JISKA (Jurnal Informatika Sunan Kalijaga)
Vol. 11, No. 2, MEI, 2026, Pp. 237 – 251
ISSN: 2527 – 5836 (print) | 2528 – 0074 (online)

Penerapan ResNeXt dan Long Short-Term Memory untuk Deteksi Video Deepfake

Chalifa Chazar [(1)], Firhan Hafiansyah [(2)*], Milda Gustiana Husada [(3)], Uung Ungkawa [(4)], Rizka Milandga Milenio [(3)]
Departemen Teknik Informatika, Institut Teknologi Nasional Bandung, Bandung, Indonesia
e-mail : firhan.hafiansyah@mhs.itenas.ac.id,
{chalifa,mghusada,uung,rizkamilandga}@itenas.ac.id.
* Penulis korespondensi.
Artikel ini diajukan 27 Januari 2026, direvisi 2 Mei 2026, diterima 2 Mei 2026, dan dipublikasikan 25 Mei 2026.

Abstract

Deepfake is a form of facial manipulation in videos that utilizes artificial intelligence-based models to generate highly realistic visual content. The increasing spread of Deepfake has the potential to cause misinformation, manipulate public opinion, and enable the misuse of digital content, making reliable detection systems increasingly necessary. This study develops a face-based Deepfake video detection system by combining spatial and temporal analysis within a unified processing framework. ResNeXt is employed to extract visual facial characteristics from each frame, while Long Short-Term Memory (LSTM) is utilized to learn facial pattern changes across frames in a video sequence. The dataset used is sourced from FaceForensics++, consisting of 1000 original videos and 1000 Deepfake videos. All data are processed through frame extraction and face detection stages. Performance evaluation is conducted using accuracy, precision, recall, and F1-score metrics. The experimental results show that ResNeXt as the baseline model achieves an accuracy of 77.67% and an F1-score of 76.66%, while the integration of LSTM improves system performance with an accuracy of 81.00% and an F1-score of 80.41%. These findings indicate that the utilization of temporal information contributes to improved stability and accuracy in face-based Deepfake video detection.

Keywords: Deepfake, Video Detection, ResNeXt, LSTM, Spatial Features, Temporal Patterns

Abstrak

Deepfake merupakan bentuk manipulasi wajah pada video yang memanfaatkan model berbasis kecerdasan buatan untuk menghasilkan konten visual yang tampak realistis. Penyebaran *Deepfake* yang semakin luas berpotensi menimbulkan misinformasi, manipulasi opini publik, serta penyalahgunaan konten digital, sehingga diperlukan sistem deteksi yang mampu mengenali manipulasi tersebut secara andal. Penelitian ini mengembangkan sistem deteksi video *Deepfake* berbasis wajah dengan menggabungkan analisis spasial dan temporal dalam satu alur pemrosesan. ResNeXt digunakan untuk mengekstraksi karakteristik visual wajah pada setiap *frame*, sedangkan *Long Short-Term Memory* (LSTM) dimanfaatkan untuk mempelajari perubahan pola wajah antar-*frame* dalam urutan video. Dataset yang digunakan berasal dari FaceForensics++, yang terdiri atas 1000 video asli dan 1000 video *Deepfake*. Seluruh data diproses melalui tahap ekstraksi *frame* dan deteksi wajah. Evaluasi kinerja dilakukan menggunakan metrik akurasi, presisi, *recall*, dan *F1-score*. Hasil pengujian menunjukkan bahwa ResNeXt sebagai acuan awal menghasilkan akurasi 77,67% dan *F1-score* 76,66%, sedangkan integrasi LSTM meningkatkan kinerja sistem dengan capaian akurasi 81,00% dan *F1-score* 80,41%. Hasil ini menunjukkan bahwa pemanfaatan informasi temporal mendukung peningkatan stabilitas dan ketepatan deteksi video *Deepfake* berbasis wajah.

Kata Kunci: *Deepfake*, Deteksi Video, ResNeXt, LSTM, Fitur Spasial, Pola Temporal

1. PENDAHULUAN

Deepfake (berasal dari gabungan kata *Deep Learning* dan *fake*) adalah teknologi berbasis kecerdasan buatan yang memanfaatkan *deep learning* untuk mengganti wajah seseorang dalam

video, sehingga menghasilkan video seolah-olah individu tersebut melakukan atau mengucapkan sesuatu yang sebenarnya dilakukan oleh orang lain (Fernandes & Fatma, 2025; Nguyen et al., 2022). *Deep Learning* memanfaatkan beberapa lapisan di antara lapisan masukan *(input layer)* dan lapisan keluaran *(output layer)* dalam proses kerjanya (Diponegoro et al., 2021). Teknologi ini menunjukkan kemampuan tinggi dalam merekonstruksi ekspresi dan gerakan wajah secara presisi, sehingga hasil manipulasi tampak sangat meyakinkan. Salah satu contoh penerapan teknologi ini secara positif adalah dalam produksi film Fast and Furious 7, di mana wajah mendiang Paul Walker digabungkan dengan tubuh sang adik untuk menyelesaikan adegan film yang belum rampung (Gandrova & Banke, 2023).

Selain digunakan dalam industri kreatif, *deepfake* juga sering disalahgunakan untuk membuat konten manipulatif yang menyerupai tokoh publik. Kualitas visual yang semakin meyakinkan menyebabkan banyak orang kesulitan membedakan antara video asli dan hasil rekayasa digital. Teknologi ini telah digunakan secara luas untuk menghasilkan gambar dan video yang menampilkan pornografi, figur politik, serta selebriti, dan sering dimanfaatkan dalam penyebaran propaganda serta pemicu gangguan sosial (Patil et al., 2023). Apabila tidak dikendalikan secara efektif, penyebaran konten *Deepfake* berpotensi mengganggu stabilitas masyarakat dan menimbulkan konflik (Son et al., 2023).

