PERANCANGAN APLIKASI BERBASIS WEB UNTUK MEMBEDAKAN VIDEO YANG DIHASILKAN OLEH (AI) DENGAN VIDEO ASLI MENGGUNAKAN METODE TRANSFORMER

Eka Nugraha Saktifany Wicaksana, Rohman Dijaya, Irwan Alnarus Kautsar, Nuril Lutvi Azizah
Teknik Informatika, Universitas Muhammadiyah Sidoarjo
Jl. Raya Gelam No.250, Pagerwaja, Gelam, Kec. Candi, Kabupaten Sidoarjo, Jawa Timur 61271
rohman.dijaya@umsida.ac.id

ABSTRAK

Perkembangan teknologi kecerdasan buatan (Artificial Intelligence/AI), khususnya dalam bidang deep learning, telah mendorong munculnya video sintetis atau deepfake yang semakin realistis dan sulit dibedakan dari video asli. Kondisi ini menimbulkan ancaman serius terhadap integritas informasi digital karena berpotensi digunakan untuk menyebarkan misinformasi, penipuan, serta pelanggaran etika, privasi, dan keamanan. Oleh karena itu, diperlukan metode deteksi deepfake yang akurat, adaptif, dan andal dalam menghadapi teknik manipulasi visual yang terus berkembang. Penelitian ini mengusulkan model deteksi deepfake berbasis arsitektur hybrid yang mengombinasikan ConvNeXt dan Vision Transformer (ViT). ConvNeXt dimanfaatkan untuk mengekstraksi fitur lokal citra wajah, sementara Vision Transformer digunakan untuk menangkap konteks global secara lebih komprehensif. Selain itu, model dilengkapi dua jalur rekonstruksi, yaitu Autoencoder (AE) dan Variational Autoencoder (VAE), guna meningkatkan sensitivitas model dalam mendeteksi artefak manipulasi visual yang bersifat halus. Dataset yang digunakan berasal dari video asli dan deepfake yang tersedia secara publik, dengan sekitar satu juta citra wajah hasil ekstraksi frame. Data dibagi menjadi set pelatihan, validasi, dan pengujian. Hasil eksperimen menunjukkan bahwa model mampu mencapai tingkat akurasi hingga 98% pada data pengujian, serta nilai F1-Score yang tinggi, untuk dataset dfdc (99,1),ff++(95,5),Timit(98,3),Celeb DF v2(91,6) mereka dirata-rata menjadi 96,125

Kata kunci : *Deepfake Detection, Deep Learning, ConvNeXt, Vision Transformer*

1. PENDAHULUAN

Fenomena deepfake atau video sintetis berbasis kecerdasan buatan (AI) menjadi isu serius dalam perkembangan teknologi informasi. Dengan kemajuan deep learning, AI kini mampu menghasilkan video yang sangat menyerupai kondisi nyata, sehingga meningkatkan risiko penyebaran misinformasi di ruang digital. Teknologi deepfake tidak hanya dimanfaatkan untuk hiburan, tetapi juga menimbulkan persoalan etika, privasi, dan keamanan informasi. Manipulasi visual semacam ini berpotensi merusak integritas komunikasi, menyulitkan proses verifikasi berita oleh jurnalis, serta menurunkan kepercayaan publik terhadap sumber informasi resmi.Menanggapi permasalahan tersebut, diperlukan pendekatan sistematis untuk mendeteksi keaslian video secara efektif. Penelitian ini mengusulkan model deteksi deepfake berbasis arsitektur Transformer yang memiliki kemampuan kuat dalam memahami hubungan spasial melalui mekanisme self-attention. Untuk memperkuat ekstraksi fitur lokal, digunakan ConvNeXt sebagai pengembangan jaringan konvolusional yang efisien dan akurat. Sementara itu, Vision Transformer (ViT) berperan dalam menangkap konteks global, sehingga mampu mendeteksi artefak manipulasi yang bersifat halus dan tersebar. Kombinasi ConvNeXt dan ViT diharapkan dapat menghasilkan sistem deteksi deepfake yang lebih akurat, adaptif, dan andal dalam menghadapi kompleksitas manipulasi visual yang semakin meningkat [1]

2. TINJAUAN PUSTAKA

Meskipun berbagai penelitian mengenai deteksi deepfake berbasis deep learning telah dilakukan, sebagian besar pendekatan yang ada masih memiliki keterbatasan dalam hal adaptivitas terhadap variasi teknik manipulasi dan jenis data. Penelitian berbasis CNN, seperti penggunaan arsitektur VGG, menunjukkan akurasi yang cukup tinggi, namun masih menghadapi masalah stabilitas pelatihan dan generalisasi. Pendekatan lain menggunakan model kompleks seperti YOLO11 mampu mengekstraksi fitur visual yang kaya, tetapi menghasilkan precision yang rendah serta berisiko mengalami overfitting akibat kompleksitas model dan ketidakseimbangan data.

