EVALUASI INTEGRASI YOLOV11N SEBAGAI PRAPROSES DETEKSI WAJAH UNTUK KLASIFIKASI DEEPFAKE BERBASIS MOBILEVIT

Habib Nurrohmad Sugiharto[1*], Faisal Muttaqin[2], Budi Mukhamad Mulyo[3]
[1,2,3]Program Studi Informatika, UPN "Veteran" Jawa Timur, Surabaya, Indonesia

**Penulis korespondensi: 22081010165@student.upnjatim.ac.id*

ABSTRAK

Perkembangan teknologi *deepfake* menyebabkan media manipulatif semakin sulit dibedakan dari media asli, terutama karena manipulasi umumnya terjadi pada area wajah. Oleh karena itu, tahap praproses yang mampu mengarahkan model pada area wajah perlu dievaluasi untuk mengetahui pengaruhnya terhadap performa klasifikasi. Penelitian ini bertujuan untuk mengevaluasi integrasi YOLOv11n sebagai detektor wajah pada tahap praproses untuk mendukung klasifikasi *deepfake* berbasis MobileViT. Metode yang digunakan meliputi pelatihan YOLOv11n menggunakan dataset deteksi wajah/kepala beranotasi *bounding box*, penerapan YOLOv11n untuk mendeteksi dan memotong area wajah, serta klasifikasi citra asli dan citra *deepfake* menggunakan MobileViT. Hasil pelatihan YOLOv11n menunjukkan performa terbaik pada *epoch* ke-40 dengan presisi 93,36%, *recall* 90,72%, *mAP@50* 96,41%, dan *mAP@50:95* 58,80%. Kurva *classification loss* YOLOv11n juga menunjukkan tren penurunan pada data latih dan validasi hingga akhir pelatihan, sehingga mengindikasikan proses pembelajaran detektor berjalan stabil. Pada tahap klasifikasi, MobileViT tanpa YOLOv11n memperoleh akurasi 81,73%, presisi 79,10%, *recall* 86,27%, dan *F1-score* 82,53%. Sementara itu, MobileViT dengan praproses YOLOv11n memperoleh akurasi 95,53%, presisi 95,96%, *recall* 95,07%, dan *F1-score* 95,51%. Integrasi YOLOv11n meningkatkan akurasi sebesar 13,80 poin persentase dan *F1-score* sebesar 12,98 poin persentase. Hasil ini menunjukkan bahwa pemfokusan area wajah melalui YOLOv11n efektif dalam meningkatkan performa klasifikasi *deepfake* berbasis MobileViT.

Kata kunci: YOLOv11n, MobileViT, *deepfake*, deteksi wajah, praproses

1 PENDAHULUAN

Perkembangan *artificial intelligence* mendorong kemunculan media sintetis yang semakin realistis, salah satunya *deepfake*. *Deepfake* dapat memanipulasi wajah seseorang pada citra atau video sehingga tampak menyerupai media asli. Kondisi ini menimbulkan masalah karena media manipulatif dapat digunakan untuk penipuan identitas, penyebaran informasi palsu, pencemaran nama baik, dan penurunan kepercayaan terhadap keaslian media digital. (Akhtar, 2023) menjelaskan bahwa manipulasi wajah berbasis *deep learning* dapat dilakukan dalam berbagai bentuk, seperti pertukaran identitas, manipulasi ekspresi, perubahan atribut, dan sintesis wajah. Sejalan dengan itu, (Malik et al., 2022) menyatakan bahwa *deepfake* pada citra dan video wajah menjadi tantangan penting karena manipulasi dapat dilakukan pada berbagai aspek visual wajah. Selain berdampak pada aspek teknis, *deepfake* juga menjadi persoalan forensik digital karena dapat digunakan untuk menyebarkan informasi palsu dan memengaruhi kepercayaan terhadap media digital (Qureshi et al., 2024). (Diel et al., 2024) juga menunjukkan bahwa kemampuan manusia dalam mengenali *deepfake* masih terbatas, sehingga diperlukan sistem deteksi otomatis yang mampu membantu membedakan media asli dan media manipulatif.

Deteksi *deepfake* umumnya berfokus pada wajah karena bagian tersebut menjadi area utama manipulasi. (Kwan et al., 2024) menjelaskan bahwa pendekatan deteksi *deepfake* modern banyak dikembangkan menggunakan metode berbasis *deep learning*, termasuk *convolutional neural network*, *transformer*, dan pendekatan berbasis sinyal biologis. (Ma et al., 2025) juga

menegaskan bahwa perkembangan deteksi pemalsuan wajah masih menghadapi tantangan pada aspek generalisasi, variasi dataset, dan kualitas manipulasi yang semakin realistis. Tantangan ini menunjukkan bahwa model deteksi tidak hanya perlu memperoleh performa tinggi pada data tertentu, tetapi juga perlu mampu mempertahankan performanya ketika berhadapan dengan variasi jenis manipulasi, kualitas visual, dan kondisi data yang berbeda (Abbasi et al., 2025). Namun, apabila citra masukan masih memuat latar belakang atau area non-wajah, model klasifikasi berpotensi mempelajari informasi yang kurang relevan. Oleh karena itu, tahap praproses yang memfokuskan citra pada area wajah diperlukan agar model lebih terarah dalam mengenali pola manipulasi.

Pemfokusan area wajah dapat dilakukan dengan model deteksi objek sebelum proses klasifikasi. (Alanazi et al., 2023) menunjukkan bahwa analisis pada wilayah wajah berperan penting dalam meningkatkan deteksi *deepfake* karena informasi manipulasi tidak selalu tersebar merata pada seluruh citra. (Lu & Ebrahimi, 2024) juga menjelaskan bahwa evaluasi deteksi *deepfake* perlu memperhatikan kondisi pemrosesan citra dan video yang lebih realistis agar sistem deteksi tidak hanya bergantung pada pola tertentu dari dataset. Dalam konteks deteksi objek, YOLOv11 telah digunakan untuk tugas lokalisasi objek berbasis citra dengan metrik seperti presisi, *recall*, dan *mean average precision* (He et al., 2025). Oleh karena itu, YOLOv11n digunakan pada penelitian ini sebagai detektor ringan untuk mendeteksi dan memotong area wajah/kepala sebelum citra masuk ke tahap klasifikasi.

Pada tahap klasifikasi, penelitian ini menggunakan MobileViT karena model tersebut relevan untuk pemrosesan citra wajah yang membutuhkan representasi fitur lokal dan global. (Rodrigo et al., 2024) menunjukkan bahwa model berbasis *vision transformer* memiliki potensi yang baik pada tugas pengenalan wajah, sedangkan (Yang et al., 2024) menunjukkan bahwa pendekatan berbasis MobileViT dapat digunakan pada tugas pengenalan ekspresi wajah dengan rancangan model yang ringan. Selain itu, (Soudy et al., 2024) menunjukkan bahwa pendekatan yang menggabungkan *convolutional neural network* dan *vision transformer* dapat digunakan dalam deteksi *deepfake*, sehingga penggunaan model yang mampu menangkap informasi lokal dan global menjadi relevan pada tugas ini. Berdasarkan uraian tersebut, penelitian ini bertujuan mengevaluasi integrasi YOLOv11n sebagai praproses deteksi wajah untuk klasifikasi *deepfake* berbasis MobileViT. Evaluasi dilakukan dengan membandingkan dua skenario, yaitu MobileViT tanpa praproses YOLOv11n dan MobileViT dengan praproses YOLOv11n berdasarkan akurasi, presisi, *recall*, dan *F1-score*.

