. 3, p. 673, Jun. 2025, doi: 10.12928/telkomnika.v23i3.26451.
- [16] M. Syarief and W. Setiawan, "Convolutional neural network for maize leaf disease image classification," *TELKOMNIKA (Telecommunication Computing Electronics and Control)*, vol. 18, no. 3, pp. 1376–1381, 2020, doi: 10.12928/TELKOMNIKA.v18i3.14840.
- [17] R. Afandi and B. Purnama, "Detecting deepfake videos using CNN and GRU methods: Evaluating performance on the Celeb-DF(v2) dataset," *Jurnal Sistem Komputer dan Informatika (JSON)*, vol. 7, no. 2, pp. 448–458, Dec. 2025, doi: 10.30865/json.v7i2.9372.
- [18] M. Amen and M. L. Ranam, "Lightweight deepfake detection on mobile devices using attention-enhanced MobileNet and frequency domain analysis," *Journal of Technology Informatics and Engineering*, vol. 4, no. 1, pp. 95–114, Apr. 2025, doi: 10.51903/jtie.v4i1.275.
- [19] F. Zhuang et al., "A comprehensive survey on transfer learning," *Proceedings of the IEEE*, vol. 109, no. 1, pp. 43–76, Jan. 2021, doi: 10.1109/JPROC.2020.3004555.