Pendekatan hybrid yang mengombinasikan CNN dan LSTM terbukti efektif dalam memanfaatkan informasi spasial dan temporal, terutama pada video berdurasi panjang. Namun, performanya sangat bergantung pada jumlah frame, sehingga kurang optimal untuk video dengan durasi pendek. Di sisi lain, arsitektur CNN ringan seperti ResNet18 dan MobileNetV3-Small menunjukkan bahwa efisiensi komputasi dapat dicapai tanpa mengorbankan performa secara signifikan. Model-model ini tidak hanya memberikan akurasi yang kompetitif, tetapi juga memungkinkan implementasi pada perangkat dengan sumber daya terbatas, termasuk aplikasi mobile dan sistem real-time.Secara keseluruhan, studi-studi terdahulu menunjukkan adanya trade-off antara kompleksitas model, efisiensi komputasi, dan

kemampuan generalisasi. Hal ini menegaskan perlunya pendekatan yang mampu menggabungkan akurasi tinggi, stabilitas pelatihan, serta adaptivitas terhadap berbagai variasi konten deepfake yang semakin berkembang..

2.1. Penelitian terdahulu

Meskipun berbagai penelitian telah dilakukan dalam bidang deteksi deepfake berbasis deep learning, sebagian besar pendekatan yang ada masih memiliki keterbatasan, baik dari segi jenis data yang digunakan maupun kemampuan adaptasi terhadap variasi teknik manipulasi deepfake yang terus berkembang. Hal ini menyebabkan performa model belum sepenuhnya stabil ketika dihadapkan pada konten deepfake yang beragam di internet.

Penelitian oleh Jesselyn Mu dkk. menggunakan arsitektur CNN berbasis VGG untuk membedakan wajah asli dan wajah deepfake. Model tersebut mampu mencapai akurasi sebesar 91%, namun menunjukkan pola loss yang tidak konsisten antara tahap pelatihan dan validasi. Kondisi ini mengindikasikan bahwa meskipun model memiliki performa yang cukup baik, masih terdapat permasalahan pada stabilitas dan kemampuan generalisasi model terhadap data yang lebih bervariasi.

Pendekatan lain dilakukan oleh Wawan Kurniawan dkk. dengan memanfaatkan YOLO11 untuk mendeteksi wajah deepfake pada gambar dan video. Model ini memiliki arsitektur yang sangat kompleks dengan jumlah parameter yang besar dan mampu mencapai nilai mAP50 sebesar 68,2%. Namun, nilai precision yang relatif rendah menunjukkan masih tingginya jumlah prediksi yang tidak akurat. Selain itu, ketidakseimbangan data serta kompleksitas model berpotensi menyebabkan overfitting, terutama ketika jumlah data pelatihan terbatas.

Sementara itu, Muhammad Indra Abidin dkk. mengombinasikan CNN ResNext dan LSTM untuk mendeteksi deepfake pada data video. Pendekatan ini efektif dalam menangkap informasi spasial dan temporal, dengan performa terbaik pada video berdurasi panjang. Namun, performa model menurun drastis ketika jumlah frame berkurang, sehingga menunjukkan bahwa metode berbasis sekuensial sangat bergantung pada kelengkapan informasi temporal [1].

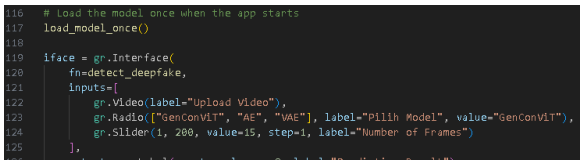
Berbeda dari penelitian sebelumnya, Cintia Putri Prasetia menawarkan solusi yang lebih efisien melalui penggunaan ResNet18, sebuah arsitektur CNN ringan. Model ini mampu mencapai akurasi 92,1% dengan waktu pelatihan yang relatif singkat dan kebutuhan komputasi yang rendah. Hasil ini membuktikan bahwa model ringan tetap mampu memberikan performa yang kompetitif, sekaligus lebih layak untuk diterapkan pada sistem real-time atau perangkat dengan sumber daya terbatas.