2 METODE

Penelitian ini menggunakan pendekatan eksperimen kuantitatif untuk mengevaluasi pengaruh YOLOv11n sebagai praproses deteksi wajah terhadap klasifikasi *deepfake* berbasis MobileViT. Evaluasi dilakukan dengan membandingkan dua skenario, yaitu MobileViT tanpa praproses YOLOv11n dan MobileViT dengan praproses YOLOv11n. Performa YOLOv11n dievaluasi menggunakan presisi, *recall*, *mAP@50*, dan *mAP@50:95*, sedangkan performa klasifikasi MobileViT dievaluasi menggunakan akurasi, presisi, *recall*, dan *F1-score*. Penggunaan metrik tersebut dilakukan karena presisi, *recall*, *F1-score*, dan *mean average precision* umum digunakan untuk menilai performa model klasifikasi dan deteksi objek (Hicks et al., 2022), (Rainio et al., 2024).

2.1 Dataset Penelitian

Penelitian ini menggunakan dua *dataset* utama. *Dataset* pertama digunakan untuk melatih YOLOv11n sebagai detektor wajah/kepala, sedangkan *dataset* kedua digunakan sebagai sumber data klasifikasi *deepfake* berbasis MobileViT. Rincian *dataset* yang digunakan dalam penelitian ini ditampilkan pada Tabel 1.

Tabel 1. Dataset yang digunakan dalam penelitian

Dataset	Bentuk data	Fungsi	Jumlah Data
Head Detection Computer Vision Dataset (Roboflow)	Citra beranotasi *bounding box*	Pelatihan YOLOv11n	2780 Citra
FaceForensics++ C23 (Kaggle)	Video Real dan Fake	Sumber data klasifikasi *deepfake*	2000 Video

Berdasarkan Tabel 1, *Head Detection Computer Vision Dataset* dari Roboflow digunakan untuk melatih YOLOv11n sebagai detektor wajah/kepala. *Dataset* ini terdiri dari 2433 citra latih, 231 citra validasi, dan 116 citra uji. Karena telah memiliki anotasi *bounding box*, data ini digunakan khusus untuk membentuk kemampuan YOLOv11n dalam mendeteksi area wajah/kepala sebelum tahap klasifikasi. Penggunaan detektor objek sebagai tahap praproses relevan karena metode deteksi objek dapat membantu melokalisasi area penting sebelum citra diproses lebih lanjut oleh model klasifikasi (He et al., 2025).

FaceForensics++ C23 dari Kaggle digunakan sebagai sumber data klasifikasi *deepfake*. *Dataset* ini terdiri atas video asli sebagai kelas Real dan video manipulatif sebagai kelas Fake. Kategori video pada FaceForensics++ C23 yang digunakan dalam penelitian ini ditampilkan pada Tabel 2. Dataset berbasis manipulasi wajah banyak digunakan pada penelitian deteksi *deepfake* karena mampu merepresentasikan variasi manipulasi wajah, seperti pertukaran identitas, manipulasi ekspresi, dan perubahan tekstur wajah (Kwan et al., 2024), (Ma et al., 2025).

Tabel 2. Kategori video pada FaceForensics++ C23

Kategori Video	Jumlah Video yang Digunakan	Label
Original	1000 video	Real
Deepfakes, Face2Face, FaceShifter, FaceSwap, dan NeuralTextures	1000 video (memakai 200 video pada masing-masing kategori)	Fake
Deepfake Detection	Tidak digunakan karena struktur datanya berbeda	-

Berdasarkan Tabel 2, kategori Original digunakan sebagai kelas Real, sedangkan Deepfakes, Face2Face, FaceShifter, FaceSwap, dan NeuralTextures digabungkan sebagai kelas Fake. Kategori *Deepfake Detection* tidak digunakan karena memiliki struktur data yang berbeda dengan data FaceForensics++ C23 yang digunakan pada penelitian ini. Secara keseluruhan, data yang digunakan dalam penelitian ini berjumlah 2000 video, terdiri atas 1000 video Original sebagai kelas Real dan 1000 video manipulatif sebagai kelas Fake.

2.2 Data Hasil Ekstraksi Frame

Data klasifikasi yang digunakan oleh MobileViT bukan berupa video secara langsung, melainkan citra hasil ekstraksi *frame* dari FaceForensics++ C23. Dari setiap video diambil 10 *frame* sebagai representasi data citra. Setiap *frame* diberi label sesuai kelas video asalnya, yaitu Real untuk video asli dan Fake untuk video manipulatif. Jumlah data hasil ekstraksi *frame* ditampilkan pada Tabel 3.

Tabel 3. Data hasil ekstraksi frame FaceForensics++ C23

Sumber Video	Jumlah Video	Jumlah Frame per Video	Total Frame	Label
Original	1000 video	10 frame	10000 citra	Real
Video Manipulatif	1000 video	10 frame	10000 citra	Fake

Berdasarkan Tabel 3, proses ekstraksi menghasilkan 20.000 citra yang terdiri atas 10.000 citra Real dan 10.000 citra Fake. Data hasil ekstraksi *frame* ini digunakan pada dua skenario pengujian. Pada skenario pertama, citra langsung digunakan sebagai masukan MobileViT setelah melalui proses pengubahan ukuran. Pada skenario kedua, citra terlebih dahulu diproses menggunakan YOLOv11n untuk mendeteksi dan memotong area wajah/kepala sebelum diklasifikasikan menggunakan MobileViT.

Setelah proses ekstraksi *frame*, data citra dibagi menjadi data latih, validasi, dan uji dengan rasio 70:15:15. Pembagian tersebut menghasilkan 14.000 citra untuk data latih, 3.000 citra untuk data validasi, dan 3.000 citra untuk data uji. Pada setiap bagian data, komposisi kelas dibuat seimbang, yaitu setengah data berasal dari citra Original berlabel Real dan setengah lainnya berasal dari citra manipulatif berlabel Fake. Dengan demikian, data latih terdiri atas 7.000 citra Real dan 7.000 citra Fake, data validasi terdiri atas 1.500 citra Real dan 1.500 citra Fake, serta data uji terdiri atas 1.500 citra Real dan 1.500 citra Fake.

2.3 Skenario Pengujian

Skenario pengujian disusun untuk mengetahui pengaruh praproses deteksi wajah terhadap performa klasifikasi *deepfake*. Pada skenario pertama, MobileViT menerima citra hasil ekstraksi *frame* tanpa proses deteksi wajah. Pada skenario kedua, YOLOv11n digunakan terlebih dahulu untuk mendeteksi dan memotong area wajah/kepala sebelum citra masuk ke MobileViT. Rincian skenario pengujian ditampilkan pada Tabel 4.