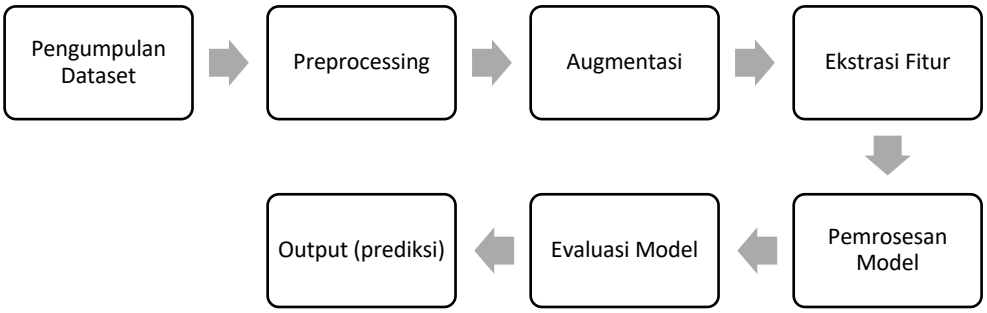
Penelitian deteksi *Deepfake* telah dilakukan melalui berbagai pendekatan. Penelitian oleh Karandikar (2020) menggunakan CNN berbasis *transfer learning* untuk mendeteksi video asli dan video manipulasi melalui analisis setiap *frame*, dengan akurasi sekitar 70%. Pendekatan tersebut berfokus pada analisis visual per *frame*, sehingga pemodelan hubungan temporal antar*frame* belum menjadi fokus utama (Karandikar, 2020). Sementara itu penelitian oleh Maksutov et al. (2020) menerapkan DenseNet169 untuk mendeteksi artefak visual hasil manipulasi dan memperoleh AUC sebesar 60,1%. Pendekatan ini mengandalkan ciri visual statis, sehingga performa deteksi dapat dipengaruhi oleh kejelasan artefak manipulasi yang muncul pada video (Maksutov et al., 2020). Keterbatasan tersebut menunjukkan bahwa pendekatan berbasis spasial saja belum optimal dalam mengenali pola manipulasi *Deepfake* secara menyeluruh.

Berdasarkan keterbatasan tersebut, penelitian ini menggunakan ResNeXt untuk mengekstraksi fitur spasial dari setiap frame, sebagai pengembangan dari *Residual Network* (ResNet) yang memanfaatkan parameter *cardinality* untuk meningkatkan representasi fitur (Kularkar et al., 2023). Selanjutnya, *Long Short-Term Memory* (LSTM) digunakan untuk menganalisis pola temporal antarframe, mengingat kemampuannya dalam memproses data berurutan atau *time-series* (Rosyd et al., 2024). Keterbatasan tersebut menunjukkan bahwa pendekatan berbasis spasial saja belum optimal dalam mengenali pola manipulasi *Deepfake* secara menyeluruh, terutama ketika indikasi manipulasi muncul melalui ketidakkonsistenan antar*frame*.

Penelitian ini bertujuan merancang sistem deteksi video *Deepfake* dengan ResNeXt dan LSTM. Sistem ini dirancang untuk menghasilkan deteksi real atau fake, sekaligus memanfaatkan pola spasial dan temporal untuk menangkap perubahan antarframe pada video. Dengan pendekatan ini, sistem diharapkan mampu meningkatkan akurasi dalam mendeteksi video *Deepfake*.

2. METODE PENELITIAN

Dalam penelitian ini terdapat tujuh tahap proses kerja, yaitu pengumpulan dataset, preprocessing, ekstraksi fitur, augmentasi, pemrosesan model, evaluasi model, dan output (prediksi), sebagaimana ditampilkan pada Gambar 1. Ketujuh tahapan tersebut dilakukan secara berurutan untuk menghasilkan sistem klasifikasi yang mampu mendeteksi pola pada data secara optimal. Alur penelitian ini menggambarkan keseluruhan proses mulai dari pengolahan data hingga menghasilkan prediksi akhir.



Gambar 1 Alur Kerja Penelitian

2.1 Pengumpulan Dataset

Dataset yang digunakan pada penelitian ini adalah FaceForensics++ yang dikembangkan oleh Rossler et al. (2019) (Rossler et al., 2019). Dataset tersebut menyediakan dua kelompok data utama, yaitu *original_sequences* yang berisi video asli dan *manipulated_sequences* yang berisi video hasil manipulasi. Pada penelitian ini, video asli diambil dari *original_sequences*, sedangkan video manipulasi diambil dari kategori *DeepFakes* pada *manipulated_sequences*. Penggunaan dataset yang berasal dari sumber yang sama bertujuan untuk mengurangi perbedaan karakteristik data antar sumber, sehingga model lebih diarahkan untuk mempelajari pola manipulasi wajah daripada perbedaan distribusi dataset.

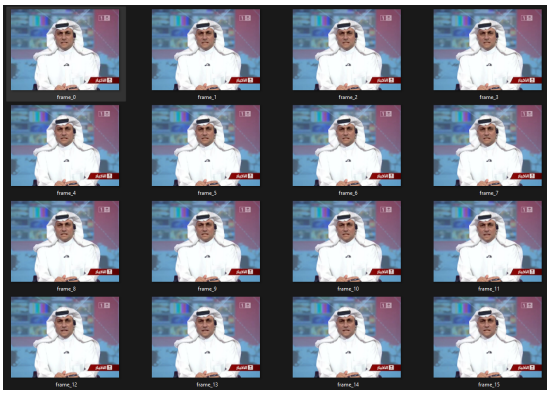
Gambar 2 Contoh Dataset

Sebanyak 1000 video asli dan 1000 video *Deepfake* digunakan, sehingga total data yang dipakai berjumlah 2000 video. Seluruh video berformat .mp4 dengan tingkat kompresi c23, yaitu kompresi berbasis H.264 dengan kualitas visual yang masih cukup baik serta ukuran file yang relatif efisien. Dataset disusun secara terpisah berdasarkan kelas, yaitu real dan fake, untuk mendukung proses *preprocessing* dan analisis lanjutan.

2.2 Preprocessing

Sebelum dataset dapat digunakan untuk pelatihan model, dilakukan tahap preprocessing. Proses ini mencakup ekstraksi frame dari video, deteksi wajah, dan pemotongan wajah agar data sesuai dengan kebutuhan model. Tahapan preprocessing dilakukan untuk memastikan bahwa data yang digunakan memiliki format dan kualitas yang sesuai dengan kebutuhan arsitektur model.

Dataset berupa video dibaca menggunakan OpenCV dan diubah menjadi sekumpulan frame. Untuk menjaga konsistensi temporal, diterapkan uniform sampling sehingga sistem menghasilkan 100 frame yang merepresentasikan keseluruhan video. Hasil dari tahap ini digunakan sebagai input pada proses deteksi wajah. Contoh hasil *frame extraction* terdapat pada Gambar 3.



