Analisis Perbandingan ResNet50-LSTM dan EfficientNetB0-LSTM pada Sistem Deteksi Video Deepfake Berbasis *Explainable AI*

Muhammad Fathir Ikhsan
Teknik Informatika, Teknik Informatika dan Komputer, Politeknik Negeri Jakarta
Depok, Jawa Barat
muhammad.fathir.ikhsan.tik22@mhsw.pnj.ac.id

Abstract* - Advances in artificial intelligence technology, particularly in the field of generative deep learning, have enabled the synthesis of digital media that is increasingly difficult to distinguish from the original content. Deepfakes a video manipulation technique based on deep neural networks capable of convincingly replacing a person's face in a source video with a target face no longer require specialized computing infrastructure beyond consumer-grade graphics processing units. This study adopts a temporal-based approach using a hybrid CNN-LSTM architecture enhanced by Explainable AI (Grad-CAM) methods. The research methodology is organized into five sequential stages: Dataset Collection, Data Preprocessing, Model Training, Explainable AI, and Model Evaluation. In the overall evaluation of the test data, the ResNet50-LSTM model consistently outperformed other models in four out of five manipulation categories, with accuracy differences ranging from 5.83% (NT) to 10.00% (DF and F2F). In the category-by-category evaluation, both models performed best in the FaceShifter category: ResNet50-LSTM achieved an accuracy of 90.83% and an F1-score of 91.06%, while EfficientNetB0-LSTM achieved an accuracy of 85.00% and an F1-score of 85.71%. This study successfully developed and compared two hybrid CNN-LSTM architectures: ResNet50-LSTM and EfficientNetB0-LSTM. The evaluation results show that the addition of the LSTM component consistently improved detection accuracy across four of the five manipulation categories, with improvements ranging from 7.50% to 20.83%, confirming the effectiveness of temporal analysis in exploiting inconsistencies between frames which are inherent artifacts of the deepfake manipulation process.

Keywords: Deepfake, Deepfake Detection, CNN-LSTM, Explainable AI, Grad-CAM, FaceForensics++

Abstrak*-- Perkembangan teknologi kecerdasan buatan, khususnya pada bidang pembelajaran mendalam generatif, telah memungkinkan sintesis media digital yang semakin sulit dibedakan dari konten aslinya. *Deepfake* teknik manipulasi video berbasis deep neural network yang mampu menggantikan identitas wajah seseorang dalam video sumber dengan wajah target secara meyakinkan kini tidak lagi memerlukan infrastruktur komputasi khusus selain unit pemrosesan grafis tingkat konsumen. Penelitian ini mengadopsi pendekatan berbasis temporal dengan arsitektur hybrid CNN-LSTM yang diperkuat metode *Explainable AI* (Grad-CAM). Metodologi penelitian disusun dalam lima tahap berurutan, meliputi Pengumpulan Dataset, Pra-pemrosesan Data, Pelatihan Model, Explainable AI, dan Evaluasi ModelPada evaluasi keseluruhan data uji, Keunggulan ResNet50-LSTM konsisten di empat dari lima kategori manipulasi, dengan perbedaan akurasi berkisar antara 5,83% (NT) hingga 10,00% (DF dan F2F). Pada evaluasi per-kategori, kedua model menunjukkan performa terbaik pada kategori FaceShifter ResNet50-LSTM mencatat akurasi 90,83% dan F1-Score 91,06%, sementara EfficientNetB0-LSTM mencatat akurasi 85,00% dan F1-Score 85,71%. Penelitian ini berhasil mengembangkan dan membandingkan dua arsitektur hybrid CNN-LSTM ResNet50-LSTM dan EfficientNetB0-LSTM. Hasil evaluasi menunjukkan bahwa penambahan komponen LSTM terbukti meningkatkan akurasi deteksi secara konsisten pada empat dari lima kategori manipulasi dengan kisaran peningkatan 7,50%–20,83%, mengkonfirmasi efektivitas analisis temporal dalam mengeksploitasi inkonsistensi antara frame yang merupakan artefak inheren proses manipulasi *deepfake*.