Tabel 4. Skenario Pengujian

Skenario Pengujian	Praproses	Model Klasifikasi	Tujuan
MobileViT tanpa YOLOv11n	*Resize* citra ke 224 × 224 piksel	MobileViT	Menguji performa klasifikasi tanpa pemfokusan wajah.
MobileViT dengan YOLOv11n	Deteksi dan pemotongan wajah/kepala, lalu *resize* ke 224 × 224 piksel	MobileViT	Menguji pengaruh pemfokusan wajah terhadap klasifikasi.

Berdasarkan Tabel 4, perbedaan utama antara kedua skenario terletak pada penggunaan YOLOv11n sebagai praproses deteksi wajah/kepala. Konfigurasi model klasifikasi dibuat sama pada kedua skenario agar perbandingan performa lebih adil.

2.4 Pelatihan YOLOv11n Sebagai Detektor Wajah

YOLOv11n digunakan sebagai model deteksi wajah/kepala pada tahap praproses. Model ini dilatih menggunakan *Head Detection Computer Vision Dataset* Roboflow yang telah memiliki anotasi *bounding box*. Setelah proses pelatihan selesai, YOLOv11n digunakan untuk mendeteksi area wajah/kepala pada citra hasil ekstraksi *frame* dari FaceForensics++ C23. Deteksi objek berbasis YOLO relevan digunakan pada penelitian ini karena model YOLO dirancang untuk melakukan lokalisasi objek secara efisien melalui prediksi *bounding box* dan skor kepercayaan (He et al., 2025). Konfigurasi pelatihan YOLOv11n yang digunakan dalam penelitian ini ditampilkan pada Tabel 5.

Tabel 5. Konfigurasi pelatihan YOLOv11n

Parameter	Nilai
Model	YOLOv11n
Ukuran Citra	640 × 640 piksel
Batch Size	16
Jumlah Epoch	50

Parameter	Nilai
Optimizer	AdamW
Pretrained model	Ya

Berdasarkan Tabel 5, YOLOv11n dilatih selama 50 *epoch* dengan ukuran citra masukan 640 × 640 piksel dan *batch size* 16. Proses pelatihan menggunakan bobot *pretrained* agar model memiliki representasi fitur awal yang lebih baik sebelum dilakukan penyesuaian terhadap dataset deteksi wajah/kepala. Optimasi model dilakukan menggunakan optimizer AdamW untuk memperbarui bobot model selama proses pelatihan. Konfigurasi tersebut digunakan untuk menghasilkan model deteksi wajah/kepala yang selanjutnya diterapkan pada tahap praproses citra sebelum proses klasifikasi menggunakan MobileViT.

Pada tahap deteksi, keluaran YOLOv11n berupa *bounding box* area wajah/kepala. Jika terdapat lebih dari satu hasil deteksi dalam satu citra, maka *bounding box* dengan nilai kepercayaan tertinggi dipilih sebagai area utama. Area tersebut kemudian dipotong dan diubah ukurannya menjadi 224 × 224 piksel sebelum digunakan sebagai masukan MobileViT. Pemfokusan area wajah dilakukan karena informasi manipulasi *deepfake* tidak selalu tersebar merata pada seluruh citra, melainkan banyak muncul pada wilayah wajah tertentu (Alanazi et al., 2023), (Lu & Ebrahimi, 2024). Alur praproses deteksi wajah menggunakan YOLOv11n ditampilkan pada Gambar 1.

Gambar 1. Alur praproses deteksi wajah menggunakan YOLOv11n

Gambar 1 menunjukkan alur praproses YOLOv11n sebelum tahap klasifikasi. Alur tersebut dimulai dari citra hasil ekstraksi *frame*, deteksi wajah/kepala menggunakan YOLOv11n, pemilihan *bounding box* dengan nilai kepercayaan tertinggi, pemotongan area wajah/kepala, pengubahan ukuran citra menjadi 224 × 224 piksel, hingga citra digunakan sebagai masukan MobileViT.

2.5 Klasifikasi Deepfake Menggunakan MobileViT

MobileViT digunakan sebagai model klasifikasi biner untuk membedakan citra Real dan Fake. Model ini digunakan pada kedua skenario pengujian dengan konfigurasi pelatihan yang sama agar perbandingan hasil tetap adil. Pada skenario tanpa YOLOv11n, citra hasil ekstraksi *frame* langsung diubah ukurannya menjadi 224 × 224 piksel. Pada skenario dengan YOLOv11n, citra yang masuk ke MobileViT merupakan hasil pemotongan wajah/kepala dari detektor YOLOv11n.

MobileViT dipilih karena pendekatan berbasis model ringan dan *vision transformer* relevan untuk pemrosesan citra wajah yang membutuhkan representasi lokal dan global. (Rodrigo et al., 2024) menunjukkan bahwa *vision transformer* memiliki potensi pada tugas pengenalan wajah, sedangkan (Yang et al., 2024) menunjukkan bahwa MobileViT dapat digunakan pada tugas pengenalan ekspresi wajah dengan rancangan model yang ringan. Konfigurasi pelatihan MobileViT yang digunakan dalam penelitian ini ditampilkan pada Tabel 6.

Tabel 6. Konfigurasi pelatihan MobileViT

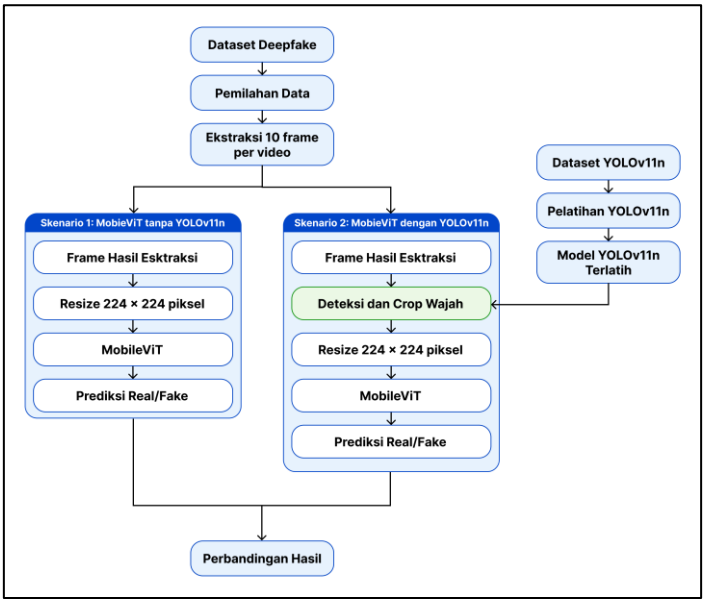
Parameter	Nilai
Model	MobileViT-XXS
Ukuran Citra	224 × 224 piksel

Parameter	Nilai
Batch Size	32
Jumlah Epoch	30
Optimizer	AdamW
Learning Rate	0.0001
Weight Decay	0,01
Kelas	Real dan fake

Berdasarkan Tabel 6, kedua skenario menggunakan konfigurasi pelatihan yang sama, yaitu MobileViT-XXS dengan ukuran masukan 224 × 224 piksel, *batch size* 32, 30 *epoch*, *optimizer* AdamW, *learning rate* 0.0001, dan *weight decay* 0,01. Kesamaan konfigurasi tersebut digunakan agar perbedaan performa yang muncul lebih merepresentasikan pengaruh praproses YOLOv11n.