Gambar 3 Contoh Frame Extraction

Setiap frame terpilih diproses menggunakan Multi-task Cascaded Convolutional Networks (MTCNN) untuk mengidentifikasi area yang mengandung wajah dan menghasilkan bounding box pada frame yang relevan. Proses ini dilakukan agar sistem dapat memfokuskan analisis hanya pada area wajah yang menjadi objek utama deteksi Deepfake. Contoh hasil face detection ditunjukkan pada Gambar 4.

Gambar 4 Contoh Face Detection

Wajah yang terdeteksi kemudian dipotong berdasarkan bounding box dan dilakukan resize menjadi 224 × 224 piksel agar sesuai dengan masukan model CNN. Hasil cropping disimpan sebagai berkas .jpg pada direktori sesuai kelas (real atau fake), sehingga data berfokus pada objek wajah. Tahap ini membantu model memperoleh representasi fitur wajah yang lebih konsisten dan terstandarisasi. Contoh hasil face cropping ditunjukkan pada Gambar 5.

Gambar 5 Contoh Face Cropping

2.3 Split Data

Setelah melalui tahap *preprocessing*, dataset yang terdiri atas 2000 video, yaitu 1000 video *real* dan 1000 video *fake*, dibagi menjadi tiga bagian: pelatihan, validasi, dan pengujian. Pembagian

dilakukan dengan rasio 70% untuk pelatihan sebanyak 1400 video, 15% untuk validasi sebanyak 300 video, dan 15% untuk pengujian sebanyak 300 video. Pada setiap bagian, jumlah video *real* dan *fake* tetap seimbang, sehingga pada data pelatihan terdapat 700 video *real* dan 700 video *fake*, pada data validasi terdapat 150 video *real* dan 150 video *fake*, dan pada data pengujian terdapat 150 video *real* dan 150 video *fake*. Hanya data pelatihan yang melalui tahap augmentasi, sedangkan data validasi dan pengujian tetap menggunakan citra asli tanpa proses augmentasi.

2.4 Augmentasi

Augmentasi data diterapkan pada citra wajah hasil *preprocessing* yang berasal dari data latih untuk meningkatkan keragaman data serta membantu mengurangi risiko *overfitting* selama proses pelatihan model. Teknik augmentasi yang digunakan meliputi *random horizontal flip* dengan probabilitas 30% untuk menambah variasi orientasi wajah, *color jitter* untuk memberikan variasi kondisi pencahayaan dan warna, serta *gaussian blur* dengan probabilitas 20% guna meningkatkan ketahanan model terhadap penurunan kualitas visual. Seluruh proses augmentasi dilakukan secara acak pada setiap citra selama pelatihan.

2.5 Ekstraksi Fitur

Setelah wajah pada setiap frame berhasil dipotong dan melalui tahap *preprocessing*, termasuk augmentasi pada data latih, tahap selanjutnya adalah ekstraksi fitur spasial menggunakan model ResNeXt. Model ini menerima citra wajah hasil *preprocessing* sebagai input dan memanfaatkan *grouped convolution* untuk menangkap pola visual penting seperti bentuk wajah, kontur, dan tekstur kulit. Setiap citra menghasilkan vektor fitur berdimensi 2048, yaitu kumpulan nilai numerik yang merepresentasikan karakteristik visual wajah pada frame tersebut. Pada tahap ini, fitur belum memiliki label *real* atau *fake*, karena proses ekstraksi hanya bertujuan menghasilkan representasi numerik yang akan digunakan pada tahap pelabelan dan pelatihan model berikutnya

2.6 Pemrosesan Model

Pemrosesan model merupakan tahapan penting dalam membangun arsitektur deteksi *Deepfake* berdasarkan data yang telah melalui *proprocessing*. Pada penelitian ini, pemodelan dilakukan dengan menggabungkan arsitektur *Convolutional Neural Network* (CNN) dan *Recurrent Neural Network* (RNN) untuk mengidentifikasi pola spasial dan temporal dalam data video. Model ResNeXt digunakan untuk mengekstraksi fitur dari masing masing frame, sedangkan LSTM digunakan untuk menangkap pola berurutan antarfame

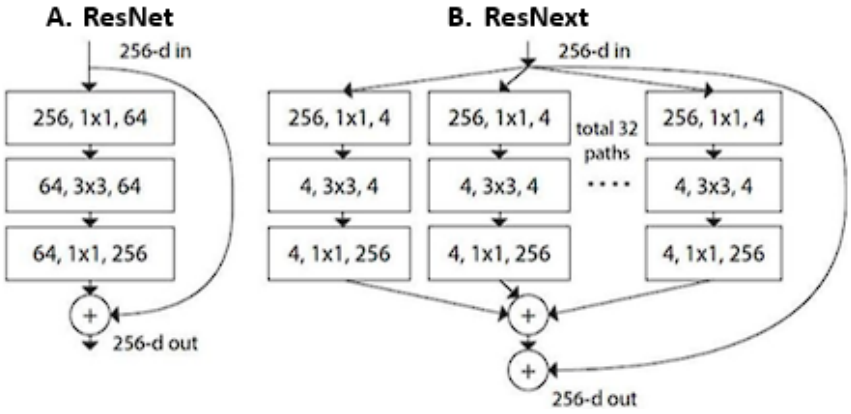
2.6.1 Arsitektur ResNeXt

ResNeXt digunakan sebagai ekstraksi fitur spasial dari setiap frame wajah yang telah melalui proses *preprocessing*. Model ini menerima masukan berupa citra wajah dan menghasilkan vektor fitur berdimensi 2048 untuk setiap frame. Vektor-vektor ini merepresentasikan informasi visual penting yang diperlukan untuk proses klasifikasi.

ResNeXt merupakan pengembangan dari arsitektur ResNet, dengan konsep utama berupa penggunaan jalur paralel *(cardinality)* dalam blok residual (Badruzzaman & Arymurthy, 2024). Hal ini memungkinkan peningkatan kemampuan dalam mengekstraksi fitur tanpa menambah kompleksitas model secara signifikan. Vektor fitur hasil dari ResNeXt selanjutnya digunakan sebagai input dalam pemrosesan urutan oleh LSTM.