Kata kunci: Deepfake, Deteksi Deepfake, CNN-LSTM, Explainable AI, Grad-CAM, FaceForensics++

I. PENDAHULUAN

Perkembangan teknologi kecerdasan buatan, khususnya pada bidang pembelajaran mendalam generatif (*generative deep learning*), telah memungkinkan sintesis media digital yang semakin sulit dibedakan dari konten aslinya. *Deepfake* teknik manipulasi video berbasis *deep neural network* yang mampu menggantikan identitas wajah seseorang dalam video sumber dengan wajah target secara meyakinkan kini tidak lagi memerlukan infrastruktur komputasi khusus selain unit pemrosesan grafis tingkat konsumen [1] [2]. Kemajuan ini membawa konsekuensi serius bagi integritas informasi, Kementerian Komunikasi dan Digital Indonesia mencatat lonjakan temuan konten *deepfake* hingga

550% dalam lima tahun terakhir [3], menandai pergeseran teknologi ini dari eksperimen ilmiah menjadi instrumen penyebaran disinformasi berskala besar. Kasus beredarnya video *deepfake* yang seolah menampilkan pernyataan Menteri Keuangan Sri Mulyani tentang "guru sebagai beban negara" pada Agustus 2025 yang dikonfirmasi melalui penelusuran forensik tim Cek Fakta Tempo dengan menemukan tanda air model AI generatif pada video tersebut [4] menjadi cerminan nyata urgensi sistem deteksi yang andal di tengah ekosistem informasi digital Indonesia.

Berbagai pendekatan berbasis pembelajaran mendalam telah dikembangkan untuk mendeteksi konten *deepfake*. *Convolutional Neural Network* (CNN) menjadi fondasi yang paling umum digunakan karena kemampuannya mengekstraksi fitur visual spasial secara otomatis dari setiap frame video [5]. Namun, CNN tunggal memiliki keterbatasan fundamental arsitektur ini tidak dirancang untuk menangkap inkonsistensi temporal antara frame seperti kedipan mata yang tidak natural atau gerakan bibir yang tidak sinkron yang merupakan artefak khas proses manipulasi *deepfake* yang dilakukan secara *frame-by-frame* [1]. *Long Short-Term Memory* (LSTM), sebagai arsitektur *Recurrent Neural Network* yang dirancang khusus untuk pemodelan dependensi sekuensial jangka panjang, terbukti efektif dalam menangkap pola temporal semacam itu [1]. Kombinasi keduanya dalam arsitektur *hybrid* CNN-LSTM telah dilaporkan memberikan peningkatan performa yang signifikan dalam berbagai studi [2] [6] [7], termasuk dalam kompetisi *Deepfake Detection Challenge* (DFDC) di mana berbagai model *hybrid* menempati posisi teratas papan pemeringkatan [2].

Meskipun arsitektur *hybrid* CNN-LSTM telah menunjukkan akurasi yang menjanjikan, tinjauan terhadap literatur terkait mengungkap tiga kesenjangan yang belum ditangani secara memadai. Pertama, mayoritas sistem deteksi *deepfake* berbasis CNN-LSTM bersifat *blackbox* menghasilkan label klasifikasi tanpa penjelasan visual tentang area wajah mana yang menjadi dasar keputusan model [8] [9]. Ketidaktransparanan ini menjadi hambatan kritis dalam konteks forensik digital dan penegakan hukum yang mensyaratkan bukti verifikasi yang dapat dipertanggungjawabkan secara ilmiah [8]. Kedua, penelitian yang mengadopsi arsitektur hybrid CNN-LSTM umumnya menggunakan satu backbone CNN tanpa membandingkan secara sistematis efektivitas *backbone* yang berbeda sebagai pengekstrak fitur spasial dalam konteks pemodelan temporal, sehingga belum tersedia panduan empiris yang memadai bagi peneliti dalam memilih arsitektur yang optimal [6] [7]. Ketiga, belum ada studi yang secara eksplisit mengkarakterisasi batas kondisi (*boundary condition*) di mana analisis temporal berbasis LSTM kehilangan efektivitasnya khususnya dalam menghadapi teknik manipulasi yang tidak meninggalkan inkonsistensi geometris antara frame, namun hanya memodifikasi tekstur permukaan wajah melalui *neural rendering* [1] [10].