Selanjutnya, Matthew Patrick dkk. meneliti penerapan MobileNetV3-Small untuk deteksi deepfake pada perangkat smartphone. Model dengan input citra RGB menunjukkan performa terbaik dan berhasil diimplementasikan secara lokal menggunakan TensorFlow Lite. Pendekatan ini menegaskan bahwa model CNN ringan tidak hanya efisien, tetapi juga praktis untuk aplikasi mobile dan edge computing dengan tetap menjaga privasi pengguna.

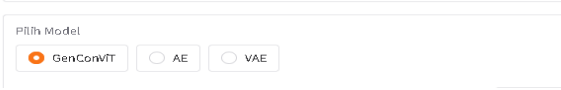
Secara keseluruhan, berbagai penelitian tersebut menunjukkan bahwa terdapat trade-off antara kompleksitas model, akurasi, dan efisiensi komputasi. Model yang kompleks cenderung memiliki masalah generalisasi dan kebutuhan sumber daya yang tinggi, sedangkan model ringan menawarkan efisiensi yang lebih baik namun tetap memerlukan peningkatan dalam hal adaptabilitas terhadap variasi konten deepfake. Oleh karena itu, masih terbuka peluang penelitian untuk mengembangkan metode deteksi deepfake yang akurat, stabil, dan efisien,
[2][3][4][1][5][6]

2.2. Kebaruan penelitian

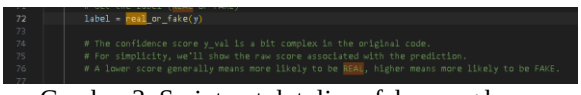
Penelitian ini mengusulkan model deteksi deepfake yang mengombinasikan ConvNeXt dan Vision Transformer (ViT) untuk memanfaatkan keunggulan ekstraksi fitur lokal dan pemahaman konteks global secara bersamaan. Model dirancang dengan pendekatan dual patch embedding dan blok inverted bottleneck yang simetris guna menjaga efisiensi komputasi tanpa menurunkan akurasi. Pendekatan ini difokuskan pada pendeteksian artefak manipulasi yang bersifat halus dan tidak mencolok, yang sering luput dari metode konvensional. Selain itu, pemanfaatan mekanisme self-attention memungkinkan model beradaptasi lebih baik terhadap beragam teknik manipulasi deepfake yang semakin kompleks.



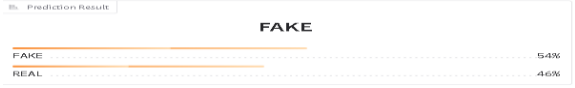
Gambar 1. script tentang model AE VAE dan genconvit



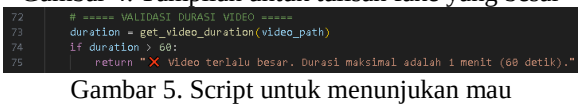
Gambar 2.tampilan model AE VAE dan genconvit



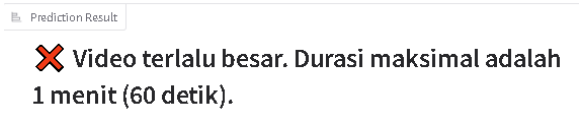
Gambar 3. Script untuk tulisan fake yang besar



Gambar 4. Tampilan untuk tulisan fake yang besar



Gambar 5. Script untuk menunjukan mau memasukkan video lebih dari 1 menit gagal



Gambar 6. Tampilan untuk menunjukan mau memasukkan video lebih dari 1 menit gagal.

3. METODE PENELITIAN

Penelitian ini menggunakan data sekunder berupa video asli dan video deepfake yang tersedia secara publik dengan variasi resolusi, sudut pengambilan gambar, dan kondisi pencahayaan. Video-video tersebut diekstraksi menjadi frame gambar dengan fokus pada area wajah menggunakan algoritma deteksi wajah berbasis Python, kemudian dilakukan proses cropping dan verifikasi kualitas. Dari proses ini dihasilkan sekitar 1.000.000 citra wajah yang terbagi seimbang antara kelas real dan fake. Selanjutnya, dataset dibagi menjadi data pelatihan (80%), validasi (15%), dan pengujian (5%)

3.1. Studi Litelatur

Studi literatur dalam penelitian ini bertujuan untuk memperoleh pemahaman komprehensif mengenai perkembangan metode deteksi deepfake berbasis deep learning. Tinjauan dilakukan terhadap teori dasar pemrosesan citra dan pengenalan pola, serta berbagai arsitektur yang telah digunakan, seperti CNN, YOLO11, dan model hybrid CNN–LSTM. Melalui kajian ini, diidentifikasi keunggulan dan keterbatasan masing-masing pendekatan, sehingga studi literatur berfungsi sebagai landasan konseptual dalam merumuskan metode deteksi deepfake yang lebih adaptif, akurat, dan relevan terhadap karakteristik manipulasi visual pada media digital