2.6 Alur Penelitian

Alur penelitian dimulai dari persiapan *dataset*, ekstraksi *frame*, pelatihan YOLOv11n, praproses deteksi wajah, pelatihan MobileViT, evaluasi, dan perbandingan hasil. Pada skenario tanpa YOLOv11n, citra hasil ekstraksi *frame* langsung digunakan untuk klasifikasi. Pada skenario dengan YOLOv11n, citra terlebih dahulu diproses untuk mendeteksi dan memotong area wajah/kepala sebelum masuk ke MobileViT. Alur penelitian secara keseluruhan ditampilkan pada Gambar 2.



Gambar 2. Alur penelitian integrasi YOLOv11n dan MobileViT

Gambar 2 menunjukkan dua jalur pengujian yang digunakan dalam penelitian. Jalur pertama menunjukkan klasifikasi langsung menggunakan MobileViT tanpa YOLOv11n. Jalur kedua menunjukkan proses deteksi dan pemotongan wajah/kepala menggunakan YOLOv11n sebelum citra diklasifikasikan menggunakan MobileViT. Hasil dari kedua jalur kemudian dibandingkan berdasarkan metrik evaluasi.

2.7 Metrik Evaluasi

Evaluasi YOLOv11n dilakukan menggunakan presisi, *recall*, *mAP@50*, dan *mAP@50:95*. Presisi digunakan untuk mengukur ketepatan model dalam mendeteksi objek,

sedangkan *recall* digunakan untuk mengukur kemampuan model menemukan objek yang seharusnya terdeteksi. *mAP@50* digunakan untuk mengukur rata-rata presisi pada ambang *Intersection over Union* 0,50, sedangkan *mAP@50:95* digunakan untuk mengukur rata-rata presisi pada rentang ambang yang lebih ketat. Metrik *mean average precision* umum digunakan dalam evaluasi deteksi objek karena mampu menilai ketepatan prediksi kelas sekaligus kualitas lokalisasi objek (Rainio et al., 2024).

Evaluasi MobileViT dilakukan menggunakan akurasi, presisi, *recall*, dan *F1-score*. Pada penelitian ini, kelas Fake digunakan sebagai kelas positif. TP menunjukkan citra Fake yang diklasifikasikan benar sebagai Fake, TN menunjukkan citra Real yang diklasifikasikan benar sebagai Real, FP menunjukkan citra Real yang salah diklasifikasikan sebagai Fake, dan FN menunjukkan citra Fake yang salah diklasifikasikan sebagai Real. Penggunaan beberapa metrik dilakukan karena satu metrik saja tidak selalu cukup untuk menggambarkan performa model klasifikasi secara menyeluruh (Hicks et al., 2022), (Rainio et al., 2024). Ringkasan metrik evaluasi yang digunakan dalam penelitian ini ditampilkan pada Tabel 7.

Tabel 7. Metrik evaluasi penelitian

Tahap Evaluasi	Metrik	Fungsi
Deteksi YOLOv11n	Presisi, *recall*, *mAP@50*, *mAP@50:95*	Mengukur kemampuan deteksi wajah/kepala
Klasifikasi MobileViT	Akurasi, presisi, *recall*, *F1-score*	Mengukur kemampuan klasifikasi Real dan Fake
Analisis kesalahan	Confusion matrix	Menunjukkan distribusi prediksi benar dan salah

Berdasarkan Tabel 7, evaluasi dilakukan pada dua tingkat, yaitu evaluasi deteksi wajah/kepala menggunakan YOLOv11n dan evaluasi klasifikasi Real/Fake menggunakan MobileViT. Selain itu, *confusion matrix* digunakan untuk melihat distribusi prediksi benar dan salah pada masing-masing skenario klasifikasi.

3 HASIL DAN PEMBAHASAN

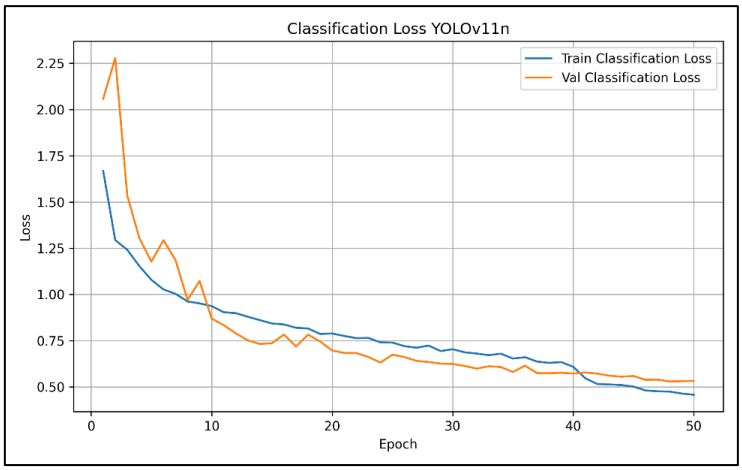
3.1 Hasil Pelatihan YOLOv11n sebagai Detektor Wajah

YOLOv11n dilatih menggunakan *Head Detection Computer Vision Dataset* Roboflow untuk mendeteksi area wajah/kepala sebelum tahap klasifikasi *deepfake*. Evaluasi model deteksi dilakukan menggunakan presisi, *recall*, *mAP@50*, dan *mAP@50:95*. Hasil pelatihan menunjukkan bahwa YOLOv11n memperoleh performa terbaik pada *epoch* ke-40.

Tabel 8. Hasil evaluasi YOLOv11n sebagai detektor wajah/kepala

Metrik	Nilai
Presisi	93,36%
Recall	90,72%
mAP@50	96,41%
mAP@50:95	58,80%
Best epoch	40

Berdasarkan Tabel 8, YOLOv11n memiliki kemampuan deteksi yang baik, ditunjukkan oleh nilai presisi sebesar 93,36% dan *recall* sebesar 90,72%. Nilai *mAP@50* sebesar 96,41% menunjukkan bahwa model mampu mendeteksi area wajah/kepala dengan baik pada ambang *Intersection over Union* 0,50. Namun, nilai *mAP@50:95* sebesar 58,80% menunjukkan bahwa kualitas lokalisasi *bounding box* masih menurun ketika ambang evaluasi dibuat lebih ketat.



Gambar 3. Kurva classification loss YOLOv11n

Gambar 3 diletakkan pada bagian ini untuk menunjukkan perubahan *classification loss* YOLOv11n selama pelatihan. Kurva tersebut memperlihatkan bahwa nilai *loss* pada data latih dan validasi cenderung menurun hingga akhir pelatihan. Hal ini menunjukkan bahwa YOLOv11n mampu mempelajari pola objek wajah/kepala pada data yang digunakan, meskipun masih terdapat fluktuasi pada beberapa *epoch*.