ResNeXt dipilih dalam penelitian ini karena kemampuannya dalam menangkap variasi pola visual yang kompleks melalui mekanisme cardinality, sehingga dapat merepresentasikan karakteristik wajah secara lebih kaya. Hal ini menjadi penting dalam deteksi *Deepfake*, karena manipulasi wajah sering melibatkan perubahan tekstur dan detail visual yang halus. Gambar 6. memperlihatkan dua arsitektur: bagian A menunjukkan blok residual pada ResNet, sedangkan bagian B menunjukkan struktur blok pada ResNeXt yang digunakan dalam penelitian ini.

Perbedaan utama terletak pada adanya beberapa jalur transformasi paralel pada ResNeXt, yang memungkinkan model mengekstraksi fitur dengan representasi lebih beragam dan efisien.

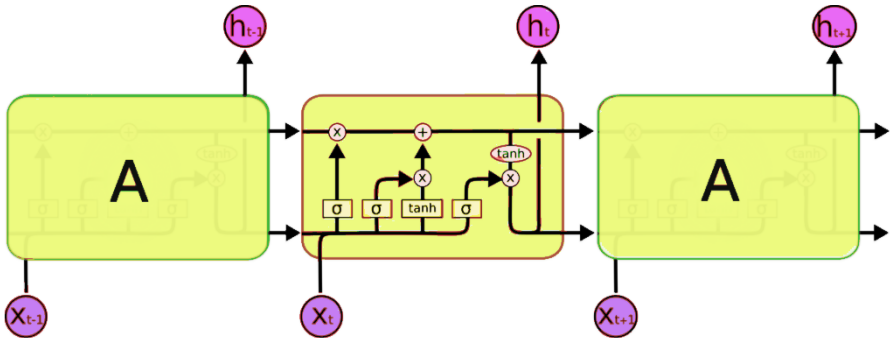


Gambar 6 Arsitektur ResNet dan ResNeXt (Yeh et al., 2021)

2.6.2 Arsitektur LSTM

Long Short-Term Memory (LSTM) merupakan salah satu model yang dirancang untuk memproses data berurutan dengan kemampuan mempertahankan informasi dari langkah sebelumnya (Wang et al., 2021). Karakteristik ini memungkinkan LSTM menangkap keterkaitan antarframe dalam video, sehingga sesuai digunakan untuk menganalisis perubahan visual yang terjadi secara berurutan. Dalam penelitian ini, LSTM digunakan untuk memproses urutan fitur hasil ekstraksi dari setiap frame wajah.

Arsitektur LSTM terdiri atas tiga komponen utama, yaitu *input gate*, *forget gate*, dan *output gate*, yang berfungsi mengatur aliran informasi di dalam memori internal (*cell state*) (Kholifatullah & Prihanto, 2023). Setiap komponen berperan dalam menentukan informasi yang disimpan, diperbarui, atau dikeluarkan pada setiap langkah pemrosesan. Struktur dan alur kerja LSTM ditunjukkan pada Gambar 7. Melalui mekanisme pengelolaan memori tersebut, LSTM mampu memanfaatkan informasi urutan frame secara efektif.



Gambar 7 Arsitektur LSTM (Abdu et al., 2021)

Kemampuan ini mendukung analisis video *Deepfake*, khususnya dalam mengenali perubahan visual antarframe yang tidak selalu dapat diidentifikasi melalui satu frame secara terpisah. LSTM dipilih dalam penelitian ini karena kemampuannya dalam memodelkan dependensi jangka panjang pada data berurutan, sehingga mampu menangkap hubungan temporal antarframe secara lebih efektif. Hal ini menjadi penting dalam deteksi *Deepfake*, karena pola manipulasi wajah sering muncul dalam bentuk perubahan antarframe yang tidak konsisten dan sulit dikenali jika hanya dianalisis secara spasial.

2.6.3 Pelatihan Model

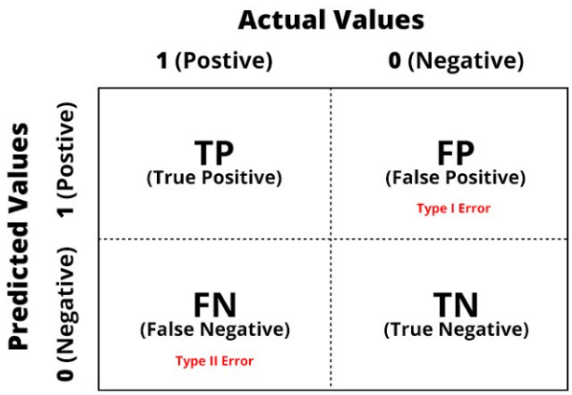
Pelatihan model dilakukan dengan menerapkan arsitektur ResNeXt yang digabungkan dengan LSTM untuk memanfaatkan informasi spasial dan urutan frame video. Setiap video direpresentasikan sebagai urutan fitur spasial dengan jumlah 100 frame yang digunakan sebagai masukan model, di mana setiap frame diproses menggunakan ResNeXt untuk menghasilkan fitur yang kemudian disusun sebagai urutan masukan ke dalam LSTM. Urutan ini selanjutnya diproses oleh LSTM untuk menangkap hubungan temporal antarframe sebelum dilakukan proses klasifikasi akhir. Pada arsitektur model, diterapkan dropout sebesar 0,3 sebagai bagian dari mekanisme regularisasi.

Proses pelatihan menggunakan optimizer AdamW dengan learning rate 0,001 dan batch size 32. Fungsi kerugian yang digunakan adalah *Binary Cross Entropy with Logits*. Pelatihan dijalankan hingga maksimum 100 epoch dengan penerapan *early stopping* pada *patience* 15 epoch. Regularisasi tambahan dilakukan melalui *weight decay* sebesar 0,0001, sedangkan kestabilan gradien dijaga menggunakan *gradient clipping* dengan batas 1,0. Penyesuaian *learning rate* dilakukan menggunakan ReduceLROnPlateau berdasarkan kinerja validasi.

Pemilihan parameter pelatihan didasarkan pada kebutuhan menjaga kestabilan proses pembelajaran model serta mengurangi risiko *overfitting*. Optimizer AdamW digunakan karena mampu menyesuaikan pembaruan parameter secara adaptif selama pelatihan. *Learning rate* 0,001 digunakan sebagai nilai awal pembelajaran, sedangkan batch size 32 dipilih untuk menyeimbangkan kebutuhan memori dan efisiensi komputasi. Penerapan *early stopping*, dropout, dan *weight decay* ditujukan sebagai mekanisme regularisasi, sementara gradient clipping dan ReduceLROnPlateau digunakan untuk menjaga kestabilan proses pelatihan.