3.2. Pengumpulan Dataset

https://www.kaggle.com/datasets/sciarrilli/dfdc-f150, Dataset DFDC terdiri dari lebih dari 100.000 video beresolusi tinggi yang mencakup video asli (real) dan video manipulasi (deepfake). Dataset ini melibatkan 3.426 partisipan dan direkam dalam beragam kondisi pencahayaan, latar belakang, serta sudut pengambilan gambar. Proses manipulasi dilakukan menggunakan delapan teknik deepfake berbeda, sehingga menghasilkan variasi data yang tinggi dan bersifat representatif. Komposisi data menunjukkan rasio sekitar 6:1 antara video deepfake dan video asli, dengan dominasi jumlah video manipulasi..

https://www.kaggle.com/datasets/greatgamedota/faceforensics, Dataset FaceForensics++ terdiri dari 1.000 video asli yang dikumpulkan dari YouTube dan dimanipulasi menggunakan empat metode manipulasi wajah otomatis, yaitu DeepFakes, Face2Face, FaceSwap, dan NeuralTextures. Dataset ini tersedia dalam berbagai tingkat kompresi, seperti c23 (kompresi sedang) dan c40 (kompresi tinggi), serta mendukung beragam resolusi video untuk keperluan evaluasi deteksi deepfake.

https://www.kaggle.com/datasets/nltkdata/timitcorpus, Dataset TIMIT (TM) mencakup 4.380 video deepfake dan 2.563 video asli yang dirancang untuk menguji sistem deteksi manipulasi audio–visual. Meskipun ukurannya lebih kecil dibandingkan DFDC, dataset ini digunakan secara eksklusif pada tahap pelatihan untuk memperluas variasi teknik manipulasi yang dipelajari oleh mode

https://www.kaggle.com/datasets/reubensuju/celeb-df-v2, Dataset Celeb-DF (v2) terdiri dari 890 video asli dan 5.639 video deepfake dengan kualitas manipulasi yang tinggi. Dataset ini menampilkan wajah figur publik, sehingga sangat sesuai untuk pengujian performa model pada skenario dunia nyata..

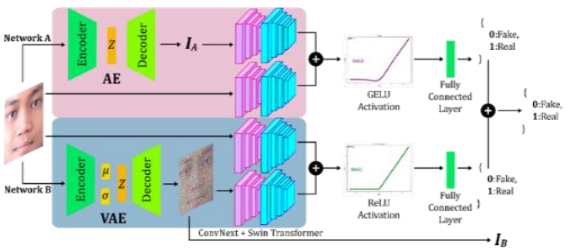
3.3. Pre-Processing Data Video

Proses pre-processing dilakukan secara sistematis untuk memastikan kualitas dan relevansi data dalam pelatihan serta pengujian model deteksi deepfake. Tahapan dimulai dengan ekstraksi region wajah dari setiap video menggunakan library computer vision, sehingga model difokuskan pada area visual yang paling informatif. Selanjutnya, citra wajah disesuaikan ke ukuran standar 224 × 224 piksel dengan format RGB guna menjaga keseimbangan antara detail visual dan efisiensi komputasi.

Setelah itu, dilakukan verifikasi kualitas gambar untuk memastikan wajah terdeteksi dengan jelas dan utuh, serta menghindari keberadaan data bermasalah yang dapat menimbulkan noise label. Frame yang tidak relevan atau gagal terdeteksi wajahnya dihapus untuk menjaga konsistensi dan integritas dataset. Pada tahap akhir, dataset hasil pre-processing yang berjumlah sekitar 1.000.000 citra wajah dibagi menjadi data pelatihan, validasi, dan pengujian dengan distribusi label yang seimbang antara kelas real dan fake. Proses ini bertujuan untuk mendukung pembelajaran model yang adil, stabil, dan memiliki kemampuan generalisasi yang baik..

3.4. Perancangan Model

Penelitian ini menerapkan pendekatan arsitektur hybrid dengan mengombinasikan kemampuan ekstraksi fitur lokal dari ConvNeXt dan pemahaman konteks global dari Vision Transformer untuk deteksi video deepfake. Model dirancang dengan dua alur pemrosesan utama, yaitu Network A yang memanfaatkan Autoencoder (AE) untuk rekonstruksi citra dan Network B yang menggunakan Variational Autoencoder (VAE) berbasis distribusi probabilistik.