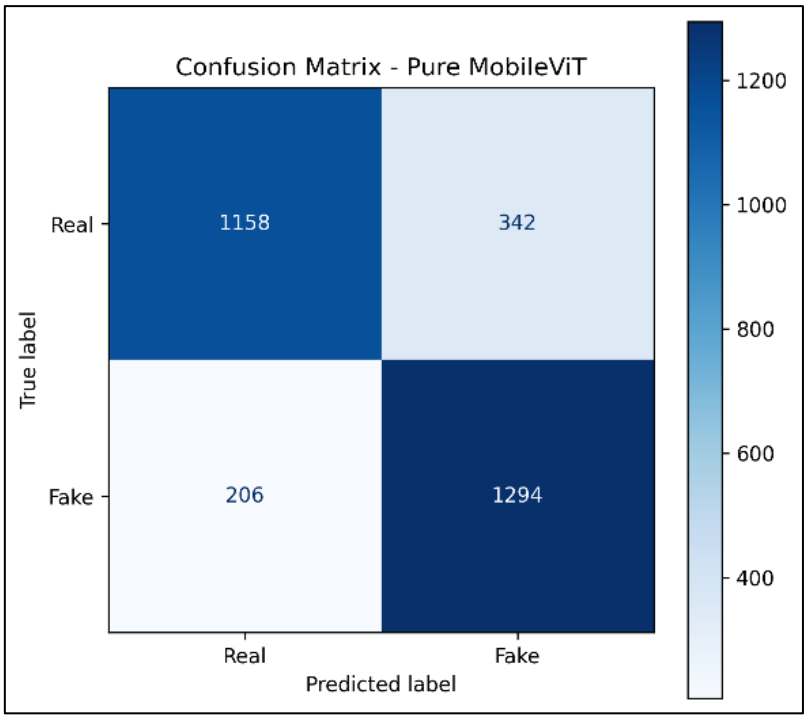
3.2 Hasil Klasifikasi MobileViT tanpa Praproses YOLOv11n

Skenario pertama dilakukan dengan menggunakan MobileViT tanpa praproses deteksi wajah. Pada skenario ini, citra hasil ekstraksi *frame* langsung diubah ukurannya menjadi 224 × 224 piksel dan digunakan sebagai masukan model klasifikasi. Hasil pengujian skenario ini ditampilkan pada Tabel 9.

Tabel 9. Hasil klasifikasi MobileViT tanpa praproses YOLOv11n

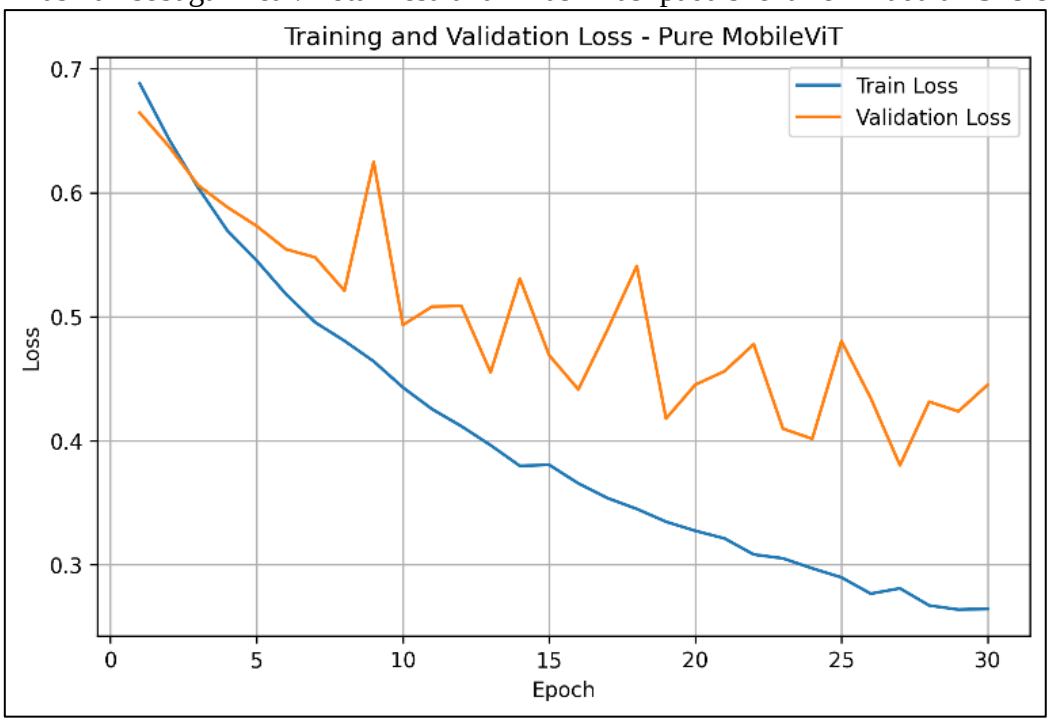
Metrik	Nilai
Akurasi	81,73%
Presisi	79,10%
Recall	86,27%
F1-Score	82,53%
Test Loss	0,3873
Best epoch	27

Berdasarkan Tabel 9, MobileViT tanpa praproses YOLOv11n memperoleh akurasi sebesar 81,73%, presisi sebesar 79,10%, *recall* sebesar 86,27%, dan *F1-score* sebesar 82,53%. Nilai *recall* yang lebih tinggi daripada presisi menunjukkan bahwa model cukup baik dalam mengenali kelas Fake, tetapi masih menghasilkan kesalahan pada citra Real yang diprediksi sebagai Fake.



Gambar 4. Confusion matrix MobileViT tanpa praproses YOLOv11n

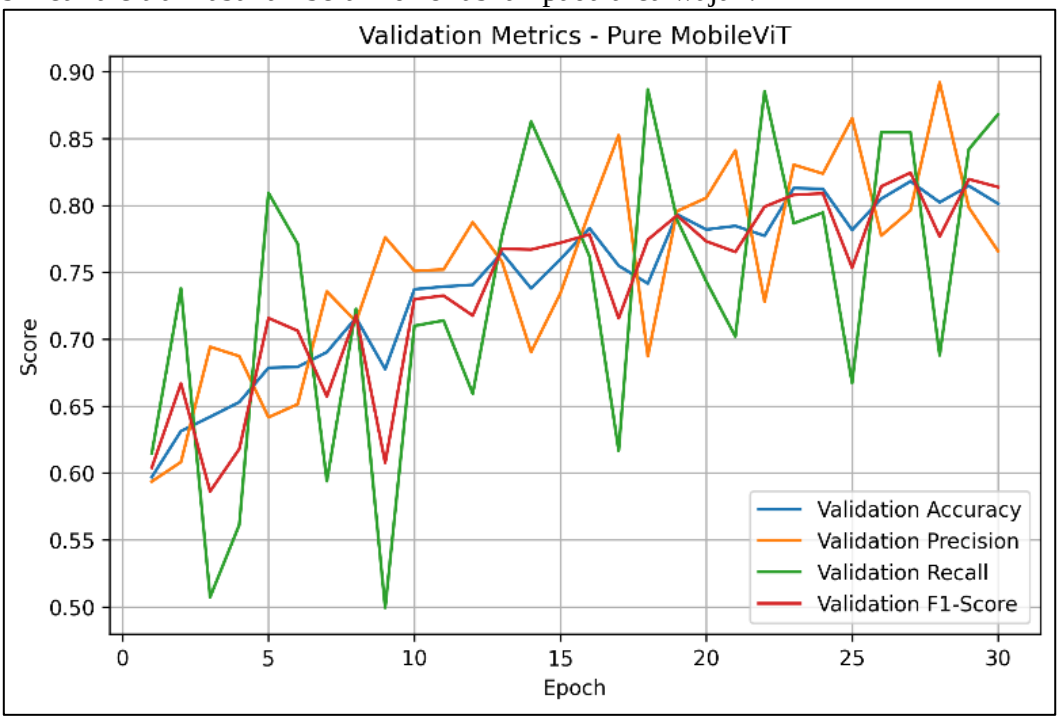
Gambar 4 diletakkan pada bagian ini untuk menunjukkan distribusi prediksi benar dan salah pada skenario tanpa YOLOv11n. Berdasarkan *confusion matrix*, model berhasil mengklasifikasikan 1158 citra Real dan 1294 citra Fake dengan benar. Namun, masih terdapat 342 citra Real yang salah diklasifikasikan sebagai Fake dan 206 citra Fake yang salah diklasifikasikan sebagai Real. Total kesalahan klasifikasi pada skenario ini adalah 548 citra.