Penelitian ini menjawab ketiga kesenjangan tersebut melalui tiga kontribusi utama. Pertama, mengembangkan dan membandingkan secara sistematis dua arsitektur *hybrid* CNN-LSTM yaitu ResNet50-LSTM dan EfficientNetB0-LSTM, keduanya dilengkapi dengan Bidirectional LSTM dan *mechanism attention* menggunakan 2.400 video dari dataset FaceForensics++ C23 [2] dengan evaluasi terpisah pada lima kategori teknik manipulasi seperti DeepFakes (DF), Face2Face (F2F), FaceShifter (FST), FaceSwap (FSP), dan NeuralTextures (NT). Kedua, mengintegrasikan metode *Gradient-weighted Class Activation Mapping* (Grad-CAM) [11] ke dalam *pipeline* deteksi untuk menghasilkan visualisasi *heatmap* per-frame yang mengidentifikasi area wajah yang menjadi dasar keputusan model secara semantik, sehingga mengubah sistem dari *blackbox* menjadi sistem yang transparan dan dapat diverifikasi. Ketiga, mengkarakterisasi secara empiris batas kondisi efektivitas analisis temporal penambahan komponen LSTM ditemukan justru menurunkan performa pada kategori NeuralTextures dibandingkan baseline CNN tunggal suatu fenomena yang belum dilaporkan secara eksplisit dalam literatur terdahulu dan memberikan implikasi penting terhadap desain sistem deteksi *deepfake* yang komprehensif.

II. METODOLOGI PENELITIAN

Deteksi video *deepfake* dapat diklasifikasikan ke dalam tiga kategori utama, yaitu metode berbasis fitur (*feature-based*), berbasis temporal (*temporal-based*), dan berbasis fitur mendalam (*deep feature-based*) [1]. Metode berbasis fitur tradisional menganalisis artefak visual seperti pola kedipan mata yang tidak wajar, inkonsistensi pose kepala, dan ketidaksesuaian geometris wajah untuk mengidentifikasi video palsu [1] [12]. Pendekatan berbasis temporal memanfaatkan *Convolutional Neural Networks*

(CNN) yang dikombinasikan dengan *Long Short-Term Memory* (LSTM) untuk menangkap inkonsistensi temporal antarframe video, di mana CNN mengekstrak fitur tingkat frame sementara LSTM mempelajari pola temporal untuk membedakan antara konten asli dan manipulasi. Metode berbasis fitur mendalam menggunakan arsitektur deep learning canggih seperti *Convolutional Vision Transformer* (CVT), yang dilatih pada dataset *deepfake* berskala besar untuk mendeteksi artefak halus yang tidak terlihat oleh mata manusia [10].

Berdasarkan ketiga kategori pendekatan tersebut, penelitian ini mengadopsi pendekatan berbasis temporal dengan arsitektur *hybrid* CNN-LSTM yang diperkuat metode *Explainable AI* (Grad-CAM). Metodologi penelitian disusun dalam lima tahap berurutan, meliputi:

1. Pengumpulan Dataset

Penelitian ini menggunakan dataset FaceForensics++ dengan tingkat kompresi C23 [13] sebagai *benchmark* utama. Dataset ini dipilih karena mencakup lima kategori teknik manipulasi *deepfake* yang memiliki karakteristik teknis berbeda secara fundamental, sehingga memungkinkan evaluasi yang komprehensif terhadap kemampuan model dalam menghadapi spektrum manipulasi yang beragam. Mengingat keterbatasan sumber daya komputasi dan untuk menjaga keseimbangan antar kelas (*class balance*), penelitian ini menggunakan subset sebanyak 2.400 video yang terdiri dari 400 video asli (original) dan 400 video untuk masing-masing dari kelima kategori manipulasi, sehingga total video deepfake berjumlah 2.000 video.

Tabel 1. Distribusi Dataset FaceForensics++ C23

Kode	Deskripsi	Jumlah Data	Label
DF	DeepFakes	400	
F2F	Face2Face	400	
FST	FaceShifter	400	Fake
FSP	FaceSwap	400	
NT	NeuralTextures	400	
ORG	original	400	Real

2. Pra-pemrosesan Data

Pada tahap pra-pemrosesan data, ekstraksi frame dilakukan dengan membagi video menjadi beberapa segmen temporal yang merata, kemudian mengambil sejumlah frame secara berurutan dari masing-masing segmen untuk merepresentasikan konten video secara keseluruhan. Setelah frame diekstrak dan *face detection* dilakukan dengan dlib, dataset video dibagi menjadi dua *class*, Asli (*class* 0) dan Palsu (*class* 1).