2.7 Evaluasi Model

Evaluasi model dilakukan untuk menilai performa sistem dalam mendeteksi video *Deepfake*. Penilaian dilakukan terhadap data uji yang telah dipisahkan dari proses pelatihan. Untuk mengukur performa model, digunakan empat metrik utama, yaitu akurasi, presisi, *recall*, dan F1-score. Seluruh metrik tersebut dihitung berdasarkan nilai-nilai yang terdapat pada *confusion matrix.*



Gambar 8 Confusion Matrix (Jefiza et al., 2023)

Gambar 8 menampilkan struktur *confusion matrix* digunakan untuk mengevaluasi kinerja model dalam proses deteksi. *Confusion matrix* berisi empat nilai utama, yakni *True Positive* (TP), *False Positive* (FP), *True Negative* (TN), dan *False Negative* (FN). Nilai-nilai tersebut dipakai sebagai acuan dalam menghitung akurasi, presisi, *recall*, serta *F1-score.*

Akurasi merupakan metrik evaluasi standar yang digunakan untuk menilai seberapa banyak prediksi yang tepat dibandingkan seluruh data uji. Metrik ini memberi gambaran umum

kemampuan model dalam menghasilkan keputusan deteksi yang benar (Vakili et al., n.d.). Nilai akurasi dihitung menggunakan Pers. (1).

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

Presisi menilai seberapa besar bagian dari prediksi positif yang benar-benar sesuai dengan kelas positif. Nilai presisi yang tinggi menunjukkan bahwa model jarang menghasilkan kesalahan positif (*false positive*), sehingga hasil deteksi yang muncul cenderung relevan dan valid (Jiwani & Satrio Waluyo Poetro, 2025)**.** Nilai presisi dihitung menggunakan Pers. (2).

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

Recall digunakan untuk menilai kemampuan model mengidentifikasi sebanyak mungkin sampel positif yang seharusnya terdeteksi. Nilai *recall* yang lebih tinggi mengurangi kemungkinan sistem melewatkan kasus positif *(false negative)*, sehingga lebih banyak data relevan dapat dikenali (Jiwani & Satrio Waluyo Poetro, 2025)**.** Nilai *recall* dihitung menggunakan Pers. (3).

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

F1-score merupakan rata-rata harmonis dari nilai presisi dan *recall* yang digunakan untuk menyeimbangkan kedua metrik tersebut. Metrik ini bermanfaat pada kondisi data yang tidak seimbang karena mampu memberikan gambaran performa model yang lebih representatif dibandingkan hanya menggunakan akurasi (Hakim et al., 2025)**.** Nilai F1-score dihitung menggunakan Pers. (4).

$$F1 - Score = 2 \text{ x } \frac{Precision \text{ x } Recall}{Precision + Recall} \quad (4)$$

3. HASIL DAN PEMBAHASAN

Penelitian ini menerapkan ResNeXt dan *Long Short-Term Memory* (LSTM) dalam proses deteksi video *Deepfake* berbasis wajah. Hasil pelatihan model, evaluasi kinerja, dan pengujian dibahas pada bagian ini sesuai dengan tahapan penelitian yang dilakukan.

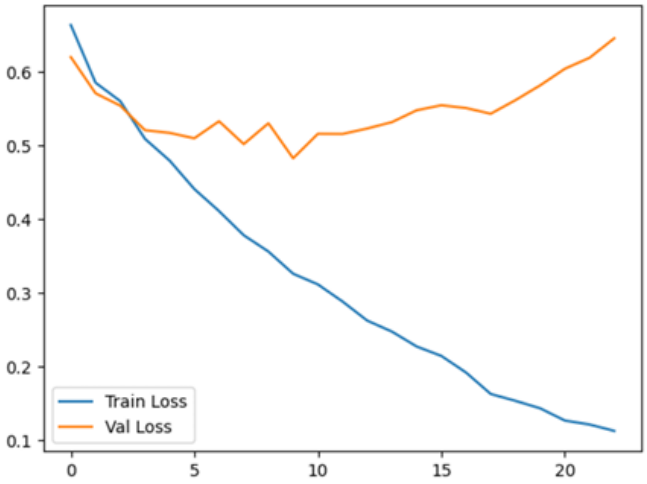
3.1 Model ResNeXt

Pelatihan model ResNeXt dilakukan menggunakan learning rate 0,0001, batch size 32, weight decay 0,0001, dan dropout sebesar 0,30 pada lapisan akhir. Proses optimisasi menggunakan AdamW dengan fungsi kerugian Binary Cross Entropy with Logits Loss, pemotongan gradien sebesar 1,0, serta penyesuaian learning rate menggunakan ReduceLROnPlateau. Mekanisme early stopping diterapkan dengan patience 15 epoch dan pelatihan dihentikan pada epoch ke-23.

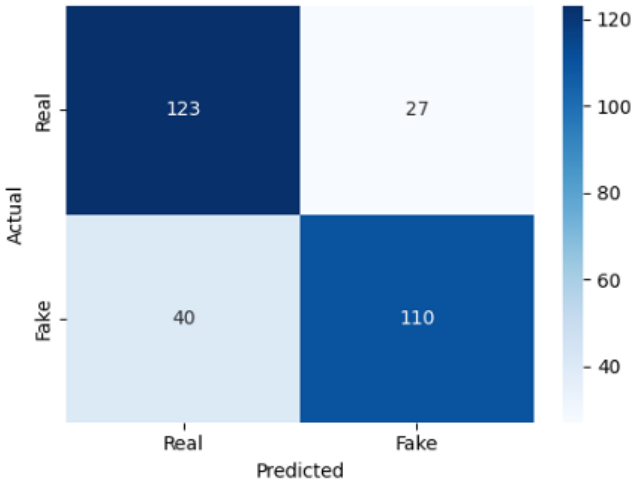
Berdasarkan Gambar 9, nilai training loss terus menurun selama proses pelatihan. Sementara itu, validation loss mencapai nilai terendah pada epoch ke-8 sebelum kembali meningkat pada epoch berikutnya. Perbedaan antara training loss dan validation loss yang semakin jelas setelahnya mengindikasikan terjadinya overfitting. Oleh karena itu, model dengan kinerja validasi terbaik dipilih sebagai model akhir.