Gambar 5. Kurva training loss dan validation loss MobileViT tanpa YOLOv11n

Gambar 5 diletakkan setelah pembahasan *confusion matrix* untuk memperlihatkan perubahan *loss* selama pelatihan. Kurva tersebut menunjukkan bahwa *training loss* mengalami penurunan secara bertahap, tetapi *validation loss* masih mengalami fluktuasi. Kondisi ini

menunjukkan bahwa model masih belum sepenuhnya stabil dalam mempelajari pola data validasi ketika citra masukan belum difokuskan pada area wajah.



Gambar 6. Kurva metrik validasi MobileViT tanpa YOLOv11n

Gambar 6 diletakkan pada bagian ini untuk menunjukkan perkembangan akurasi, presisi, *recall*, dan *F1-score* selama pelatihan. Metrik validasi cenderung meningkat, tetapi masih menunjukkan fluktuasi pada beberapa *epoch*. Hal ini mengindikasikan bahwa MobileViT tanpa praproses YOLOv11n masih dipengaruhi oleh informasi visual yang tidak selalu relevan, seperti latar belakang atau area non-wajah.

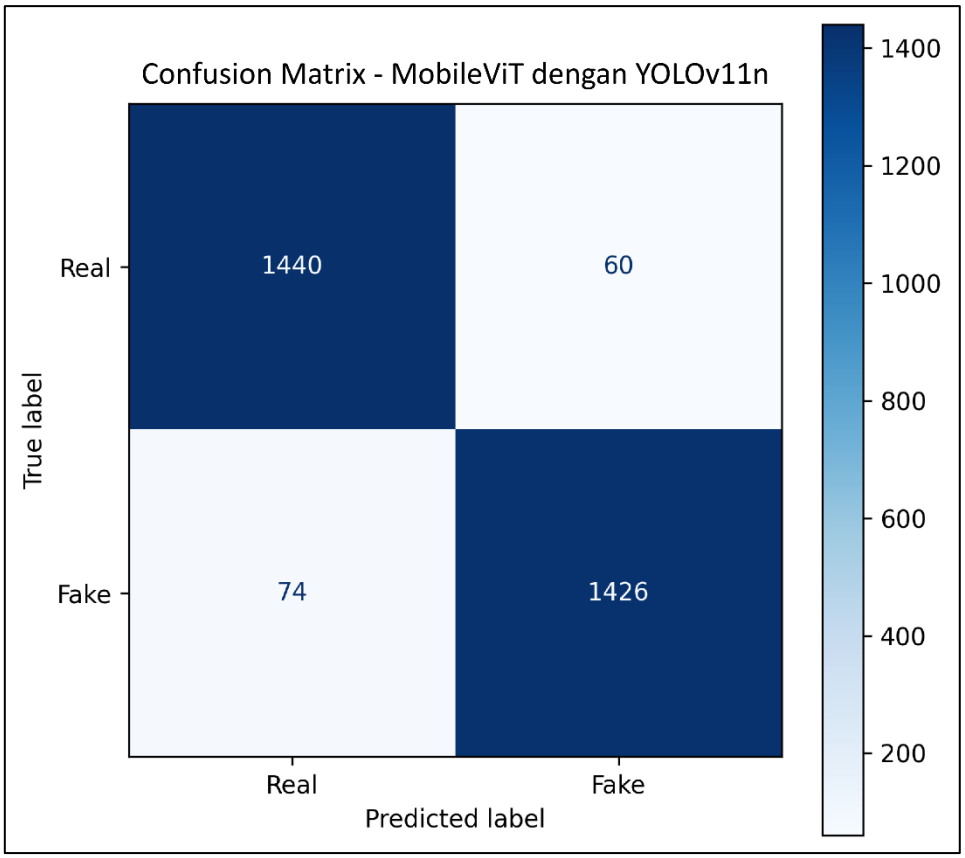
3.3 Hasil Klasifikasi MobileViT dengan Praproses YOLOv11n

Skenario kedua dilakukan dengan menambahkan YOLOv11n sebagai tahap praproses sebelum MobileViT. Pada skenario ini, citra hasil ekstraksi *frame* terlebih dahulu diproses menggunakan YOLOv11n untuk mendeteksi area wajah/kepala. Area dengan nilai kepercayaan tertinggi kemudian dipotong, diubah ukurannya menjadi 224 × 224 piksel, dan digunakan sebagai masukan MobileViT.

Tabel 10. Hasil klasifikasi MobileViT dengan praproses YOLOv11n

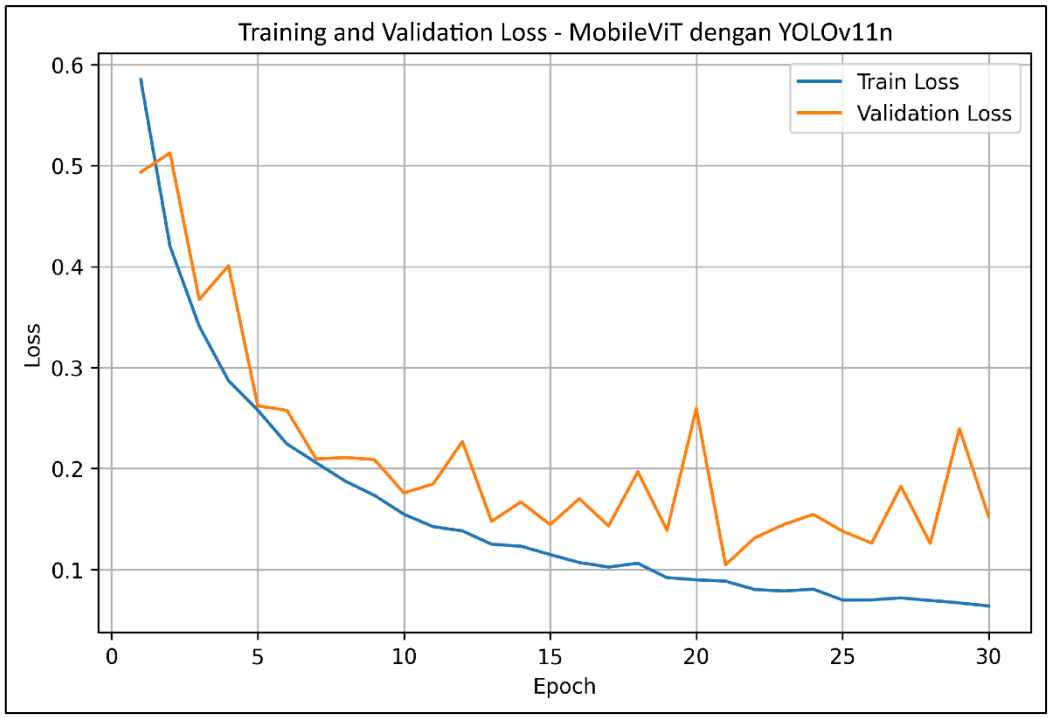
Metrik	Nilai
Akurasi	95,53%
Presisi	95,96%
Recall	95,07%
F1-Score	95,51%
Test Loss	0,1209
Best epoch	21