Gambar 6. perancangan model

Kedua alur tersebut dikombinasikan dengan ConvNeXt–Swin Transformer sebagai modul ekstraksi fitur dan klasifikasi. Hasil prediksi dari kedua jaringan kemudian digabungkan untuk menghasilkan keputusan klasifikasi akhir, sehingga meningkatkan akurasi dan ketahanan model terhadap variasi manipulasi visual

3.5. Autoencoder (AE) dan Variational Autoencoder (VAE)

Autoencoder (AE) dan Variational Autoencoder (VAE) berfungsi untuk merekonstruksi citra dan mempelajari representasi laten, sehingga membantu model menangkap karakteristik spasial dan distribusi data yang membedakan citra asli dan deepfake.

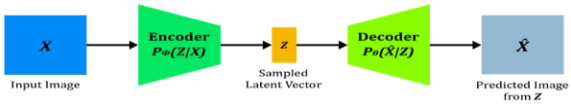
Tabel 1. Konfigurasi encoder AE dan VAE

Layer	Kernel	Stride	Channel AE (Encoder)	Channel VAE (Decoder)
1	3×3	1	3 → 16	3 → 16
2	3×3	1	16 → 32	16 → 32
3	3×3	1	32 → 64	32 → 64
4	3×3	1	64 → 128	64 → 128
5	3×3	1	128 → 256	128 → 256

Tabel 1 menunjukkan konfigurasi encoder pada model Autoencoder (AE) dan Variational Autoencoder (VAE). Setiap layer menggunakan kernel 3×3 dengan stride 1, sementara jumlah channel meningkat secara bertahap dari 3 hingga 256 untuk mengekstraksi fitur citra yang semakin kompleks sebagai representasi laten

3.6. Autoencoder (AE)

Autoencoder terdiri dari Encoder dan Decoder yang bekerja secara berurutan untuk mengompresi citra input ke dalam representasi laten dan merekonstruksinya kembali, sebagaimana diilustrasikan pada Gambar



Gambar 7. Autoenconder

Autoencoder terdiri dari dua komponen utama, yaitu encoder dan decoder, yang bekerja secara berurutan untuk mengompresi citra input ke dalam representasi laten dan merekonstruksinya kembali, sebagaimana diilustrasikan pada Gambar X.

3.7. Evaluasi dan Ablasi Studi

Tahap pelatihan dilakukan dengan mengintegrasikan seluruh komponen arsitektur yang telah dirancang guna mencapai performa optimal dalam deteksi video deepfake. Proses ini mencakup pemilihan backbone, pengaturan hyperparameter, serta strategi optimasi. Model menggunakan ConvNeXt Tiny sebagai backbone berbasis CNN untuk mengekstraksi fitur spasial lokal secara efisien, dan Swin Transformer Tiny sebagai modul Vision Transformer untuk menangkap konteks global melalui mekanisme window-based self-attention. Untuk optimasi, digunakan AdamW optimizer dengan learning rate 0,0001 dan weight decay 0,0001 guna memastikan konvergensi yang stabil sekaligus mengurangi risiko

3.8. Pengujian Validasi

Pengujian dan validasi dilakukan setelah proses pelatihan untuk mengevaluasi kemampuan generalisasi model terhadap data yang belum pernah dilihat sebelumnya. Evaluasi dilakukan menggunakan test set yang terdiri dari 50.000 citra wajah dengan distribusi seimbang antara kelas real dan fake. Kinerja model diukur menggunakan metrik evaluasi standar yang umum digunakan dalam penelitian deep learning untuk menilai performa klasifikasi secara komprehensif

4. HASIL DAN PEMBAHASAN

4.1. Konfigurasi Model

Penelitian ini menetapkan parameter pelatihan standar yang digunakan secara konsisten pada seluruh eksperimen untuk memastikan proses pelatihan berjalan stabil, terukur, dan dapat direplikasi, serta menghasilkan model dengan kemampuan generalisasi yang baik..

4.2. Hasil Pelatihan Model

Proses pelatihan dilakukan menggunakan dua jaringan, yaitu Network A berbasis Autoencoder (AE) dan Network B berbasis Variational Autoencoder (VAE), yang dilatih secara terpisah selama 30 epoch dengan bantuan GPU NVIDIA Tesla V100. Waktu pelatihan Network A sekitar 60 menit dan Network B sekitar 93 menit. Evaluasi model dilakukan pada empat dataset, yaitu DFDC, FF++, DeepfakeTIMIT, dan Celeb-DF v2, dengan membandingkan kinerja model ensemble terhadap masing-masing jaringan untuk mengukur ketahanannya terhadap berbagai teknik manipulasi deepfake.