Evaluasi pada dataset uji menghasilkan akurasi sebesar 77,67%, presisi 80,29%, recall 73,33%, dan F1-score 76,66%. Hasil ini menunjukkan bahwa model ResNeXt mampu mengenali pola manipulasi wajah pada video dengan cukup baik. Nilai presisi yang relatif tinggi menunjukkan bahwa prediksi kelas fake dilakukan secara selektif, sedangkan nilai recall yang lebih rendah mengindikasikan masih terdapat beberapa video *Deepfake* yang belum terdeteksi dengan benar. Hal ini menunjukkan bahwa model yang hanya mengandalkan analisis spasial pada setiap frame

masih memiliki keterbatasan dalam menangkap hubungan antarframe, sehingga pola manipulasi yang bersifat temporal belum dapat dikenali secara optimal.



Gambar 9 Grafik Pelatihan Model ResNeXt



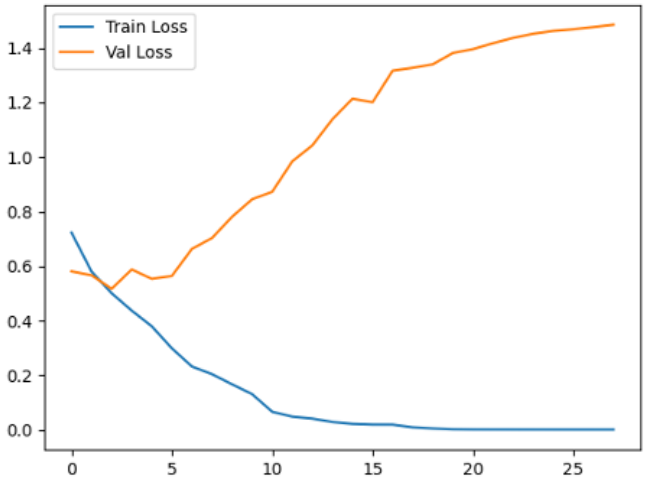
Gambar 10 Confusion Matrix Model ResNeXt

3.2 Model ResNeXt dengan LSTM

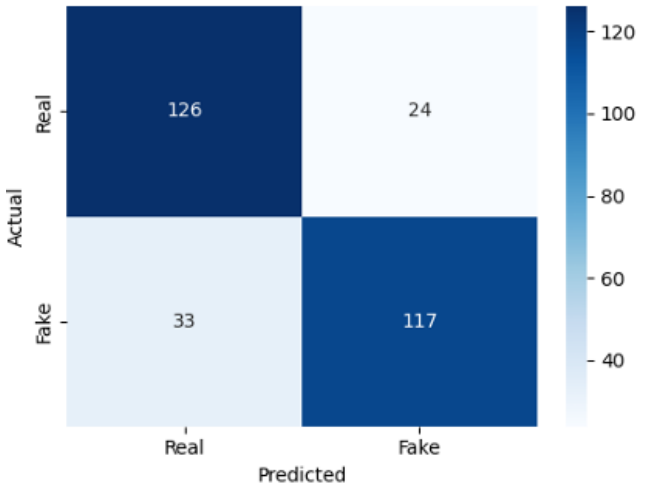
Pelatihan model ResNeXt dengan LSTM dilakukan untuk memanfaatkan informasi spasial dan temporal pada data video. Fitur spasial dari setiap *frame* terlebih dahulu diekstraksi menggunakan ResNeXt, kemudian disusun sebagai urutan sepanjang 100 *frame*. Vektor fitur berdimensi 2048 direduksi menjadi 512 sebelum diproses oleh LSTM dengan 2 *layer*, ukuran *hidden state* 256, dan *dropout* sebesar 0,30. Proses pelatihan menggunakan *optimizer* AdamW dengan *learning rate* 0,0001, *batch size* 32, *weight decay* 0,0001, *gradient clipping* sebesar 1,0, serta penyesuaian *learning rate* menggunakan *ReduceLROnPlateau*. Mekanisme *early stopping* diterapkan dengan *patience* 15 *epoch.*

Berdasarkan Gambar 11, nilai *training loss* menurun secara konsisten selama proses pelatihan. Sementara itu, *validation loss* mencapai nilai terendah pada *epoch* ke-13 dengan *F1-score* sebesar 0,7911 pada *threshold* 0,082. Setelah *epoch* tersebut, *validation loss* meningkat meskipun *training loss* terus menurun, yang mengindikasikan mulai terjadinya *overfitting*. Oleh karena itu, model pada *epoch* ke-13 dipilih sebagai model terbaik, dan proses pelatihan dihentikan secara otomatis pada *epoch* ke-28 melalui mekanisme *early stopping*.

Evaluasi pada dataset uji menghasilkan akurasi sebesar 81,00%, presisi 82,98%, recall 78,00%, dan F1-score 80,41%. Hasil ini menunjukkan bahwa model ResNeXt dengan LSTM mampu mengenali video real dan *Deepfake* dengan tingkat kesalahan yang relatif rendah. Pemanfaatan informasi urutan frame membantu model dalam mengenali pola manipulasi wajah yang tidak selalu terlihat jelas pada satu frame tunggal. Hal ini disebabkan karena manipulasi *Deepfake* tidak hanya terjadi pada satu frame, tetapi juga menimbulkan ketidakkonsistenan antarframe, seperti perubahan ekspresi yang tidak stabil atau pergerakan wajah yang kurang sinkron. Dengan memanfaatkan LSTM, model dapat menangkap pola perubahan tersebut secara berurutan, sehingga meningkatkan kemampuan dalam membedakan video asli dan hasil manipulasi. Meskipun demikian, masih terdapat beberapa kesalahan klasifikasi yang terjadi. Kesalahan ini dapat disebabkan oleh kemiripan visual antara video asli dan *Deepfake* dengan kualitas tinggi, serta variasi kondisi seperti pencahayaan, sudut pengambilan gambar, dan resolusi video yang memengaruhi proses ekstraksi fitur.