Gambar 1. Perbandingan video Real dengan Fake

3. Pelatihan Model

Pada tahap ini dilakukan proses pelatihan menggunakan data yang telah melalui tahap Pra-pemrosesan Data. Untuk mengatasi potensi ketidakseimbangan jumlah data antara kelas *Real* dan *Fake*, digunakan pendekatan *class weighting* yang diterapkan pada fungsi *loss*. Nilai bobot kelas dihitung berdasarkan distribusi data pada masing-masing kelas menggunakan metode *balanced class weight*. Penerapan class *weighting* ini bertujuan untuk memberikan penyesuaian terhadap kontribusi masing-masing kelas dalam proses pembelajaran, sehingga model tidak bias terhadap kelas dengan jumlah data yang lebih dominan.

4. *Explainable AI*

Pada tahap ini menggunakan metode *Gradient-weighted Class Activation Mapping* (Grad-CAM) untuk menjelaskan keputusan yang dihasilkan oleh model. Grad-CAM digunakan untuk menghasilkan visualisasi berupa *heatmap* yang menunjukkan area wajah pada frame video yang paling berkontribusi terhadap klasifikasi *deepfake* atau asli.

5. Evaluasi Model

Evaluasi model dilakukan pada *test set* yang terpisah menggunakan multiple metrics untuk mengukur performa secara komprehensif. *Metrics* yang digunakan meliputi:

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \quad (1)$$

$$Precision = \frac{TP}{TP + FP} \quad (2)$$

$$Recall = \frac{TP}{TP + FN} \quad (3)$$

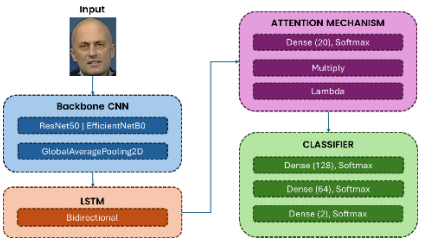
$$F1\ Score = 2 \cdot \frac{Precision \times Recall}{Precision + Recall} \quad (4)$$

III. HASIL DAN PEMBAHASAN

Pada tahap ini menyajikan hasil eksperimen dan analisis perbandingan performa keempat model ResNet50-LSTM, EfficientNetB0-LSTM, ResNet50, dan EfficientNetB0 pada dataset FaceForensics++ C23, mencakup evaluasi kuantitatif berbasis metrik klasifikasi dan validasi kualitatif melalui interpretasi heatmap Grad-CAM.

1. Arsitektur Model

Kedua model yang dibandingkan ResNet50-LSTM dan EfficientNetB0-LSTM dibangun menggunakan kerangka kerja TensorFlow/Keras dengan pendekatan transfer learning dari bobot yang telah dilatih pada dataset ImageNet. Keduanya memiliki struktur arsitektur yang identik pada setiap lapisan, dengan perbedaan hanya pada dimensi keluaran *backbone* CNN masing-masing.



Gambar 2 Arsitektur Model *Hybrid* CNN-LSTM dengan *Mechanism Attention*

2. Perbandingan Performa

Perbandingan menyeluruh terhadap keempat model dua model *hybrid* (ResNet50-LSTM dan EfficientNetB0-LSTM) dan dua model *baseline* tanpa komponen LSTM (ResNet50 dan EfficientNetB0) disajikan pada Tabel 2. Perbandingan ini dilakukan untuk mengkuantifikasi secara empiris kontribusi komponen LSTM dalam meningkatkan akurasi deteksi *deepfake* berbasis video dibandingkan pendekatan CNN tunggal.

Tabel 2. Perbandingan Akurasi (%) Keempat Model pada Seluruh Kategori Dataset FaceForensics++ C23

Model	DF	F2F	FST	FSP	NT
ResNet50	73,33	62,50	73,33	65,00	56,67
ResNet50 + LSTM	85,00	82,50	90,83	85,83	50,83
EfficientNetB0	67,50	60,83	70,00	59,17	50,00
EfficientNetB0 + LSTM	75,00	73,33	85,00	77,50	65,00