Tabel 2. Hasil pelatihan model

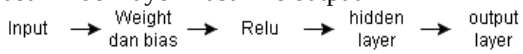
Dataset	Model (AE + VAE)			AE			VAE		
	A	R	F	A	R	F	A	R	F
DFDC	98.5	98.7	98.4	97.5	98.7	97.2	98.4	98.7	95.5
FF++	97.0	95.5	98.5	95.5	94.1	95.5	96.8	95.5	98.0
TIMIT	98.2	-	98.2	97.6	-	97.5	97.8	-	97.8
Celeb-DF (v2)	90.9	92.4	98.5	94.0	87.6	95.9	94.2	83.0	97.6

Pada tabel 2 dijelaskan setiap dataset dimasukan dan uji coba dengan metode ae vae dan keduanya menghasilkan niali sebagai berikut

a. Dataset DFDC pada pengujian metode keduanya model mendeteksei 98, Real nilai 98 dan fake dideteksi 98
b. Dataset FF++ pada pengujian metode keduanya model mendeteksi 97, realnya rata-rata 95 dan Fake dideteksi rata-rata98
c. Dataset Timit pada pengujian metode keduanya model mendeteksi 98 sementara realnya rata-rata 98dan Fake rata-rata dideteksi 97

d. Dataset Celeb b-DF (v2) pada pengujian metode keduanya model mendeteksi 90 sementara realnya 92 dan fakenya rata-rata 98

Perhitungan ini dihasilkan otputnya diteksi deepleraning sendiri perkalian bobot weigth penambahan bias kemudian dilakuakan fungsi aktifasi relu kemudian masuk bagian hidden layer setelah masuk hiden layer masuk ke output



4.3. Pengujian Model

Tabel 3. pengujian model

Parameter	Nilai / Konfigurasi
Optimizer	Adam
Learning rate	0.0001
Epoch	30
Batch size (AE)	32
Batch size (VAE)	16
Augmentation	Albumentations
Input image	224 × 224 × 3

Parameter pelatihan disebut optimal karena dipilih pada titik keseimbangan antara nilai yang terlalu besar dan terlalu kecil, sehingga menghasilkan proses pelatihan yang stabil dan efektif. Secara empiris, konfigurasi ini menunjukkan penurunan loss training dan validation yang konsisten, tanpa divergensi maupun overfitting, serta menghasilkan akurasi dan F1-score tertinggi dibandingkan konfigurasi parameter lainnya.

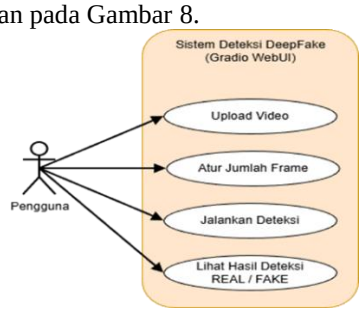
Secara teoretis, pemilihan parameter telah disesuaikan dengan karakteristik arsitektur CNN–Transformer serta kompleksitas Autoencoder (AE) dan Variational Autoencoder (VAE), di mana optimizer Adam, learning rate kecil, dan weight decay membantu menjaga stabilitas pembaruan bobot serta mencegah memorisasi data. Selain itu, penggunaan parameter yang sama pada berbagai dataset (DFDC, FF++, TIMIT, dan Celeb-DF) tetap memberikan performa yang stabil dan konsisten, sehingga menunjukkan bahwa konfigurasi ini bersifat robust dan bukan hasil kebetulan.

https://www.kaggle.com/code/clonesang/deepfake-video-detection-major, Dataset yang diambil pada situs ini terdiri dari 1000 video asli yang diambil dari YouTube dan kemudian dimanipulasi menggunakan empat metode manipulasi wajah otomatis seperti Deepfakes, Face2Face, FaceSwap, dan NeuralTextures. Dataset ini disediakan dalam beberapa tingkat kompresi di antaranya c23 (compression medium) dan c40 (compression high), serta berbagai resolusi video

https://www.kaggle.com/code/kerneler/starter-dfdc-frame-150-6c547762-e/output, Dataset DFDC berisi lebih dari 100.000 video resolusi tinggi yang mencakup video asli dan deepfake, melibatkan 3.426 partisipan. Video direkam dalam beragam kondisi pengambilan dan dimanipulasi menggunakan delapan metode deepfake, sehingga memiliki tingkat variasi yang tinggi. Komposisi data didominasi oleh video manipulasi dengan rasio sekitar 6:1 dibandingkan video asli.

4.4. Use case penggunaan system

Aplikasi deteksi deepfake dirancang untuk pengguna umum dan diuji berdasarkan alur interaksi yang realistis. Hubungan antara pengguna dan sistem digambarkan melalui Use Case Diagram yang ditampilkan pada Gambar 8.