Gambar 11 Grafik Pelatihan Model ResNext dengan LSTM



Gambar 12 Confusion Matrix Model ResNeXt dengan LSTM

Tabel 1 menunjukkan bahwa model yang mengintegrasikan ResNeXt dan LSTM memperoleh kinerja lebih baik dibandingkan model ResNeXt. Peningkatan terlihat pada akurasi dari 77,67% menjadi 81,00%, presisi dari 80,29% menjadi 82,98%, recall dari 73,33% menjadi 78,00%, serta F1-score dari 76,66% menjadi 80,41%. Hasil ini menunjukkan bahwa penambahan LSTM memberikan kontribusi dalam memanfaatkan informasi temporal antarframe sehingga meningkatkan performa deteksi *Deepfake*.

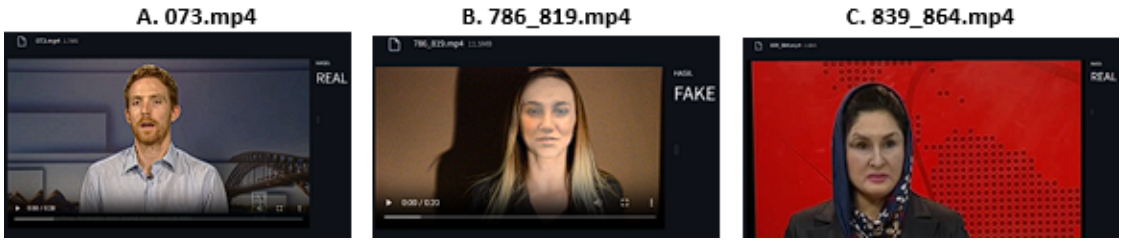
Tabel 1 Hasil Evaluasi Model

Metrik	ResNeXt	ResNeXt + LSTM
Akurasi	77,67%	81,00%
Presisi	80,29%	82,98%
Recall	73,33%	78,00%
F1-Score	76,66%	80,41%

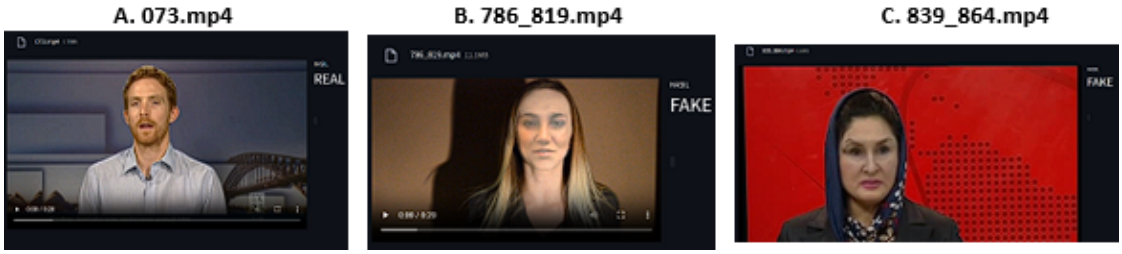
3.3 Pengujian Model

Pengujian dilakukan untuk menilai kinerja sistem deteksi video *Deepfake* setelah seluruh komponen pemrosesan terintegrasi. Pada tahap ini, sistem diuji menggunakan video uji yang tidak terlibat dalam proses pelatihan dan validasi, dengan tujuan memastikan bahwa alur pemrosesan video hingga pemberian prediksi dapat berjalan secara utuh. Setiap video uji diproses melalui ekstraksi frame dan preprocessing citra wajah sebelum dilakukan ekstraksi fitur menggunakan ResNeXt. Pengujian dilakukan menggunakan dua konfigurasi model, yaitu ResNeXt dan ResNeXt yang dikombinasikan dengan LSTM, untuk mengamati perbedaan hasil prediksi pada kondisi pengujian yang sama. Hasil prediksi ditampilkan dalam bentuk label kelas real atau fake melalui antarmuka sistem.

Berdasarkan hasil pengujian pada Gambar 13 dan Gambar 14, model ResNeXt menghasilkan satu kesalahan prediksi pada video C yang memiliki label asli *fake*, sedangkan model ResNeXt dengan LSTM memberikan prediksi yang sesuai dengan label asli pada seluruh video uji. Perbedaan hasil ini menunjukkan bahwa pemanfaatan informasi urutan frame membantu model dalam mengenali pola manipulasi yang tidak selalu terlihat pada satu frame. Pengujian lanjutan menggunakan 10 video uji dilakukan untuk melihat konsistensi hasil prediksi pada masing-masing konfigurasi model, yang selanjutnya dirangkum dalam tabel hasil pengujian.



Gambar 13 Hasil Pengujian Model ResNeXt Menggunakan 3 Video

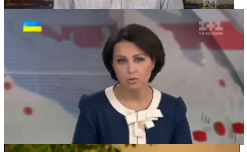
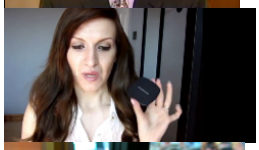
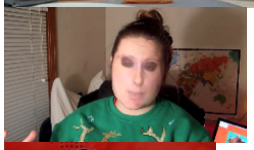
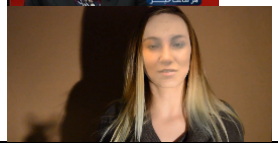


Gambar 14 Hasil Pengujian Model ResNeXt LSTM Menggunakan 3 Video

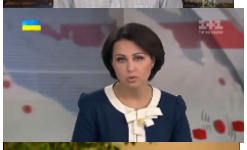
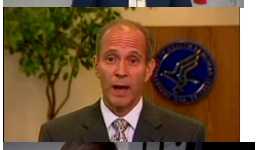
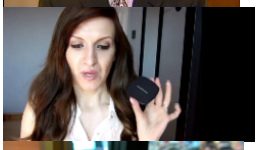
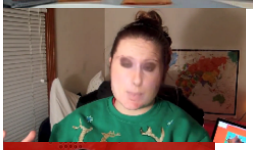
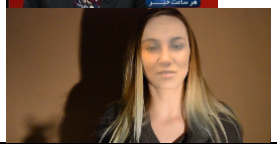
Berdasarkan hasil pengujian tersebut, model ResNeXt pada Tabel 2 menghasilkan prediksi yang benar pada 6 dari 10 video uji, sedangkan model ResNeXt dengan LSTM pada Tabel 3 menghasilkan prediksi yang benar pada 9 dari 10 video uji. Hasil ini menunjukkan bahwa pemanfaatan informasi urutan frame membantu sistem dalam menghasilkan prediksi yang lebih stabil. Secara keseluruhan, hasil pengujian menunjukkan bahwa sistem mampu melakukan proses deteksi video *Deepfake*. Integrasi ResNeXt dengan LSTM memberikan hasil prediksi yang lebih konsisten dibandingkan dengan penggunaan analisis spasial saja, sehingga sistem dapat

digunakan untuk mendukung proses deteksi video *Deepfake* berbasis wajah pada skenario pengujian langsung.