Berdasarkan Tabel 10, MobileViT dengan praproses YOLOv11n memperoleh akurasi sebesar 95,53%, presisi sebesar 95,96%, *recall* sebesar 95,07%, dan *F1-score* sebesar 95,51%. Hasil ini menunjukkan bahwa pemfokusan area wajah/kepala sebelum klasifikasi mampu meningkatkan kemampuan MobileViT dalam membedakan citra Real dan Fake.



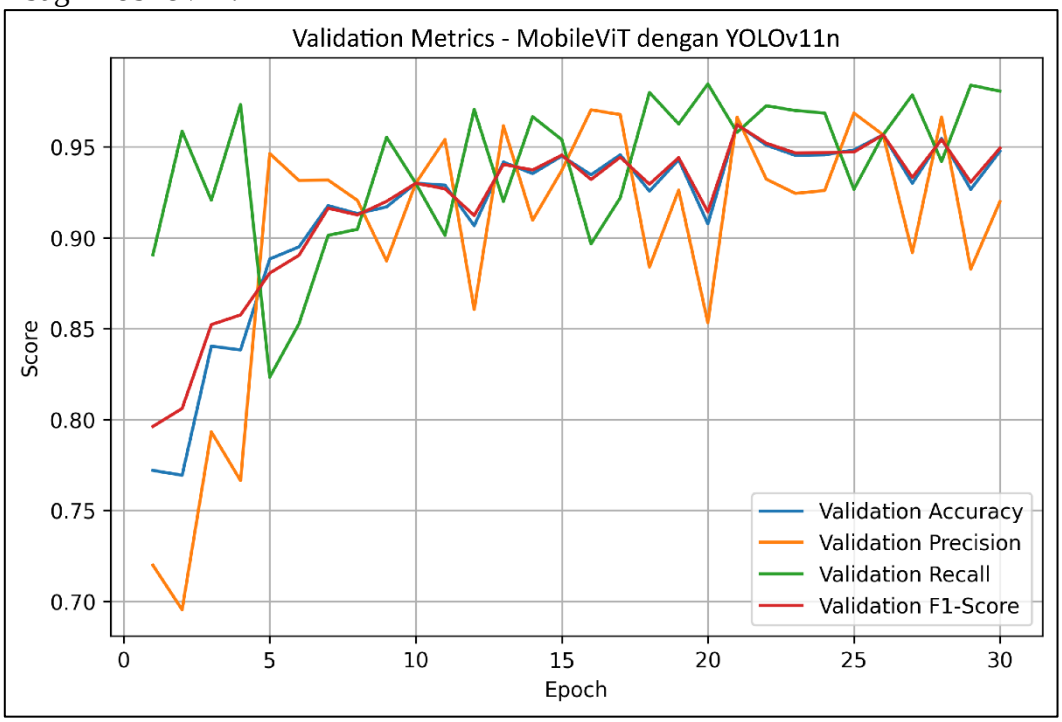
Gambar 7. Confusion matrix MobileViT dengan praproses YOLOv11n

Gambar 7 diletakkan pada bagian ini untuk menunjukkan distribusi hasil klasifikasi pada skenario dengan YOLOv11n. Berdasarkan *confusion matrix*, model berhasil mengklasifikasikan 1440 citra Real dan 1426 citra Fake dengan benar. Kesalahan klasifikasi menurun menjadi 60 citra Real yang salah diklasifikasikan sebagai Fake dan 74 citra Fake yang salah diklasifikasikan sebagai Real. Total kesalahan klasifikasi pada skenario ini adalah 134 citra.



Gambar 8. Kurva training loss dan validation loss MobileViT dengan YOLOv11n

Gambar 8 diletakkan setelah *confusion matrix* untuk menunjukkan perubahan *loss* selama pelatihan. Kurva menunjukkan bahwa *training loss* menurun secara konsisten dan *validation loss* berada pada nilai yang lebih rendah dibandingkan skenario tanpa YOLOv11n. Hal ini menunjukkan bahwa citra hasil pemotongan wajah/kepala memberikan masukan yang lebih terarah bagi MobileViT.



Gambar 9. Kurva metrik validasi MobileViT dengan YOLOv11n

Gambar 9 diletakkan pada bagian ini untuk menunjukkan perkembangan metrik validasi selama pelatihan. Akurasi, presisi, *recall*, dan *F1-score* pada skenario dengan YOLOv11n berada pada nilai yang lebih tinggi dan cenderung stabil. Nilai *F1-score* validasi terbaik

diperoleh pada *epoch* ke-21, sehingga model dengan praproses YOLOv11n menunjukkan performa validasi yang lebih baik.

3.4 Perbandingan Hasil Kedua Skenario

Perbandingan dilakukan untuk mengetahui pengaruh integrasi YOLOv11n sebagai praproses deteksi wajah terhadap performa klasifikasi *deepfake* berbasis MobileViT. Hasil perbandingan kedua skenario ditampilkan pada Tabel 11.

Tabel 11. Perbandingan performa MobileViT tanpa dan dengan praproses YOLOv11n

Skenario	Akurasi	Presisi	Recall	F1-Score	Salah Klasifikasi
MobileViT tanpa YOLOv11n	81,73%	79,10%	86,27%	82,53%	548
MobileViT dengan YOLOv11n	95,53%	95,96%	95,07%	95,51%	134
Peningkatan	+13,80 poin	+16,86 poin	+8,80 poin	+12,98 poin	-414 citra

Berdasarkan Tabel 11, integrasi YOLOv11n sebagai praproses deteksi wajah meningkatkan seluruh metrik evaluasi. Akurasi meningkat sebesar 13,80 poin persentase, presisi meningkat sebesar 16,86 poin persentase, *recall* meningkat sebesar 8,80 poin persentase, dan *F1-score* meningkat sebesar 12,98 poin persentase. Selain itu, jumlah kesalahan klasifikasi menurun dari 548 citra menjadi 134 citra.

Peningkatan tersebut menunjukkan bahwa pemfokusan area wajah/kepala membantu MobileViT memperoleh informasi yang lebih relevan untuk klasifikasi *deepfake*. Pada skenario tanpa YOLOv11n, citra masukan masih memuat area non-wajah yang dapat memengaruhi proses pembelajaran model. Sebaliknya, pada skenario dengan YOLOv11n, citra masukan telah difokuskan pada area wajah/kepala sehingga model lebih terarah dalam mengenali perbedaan antara citra Real dan Fake.