Gambar 8. usecase penggunaan system

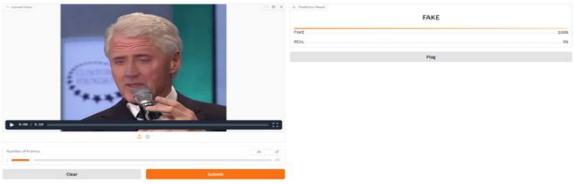
Gambar 8 menunjukkan Use Case Diagram aplikasi deteksi deepfake yang menggambarkan interaksi antara pengguna dan sistem dalam proses penggunaan aplikasi secara umum

4.5. Senario Pengujian Melalui Antarmuka Gradio

berbasis Gradio yang memungkinkan pengguna mengunggah video, menjalankan proses deteksi, dan melihat hasil prediksi secara langsung. Gradio dipilih karena kemampuannya dalam menyajikan pipeline deep learning secara intuitif serta dapat dijalankan baik secara lokal maupun melalui layanan hosting. Proses pengujian meliputi inisialisasi model pre-trained, pengunggahan video oleh pengguna, pengaturan jumlah frame yang dianalisis, serta eksekusi deteksi yang mencakup ekstraksi frame, deteksi wajah, praproses data, inferensi model berbasis AE dan VAE, hingga perhitungan skor probabilitas akhir. Hasil deteksi ditampilkan dalam bentuk label real atau fake beserta nilai confidence score, sehingga pengguna dapat memperoleh hasil klasifikasi secara jelas dan informatif.

4.6. Hasil Pengujian pada antarmuka gradio

Gambar 9. Hasil Deteksi Video Asli (REAL)



Gambar 10. Hasil Deteksi Video Manipulasi (FAKE)

Pengujian dilakukan menggunakan sejumlah Pengujian dilakukan menggunakan sejumlah video real dan deepfake untuk memvalidasi performa model dalam lingkungan aplikasi. Contoh tampilan antarmuka pengujian ditampilkan pada Gambar 9 dan Gambar 10.

4.7. Evaluasi Metrik F1-Score

Selain pengujian melalui antarmuka aplikasi, performa model juga dievaluasi menggunakan metrik F1-Score yang mengombinasikan precision dan recall. Evaluasi ini dilakukan pada model ensemble (AE + VAE) serta masing-masing jaringan internal untuk mengukur kemampuan klasifikasi pada kedua kelas.

Tabel 4. Nilai F1-Score Berdasarkan Dataset

Dataset	Model (AE + VAE)	AE	VAE
DFDC	99.1	98.4	98.4
FF++	95.5	94.9	96.8
TIMIT	98.3	97.5	97.8
Celeb-DF (v2)	91.6	95.2	89.0

Tabel 4 menyajikan nilai F1-Score pada beberapa dataset untuk model ensemble (AE + VAE) serta masing-masing model AE dan VAE. Hasil ini menunjukkan bahwa pendekatan ensemble secara umum memberikan performa yang lebih stabil dan unggul dalam mendeteksi deepfake pada berbagai dataset. untuk dataset dfdc (99,1),ff++(95,5),Timit(98,3),Celeb DF v2(91,6) mereka dirata-rata menjadi 96,125

5. KESIMPULAN DAN SARAN

Penelitian ini berhasil merancang dan mengimplementasikan aplikasi deteksi deepfake berbasis web menggunakan pendekatan arsitektur hybrid yang mengombinasikan ConvNeXt dan Vision Transformer (ViT), serta diperkuat dengan dua jalur rekonstruksi menggunakan Autoencoder (AE) dan Variational Autoencoder (VAE). Pendekatan ini memungkinkan model mengekstraksi fitur lokal secara efektif sekaligus memahami konteks global melalui mekanisme self-attention, sehingga mampu mendeteksi artefak manipulasi visual yang bersifat halus.

Berdasarkan hasil pengujian pada empat dataset publik, yaitu DFDC, FF++, DeepfakeTIMIT, dan Celeb-DF v2, model ensemble (AE + VAE) menunjukkan performa yang stabil dan unggul dibandingkan masing-masing jaringan secara terpisah. Model mencapai nilai F1-Score sebesar 99,1% pada DFDC, 95,5% pada FF++, 98,3% pada TIMIT, dan 91,6% pada Celeb-DF v2, dengan rata-rata keseluruhan sebesar 96,125%. Hasil ini menunjukkan bahwa pendekatan hybrid yang diusulkan memiliki kemampuan generalisasi yang baik terhadap berbagai variasi teknik manipulasi deepfake..