Berdasarkan Tabel 2, penambahan komponen LSTM pada arsitektur ResNet50 menghasilkan peningkatan akurasi yang substansial pada empat dari lima kategori manipulasi, dengan kisaran peningkatan antara 11,67% (DF) hingga 20,83% (FSP). Pola serupa teramati pada EfficientNetB0-LSTM, yang menunjukkan peningkatan antara 7,50% (DF) hingga 18,33% (FSP) dibandingkan *baseline*-nya. Konsistensi peningkatan ini pada kedua *backbone* mengonfirmasi bahwa analisis temporal berbasis LSTM memberikan kontribusi yang signifikan dan bukan merupakan artefak arsitektur spesifik. Secara mekanistis, proses manipulasi deepfake yang dilakukan secara *frame-by-frame* meninggalkan inkonsistensi temporal antara frame seperti fluktuasi tekstur di area transisi wajah dan ketidakkonsistenan geometri antar waktu yang hanya dapat terdeteksi melalui analisis sekuensial [1] [7]. LSTM, melalui mekanisme gate yang mampu mempertahankan dan memperbarui informasi konteks jangka panjang, secara efektif menangkap pola inkonsistensi tersebut dalam dimensi temporal yang tidak dapat diakses oleh CNN tunggal.

3. Perbandingan Performa Kedua Arsitektur Hybrid

Setelah kontribusi LSTM dikonfirmasi, analisis difokuskan pada perbandingan antara kedua arsitektur *hybrid* untuk menentukan *backbone* CNN yang paling efektif dalam konteks deteksi video *deepfake*. Metrik evaluasi lengkap untuk ResNet50-LSTM dan EfficientNetB0-LSTM disajikan masing-masing pada Tabel 3 dan Tabel 4.

Tabel 3. Metrik Evaluasi Model ResNet50-LSTM per Kategori Manipulasi (%)

Kode	Accuracy	Precission	Recall	F1- Score
DF	85,00	90,38	78,33	83,93
F2F	82,50	86,79	76,67	81,42
FST	90,83	88,89	93,33	91,06
FSP	85,83	84,13	83,33	86,18
NT	50,83	50,44	95,00	65,90

Tabel 4. Metrik Evaluasi Model EfficientNetB0-LSTM per Kategori Manipulasi (%)

Kode	Accuracy	Precission	Recall	F1- Score
DF	75,00	74,19	76,67	75,41
F2F	73,33	74,14	71,67	72,88
FST	85,00	81,82	90,00	85,71

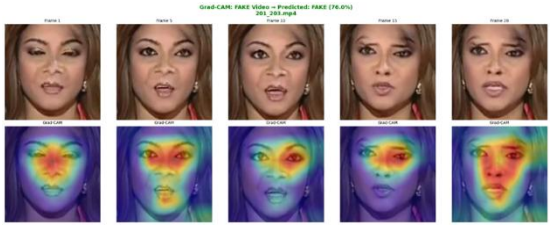
FSP	77,50	75,38	81,67	78,40
NT	65,00	87,50	35,00	50,00

Pada evaluasi keseluruhan data uji, Keunggulan ResNet50-LSTM konsisten di empat dari lima kategori manipulasi, dengan perbedaan akurasi berkisar antara 5,83% (NT) hingga 10,00% (DF dan F2F). Perbedaan ini dapat diatribusikan pada kapasitas representasi fitur ResNet50 yang lebih besar menghasilkan feature map berdimensi 2.048 dibandingkan 1.280 pada EfficientNetB0 [14] [15] yang memungkinkan ekstraksi artefak visual halus yang lebih komprehensif dari setiap frame video. Pada evaluasi per-kategori, kedua model menunjukkan performa terbaik pada kategori FaceShifter ResNet50-LSTM mencatat akurasi 90,83% dan *F1-Score* 91,06%, sementara EfficientNetB0-LSTM mencatat akurasi 85,00% dan F1-Score 85,71%. Pada kategori DF dan F2F, nilai *Precision* ResNet50-LSTM (90,38% dan 86,79%) secara konsisten lebih tinggi dibandingkan *Recall* (78,33% dan 76,67%), mengindikasikan bahwa model bersikap konservatif dalam memprediksi kelas Real sejumlah video asli diklasifikasikan sebagai *Fake* (*False Negative* tinggi).

4. Validasi Semantik via Grad-CAM

Selain evaluasi kuantitatif, validasi kualitatif dilakukan melalui analisis *heatmap* Grad-CAM untuk mengkonfirmasi bahwa keputusan kedua model didasarkan pada fitur visual yang relevan secara semantik terhadap proses manipulasi *deepfake*, bukan pada artefak kompresi, latar belakang, atau bias statistik lainnya. Gambar 3 dan Gambar 4 menyajikan visualisasi Grad-CAM representatif dari model ResNet50-LSTM pada video *Real* dan video *Fake*.