Dengan demikian, hasil penelitian menunjukkan bahwa YOLOv11n tidak hanya berperan sebagai detektor wajah/kepala, tetapi juga membantu meningkatkan kualitas masukan sebelum proses klasifikasi. Integrasi YOLOv11n sebagai praproses terbukti memberikan pengaruh positif terhadap klasifikasi *deepfake* berbasis MobileViT.

4 KESIMPULAN

Berdasarkan hasil pengujian, integrasi YOLOv11n sebagai praproses deteksi wajah/kepala terbukti mampu meningkatkan performa klasifikasi *deepfake* berbasis MobileViT. YOLOv11n memperoleh presisi 93,36%, *recall* 90,72%, *mAP@50* 96,41%, dan *mAP@50:95* 58,80%, sehingga dapat digunakan untuk memfokuskan citra pada area wajah/kepala sebelum tahap klasifikasi.

Pada pengujian klasifikasi, MobileViT tanpa praproses YOLOv11n memperoleh akurasi 81,73%, presisi 79,10%, *recall* 86,27%, dan *F1-score* 82,53%. Setelah menggunakan praproses YOLOv11n, performa meningkat menjadi akurasi 95,53%, presisi 95,96%, *recall* 95,07%, dan *F1-score* 95,51%. Selain itu, jumlah kesalahan klasifikasi menurun dari 548 citra menjadi 134 citra.

Hasil tersebut menunjukkan bahwa pemfokusan area wajah/kepala melalui YOLOv11n membantu MobileViT memperoleh masukan yang lebih relevan untuk membedakan citra Real dan Fake. Dengan demikian, YOLOv11n efektif digunakan sebagai tahap praproses dalam klasifikasi *deepfake* berbasis MobileViT.

Saran untuk penelitian selanjutnya adalah menggunakan data uji dari sumber berbeda untuk mengetahui kemampuan generalisasi model, serta membandingkan YOLOv11n dengan detektor wajah lain agar efektivitas tahap praproses dapat dianalisis lebih luas.

UCAPAN TERIMAKASIH

Penulis mengucapkan terima kasih kepada Program Studi Informatika UPN "Veteran" Jawa Timur atas dukungan yang diberikan dalam pelaksanaan penelitian ini. Ucapan terima kasih juga disampaikan kepada dosen pembimbing yang telah memberikan arahan, masukan, dan evaluasi selama proses penelitian serta penyusunan artikel. Selain itu, penulis berterima kasih kepada pihak-pihak yang telah membantu dalam proses pengolahan data, pelatihan model, dan evaluasi hasil penelitian.

DAFTAR PUSTAKA

Abbasi, M., Váz, P., Silva, J., & Martins, P. (2025). Comprehensive Evaluation of Deepfake Detection Models: Accuracy, Generalization, and Resilience to Adversarial Attacks. *Applied Sciences 2025, Vol. 15, Page 1225*, *15*(3), 1225. https://doi.org/10.3390/APP15031225

Akhtar, Z. (2023). Deepfakes Generation and Detection: A Short Survey. *Journal of Imaging 2023, Vol. 9, Page 18*, *9*(1), 18. https://doi.org/10.3390/JIMAGING9010018

Alanazi, F., Ushaw, G., & Morgan, G. (2023). Improving Detection of DeepFakes through Facial Region Analysis in Images. *Electronics 2024, Vol. 13, Page 126*, *13*(1), 126. https://doi.org/10.3390/ELECTRONICS13010126

Diel, A., Lalgi, T., Schröter, I. C., MacDorman, K. F., Teufel, M., & Bäuerle, A. (2024). Human performance in detecting deepfakes: A systematic review and meta-analysis of 56 papers. *Computers in Human Behavior Reports*, *16*, 100538. https://doi.org/10.1016/J.CHBR.2024.100538

He, L. H., Zhou, Y. Z., Liu, L., Cao, W., & Ma, J. H. (2025). Research on object detection and recognition in remote sensing images based on YOLOv11. *Scientific Reports 2025 15:1*, *15*(1), 14032-. https://doi.org/10.1038/s41598-025-96314-x

Hicks, S. A., Strümke, I., Thambawita, V., Hammou, M., Riegler, M. A., Halvorsen, P., & Parasa, S. (2022). On evaluation metrics for medical applications of artificial intelligence. *Scientific Reports 2022 12:1*, *12*(1), 5979-. https://doi.org/10.1038/s41598-022-09954-8

Kwan, C., Li, B., Yu Gong, L., & Jun Li, X. (2024). A Contemporary Survey on Deepfake Detection: Datasets, Algorithms, and Challenges. *Electronics 2024, Vol. 13, Page 585*, *13*(3), 585. https://doi.org/10.3390/ELECTRONICS13030585

Lu, Y., & Ebrahimi, T. (2024). Assessment framework for deepfake detection in real-world situations. *EURASIP Journal on Image and Video Processing 2024 2024:1*, *2024*(1), 6-. https://doi.org/10.1186/S13640-024-00621-8

Ma, L., Yang, P., Xu, Y., Yang, Z., Li, P., & Huang, H. (2025). Deep learning technology for face forgery detection: A survey. *Neurocomputing*, *618*, 129055. https://doi.org/10.1016/J.NEUCOM.2024.129055

Malik, A., Kuribayashi, M., Abdullahi, S. M., & Khan, A. N. (2022). DeepFake Detection for Human Face Images and Videos: A Survey. *IEEE Access*, *10*, 18757–18775. https://doi.org/10.1109/ACCESS.2022.3151186

Qureshi, S. M., Saeed, A., Almotiri, S. H., Ahmad, F., & Ghamdi, M. A. A. (2024). Deepfake forensics: a survey of digital forensic methods for multimodal deepfake identification on social media. *PeerJ Computer Science*, *10*, e2037. https://doi.org/10.7717/PEERJ-CS.2037

Rainio, O., Teuho, J., & Klén, R. (2024). Evaluation metrics and statistical tests for machine learning. *Scientific Reports 2024 14:1*, *14*(1), 6086-. https://doi.org/10.1038/s41598-024-56706-x

Rodrigo, M., Cuevas, C., & García, N. (2024). Comprehensive comparison between vision transformers and convolutional neural networks for face recognition tasks. *Scientific Reports 2024 14:1*, *14*(1), 21392-. https://doi.org/10.1038/s41598-024-72254-w

Soudy, A. H., Sayed, O., Tag-Elser, H., Ragab, R., Mohsen, S., Mostafa, T., Abohany, A. A., & Slim, S. O. (2024). Deepfake detection using convolutional vision transformers and convolutional neural networks. *Neural Computing and Applications 2024 36:31*, *36*(31), 19759–19775. https://doi.org/10.1007/S00521-024-10181-7

Yang, X., Lan, Z., Wang, N., Li, J., Wang, Y., & Meng, Y. (2024). LiteFer: An Approach Based on MobileViT Expression Recognition. *Sensors 2024, Vol. 24, Page 5868*, *24*(18), 5868. https://doi.org/10.3390/S24185868