DAFTAR PUSTAKA

[1] R. A. Prawiratama, ‘Design of a Generative AI Image Similarity Test Application and Handmade Images Using Deep Learning Methods’, Telematika, vol. 20, no. 3, p. 326, Nov. 2023, doi: 10.31315/telematika.v20i3.10096.

[2] M. A. I. H. Khusna and S. Pangestuti, ‘DEEPFAKE, TANTANGAN BARU UNTUK NETIZEN (DEEPFAKE, A NEW CHALLENGE FOR NETIZEN)’, PROMEDIA (PUBLIC RELATION DAN MEDIA KOMUNIKASI), vol. 5, Jun. 2019, doi: 10.52447/promedia.v5i2.2300.

[3] Y. Arif Fernandes and Y. Fatma, ‘METODE DEEP LEARNING DALAM TEKNOLOGI DEEPFAKE : SYSTEMATIC LITERATURE REVIEW’, JATI (Jurnal Mahasiswa Teknik Informatika), vol. 9, no. 2, pp. 3403–3410, Apr. 2025, doi: 10.36040/jati.v9i2.12987.

[4] I. Leliana, G. Irhamdhika, A. Haikal, R. Septian, and E. Kusnadi, ‘ETIKA DALAM ERA DEEPFAKE: BAGAIMANA MENJAGA INTEGRITAS KOMUNIKASI’, Jurnal Visi Komunikasi, vol. 22, no. 02, p. 234, Jan. 2024, doi: 10.22441/visikom.v22i02.24229.

[5] D. Putra, S. Sania, and A. Mitrin, ‘Pengaruh Deepfake terhadap Kredibilitas Media Tradisional: Tantangan dan Implikasi di Era Digital’, Sagara Komunika, vol. 1, pp. 13–18, 2024, doi: 10.25311/sagara/Vol1.Iss1.2022.

[6] Y. Hao, L. Dong, F. Wei, and K. Xu, ‘Self-Attention Attribution: Interpreting Information Interactions Inside Transformer’, Proceedings of the AAAI Conference on Artificial Intelligence, vol. 35, no. 14, pp. 12963–12971, May 2021, doi: 10.1609/aaai.v35i14.17533.

[7] J. Feng, H. Tan, W. Li, and M. Xie, ‘Conv2NeXt: Reconsidering Conv NeXt Network Design for Image Recognition’, in 2022 International Conference on Computers and Artificial Intelligence Technologies (CAIT), IEEE, Nov. 2022, pp. 53–60. doi: 10.1109/CAIT56099.2022.10072172.

[8] K. Han et al., ‘A Survey on Vision Transformer’, IEEE Trans Pattern Anal Mach Intell, vol. 45, no. 1, pp. 87–110, Jan. 2023, doi: 10.1109/TPAMI.2022.3152247.

[9] T. Raharjo et al., ‘ANALISIS FORENSIK DEEPFAKE BERBASIS CONVOLUTIONAL NEURAL NETWORK (CNN) UNTUK DETEKSI INKONSISTENSI TEKSTUR DAN POLA PADA CITRA WAJAH’, JATI (Jurnal Mahasiswa Teknik Informatika), vol. 9, no. 2, pp. 2731–2738, Mar. 2025, doi: 10.36040/jati.v9i2.13058.

[10] 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.

[11] Wawan Kurniawan, A. Kurniasih, and Muhamad Abdul Ghani, ‘Real or Deepfake Face Detection in Images and Video Data using YOLO11 Algorithm’, Journal of Artificial Intelligence and Engineering Applications (JAIEA), vol. 4, no. 2, pp. 1514–1521, Feb. 2025, doi: 10.59934/jaiea.v4i2.939.

[12] M. I. Abidin, I. Nurtanio, and A. Achmad, ‘Deepfake Detection in Videos Using Long Short-Term Memory and CNN ResNext’, ILKOM Jurnal Ilmiah, vol. 14, no. 3, pp. 178–185, Dec. 2022, doi: 10.33096/ilkom.v14i3.1254.178-185.

[13] C. P. Prasetia, ‘JITE (Journal of Informatics and Telecommunication Engineering) Efficient Real and Fake Face detection Using ResNet18’, JITE, vol. 4, no. 2, 2025, doi: 10.31289/jite.v9i1.15128.

[14] M. Patrick, C. Lubis,) Agus, and B. Dharmawan, ‘Jurnal Ilmu Komputer dan Sistem Informasi PENDETEKSIAN CITRA DEEPFAKE WAJAH DI SMARTPHONE MENGGUNAKAN MOBILENETV3-SMALL DAN LBP’.