Gambar 3. video Real diprediksi sebagai Real dengan keyakinan 99,4%



Gambar 4. video Fake diprediksi sebagai Fake dengan keyakinan 76,0%

Pada Gambar 3, *heatmap* Grad-CAM menunjukkan distribusi aktivasi yang relatif merata di seluruh permukaan wajah tanpa konsentrasi pada area tertentu, dengan intensitas tertinggi pada area mata dan hidung. Pola distribusi yang merata ini konsisten dengan karakteristik video asli yang tidak memiliki artefak manipulasi terlokalisasi model mengambil keputusan berdasarkan evaluasi tekstur dan geometri wajah secara holistik, bukan berdasarkan anomali spesifik di area tertentu. Sebaliknya, pada Gambar 4, aktivasi terkonsentrasi secara signifikan pada area-area yang memiliki relevansi teknis langsung dengan proses *face-swapping* khususnya batas blending antara wajah sumber dan target, area sekitar hidung, pipi, dan tepi lateral wajah. Area-area ini merupakan titik-titik yang secara teknis paling rentan meninggalkan artefak pada algoritma *deepfake* generatif, karena proses penggabungan (*blending*) antara wajah sumber dan frame target seringkali tidak sempurna pada zona transisi tersebut [1] [10].

IV. KESIMPULAN DAN SARAN

Penelitian ini berhasil mengembangkan dan membandingkan dua arsitektur *hybrid* CNN-LSTM ResNet50-LSTM dan EfficientNetB0-LSTM, keduanya dilengkapi Bidirectional LSTM dan *mechanism attention* untuk deteksi video *deepfake* pada dataset FaceForensics++ C23 dengan integrasi Grad-CAM sebagai metode *Explainable AI*. Hasil evaluasi menunjukkan bahwa penambahan komponen LSTM terbukti meningkatkan akurasi deteksi secara konsisten pada empat dari lima kategori manipulasi dengan kisaran peningkatan 7,50%–20,83%, mengkonfirmasi efektivitas analisis temporal dalam mengeksploitasi inkonsistensi antara frame yang merupakan artefak inheren proses manipulasi *deepfake*. Dalam perbandingan antar-arsitektur, ResNet50-LSTM secara konsisten mengungguli EfficientNetB0-LSTM dan performa terbaik pada kategori FaceShifter (akurasi 90,83%, F1-Score 91,06%), sementara validasi Grad-CAM mengkonfirmasi bahwa keputusan model didasarkan pada fitur yang relevan secara semantik khususnya

area batas blending wajah pada video *Fake* sehingga berhasil mengubah sistem dari *blackbox* menjadi sistem yang transparan dan dapat diverifikasi. Sebagai temuan teoritis baru, penelitian ini mengkarakterisasi secara empiris batas kondisi efektivitas LSTM pada kategori NeuralTextures, komponen LSTM justru menurunkan akurasi sebesar 5,84% dibandingkan *baseline* CNN tunggal karena teknik manipulasi ini tidak meninggalkan inkonsistensi geometris antara frame yang menjadi sumber kekuatan analisis temporal.

Berdasarkan keterbatasan yang diidentifikasi, penelitian selanjutnya disarankan untuk memperluas dataset pelatihan dengan sumber yang lebih beragam seperti Celeb-DF v2 dan *Deepfake Detection Challenge* (DFDC) guna meningkatkan kemampuan generalisasi model terhadap skenario dunia nyata, mengeksplorasi *backbone* CNN yang lebih mutakhir seperti *Vision Transformer* (ViT) atau EfficientNetV2 serta arsitektur sekuensial yang lebih modern seperti *Temporal Transformer* sebagai pengganti LSTM untuk meningkatkan pemodelan dependensi temporal yang lebih kompleks, sekaligus mengembangkan metode XAI yang lebih komprehensif seperti SHAP (*Shapley Additive Explanations*) atau *Integrated Gradients* sebagai pelengkap Grad-CAM khususnya untuk menangani keterbatasan deteksi pada kategori manipulasi berbasis tekstur halus seperti NeuralTextures.

V. REFERENSI

[1] A. Malik, M. Kuribayashi, S. M. Abdullahi, and A. N. Khan, "DeepFake Detection for Human Face Images and Videos: A Survey," 2022, *Institute of Electrical and Electronics Engineers Inc.* doi: 10.1109/ACCESS.2022.3151186.

[2] B. Dolhansky *et al.*, "The DeepFake Detection Challenge (DFDC) Dataset," Oct. 2020, [Online]. Available: http://arxiv.org/abs/2006.07397

[3] Komdigi, "Deepfake Naik 550%, Kemkomdigi Minta Platform Global Sediakan Fitur Cek Konten AI," Kementerian Komunikasi dan Digital. Accessed: Jan. 28, 2026. [Online]. Available: https://portal.komdigi.go.id/kanal-publik/berita-kini/9608

[4] A. Hardiantoro and I. E. Pratiwi, "Video Sri Mulyani yang Sebut Gaji Guru Beban Negara Ternyata 'Deepfake', Apa Itu?," https://www.kompas.com/tren/read/2025/08/20/134500065/video-sri-mulyani-yang-sebut-gaji-guru-beban-negara-ternyata-deepfake-apa?page=all.

[5] Y. Patel *et al.*, "An Improved Dense CNN Architecture for Deepfake Image Detection," *IEEE Access*, vol. 11, pp. 22081–22095, 2023, doi: 10.1109/ACCESS.2023.3251417.

[6] D. Liu, Z. Yang, R. Zhang, and J. Liu, "A Robust Deepfake Video Detection Method based on Continuous Frame Face-swapping," in *Proceedings - 2022 International Conference on Artificial Intelligence, Information Processing and Cloud Computing, AIIPCC 2022*, Institute of Electrical and Electronics Engineers Inc., 2022, pp. 188–191. doi: 10.1109/AIIPCC57291.2022.00048.

[7] D. Xiong, Z. Wen, C. Zhang, D. Ren, and W. Li, "BMNet: Enhancing Deepfake Detection Through BiLSTM and Multi-Head Self-Attention Mechanism," *IEEE Access*, vol. 13, pp. 21547–21556, 2025, doi: 10.1109/ACCESS.2025.3533653.

[8] A. Holzinger, A. Saranti, C. Molnar, P. Biecek, and W. Samek, "Explainable AI Methods - A Brief Overview," 2022, pp. 13–38. doi: 10.1007/978-3-031-04083-2_2.

[9] R. Dwivedi *et al.*, "Explainable AI (XAI): Core Ideas, Techniques, and Solutions," *ACM Comput. Surv.*, vol. 55, no. 9, pp. 1–33, Sep. 2023, doi: 10.1145/3561048.

[10] D. Wodajo and S. Atnafu, "Deepfake Video Detection Using Convolutional Vision Transformer," Mar. 2021, [Online]. Available: http://arxiv.org/abs/2102.11126

[11] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra, "Grad-CAM: Visual Explanations from Deep Networks via Gradient-Based Localization," in *2017 IEEE International Conference on Computer Vision (ICCV)*, IEEE, Oct. 2017, pp. 618–626. doi: 10.1109/ICCV.2017.74.

[12] D. Xiong, Z. Wen, C. Zhang, P. Xiao, and M. Wu, "Deepfake Detection Based on LSTM Networks with Facial Geometric Features," in *2025 4th International Conference on Electronics, Integrated Circuits and Communication Technology, EICCT 2025*, Institute of Electrical and Electronics Engineers Inc., 2025, pp. 627–630. doi: 10.1109/EICCT65471.2025.11099876.

[13] A. Rössler, D. Cozzolino, L. Verdoliva, C. Riess, J. Thies, and M. Nießner, "FaceForensics++: Learning to Detect Manipulated Facial Images," Aug. 2019, [Online]. Available: http://arxiv.org/abs/1901.08971

[14] K. He, X. Zhang, S. Ren, and J. Sun, "Deep Residual Learning for Image Recognition," in

2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), IEEE, Jun. 2016, pp. 770–778. doi: 10.1109/CVPR.2016.90.

[15] M. Tan and Q. V Le, “EfficientNet: Rethinking Model Scaling for Convolutional Neural Networks,” *CoRR*, vol. abs/1905.11946, 2019, [Online]. Available: http://arxiv.org/abs/1905.11946