Tabel 2 Hasil Pengujian Model ResNeXt dengan 10 Video

No.	Visual	Jenis Video	Label Asli	Prediksi Model	Status
1		322.mp4	Real	Fake	Gagal
2		073.mp4	Real	Real	Berhasil
3		103.mp4	Real	Real	Berhasil
4		432.mp4	Real	Real	Berhasil
5		104.mp4	Real	Fake	Gagal
6		274_412.mp4	Fake	Fake	Berhasil
7		550_452.mp4	Fake	Real	Gagal
8		032_944.mp4	Fake	Fake	Berhasil
9		839_864.mp4	Fake	Real	Gagal
10		786_819.mp4	Fake	Fake	Berhasil

Tabel 3 Hasil Pengujian Model ResNeXt LSTM dengan 10 Video

No.	Visual	Jenis Video	Label Asli	Prediksi Model	Status
1		322.mp4	Real	Real	Berhasil
2		073.mp4	Real	Real	Berhasil
3		103.mp4	Real	Real	Berhasil
4		432.mp4	Real	Real	Berhasil
5		104.mp4	Real	Fake	Gagal
6		274_412.mp4	Fake	Fake	Berhasil
7		550_452.mp4	Fake	Fake	Berhasil
8		032_944.mp4	Fake	Fake	Berhasil
9		839_864.mp4	Fake	Fake	Berhasil
10		786_819.mp4	Fake	Fake	Berhasil

4. KESIMPULAN

Penelitian ini mengembangkan sistem deteksi video *Deepfake* berbasis wajah dengan memanfaatkan ResNeXt untuk menangkap informasi spasial pada setiap frame dan *Long Short-Term Memory* (LSTM) untuk memanfaatkan informasi urutan frame video. Evaluasi dilakukan pada data uji yang tidak terlibat dalam proses pelatihan dan validasi menggunakan metrik akurasi, presisi, recall, dan F1-score.

Hasil evaluasi menunjukkan bahwa model ResNeXt mampu memberikan kinerja awal yang baik dalam mengenali karakteristik visual wajah pada video *Deepfake*, dengan capaian akurasi sebesar 77,67%, presisi 80,29%, recall 73,33%, dan F1-score 76,66%. Hasil ini digunakan sebagai patokan kinerja untuk menggambarkan kemampuan deteksi berbasis informasi spasial pada frame video.

Berdasarkan patokan tersebut, penerapan LSTM pada arsitektur ResNeXt menghasilkan kinerja deteksi yang lebih stabil dengan capaian akurasi sebesar 81,00%, presisi 82,98%, recall 78,00%, dan F1-score 80,41%. Hasil ini menunjukkan bahwa pemanfaatan informasi urutan frame membantu sistem dalam menangkap pola manipulasi wajah yang tidak selalu terlihat jelas pada satu frame tunggal, sehingga mendukung peningkatan kualitas prediksi pada data uji.

Meskipun demikian, penelitian ini masih memiliki keterbatasan karena pengujian dilakukan pada *dataset* dengan karakteristik tertentu, sehingga kemampuan generalisasi model pada kondisi nyata masih perlu diuji lebih lanjut. Selain itu, peningkatan performa yang diperoleh menunjukkan bahwa pemanfaatan informasi temporal belum sepenuhnya optimal dan masih berpotensi untuk ditingkatkan. Penelitian selanjutnya dapat diarahkan untuk menguji sistem pada *dataset* tambahan serta variasi kualitas video yang lebih beragam guna meningkatkan ketahanan model dalam berbagai kondisi.

DAFTAR PUSTAKA

Abdu, F. J., Zhang, Y., Fu, M., Li, Y., & Deng, Z. (2021). Application of deep learning on millimeter-wave radar signals: A review. *Sensors*, *21*(6), 1–45. https://doi.org/10.3390/s21061951

Badruzzaman, A., & Arymurthy, A. M. (2024). A Comparative Study of Convolutional Neural Network in Detecting Blast Cells for Diagnose Acute Myeloid Leukemia. *Journal of Electronics, Electromedical Engineering, and Medical Informatics*, *6*(1), 84–91. https://doi.org/10.35882/jeemi.v6i1.354

Diponegoro, M. H., Kusumawardani, S. S., & Hidayah, I. (2021). Tinjauan Pustaka Sistematis: Implementasi Metode Deep Learning pada Prediksi Kinerja Murid (Implementation of Deep Learning Methods in Predicting Student Performance: A Systematic Literature Review). *Jurnal Nasional Teknik Elektro Dan Teknologi Informasi* |, *10*(2), 131–137.

Fernandes, Y. A., & Fatma, Y. (2025). METODE DEEP LEARNING DALAM TEKNOLOGI DEEPFAKE : SYSTEMATIC LITERATURE REVIEW. In *Jurnal Mahasiswa Teknik Informatika)* (Vol. 9, Number 2).

Gandrova, S., & Banke, R. (2023). Penerapan Hukum Positif Indonesia Terhadap Kasus Kejahatan Dunia Maya Deepfake. *Madani: Jurnal Ilmiah Multidisiplin*, *1*(10), 650–657. https://doi.org/10.5281/zenodo.10201140

Hakim, L., Sobri, A., Sunardi, L., & Nurdiansyah, D. (2025). Prediksi penyakit jantung berbasis mesin learning dengan menggunakan metode k-nn. *Jurnal Digital Teknologi Informasi*, *7*(2), 14–20. https://doi.org/10.32502/digital.v7i2.9429

Jefiza, A., Zarma Putra, I., Budiana, Karlina Laila Nur Suciningtyas, I., Siregar, L., Rika Puspita, W., Bestario Harlan, F., Assegaf, I., & Hitmen Marpaung, R. (2023). Klasifikasi Wajah Manusia Menggunakan Multi Layer Perceptron. *Jurnal Integrasi*, *15*(2), 142–148.

Jiwani, F. A., & Satrio Waluyo Poetro, B. (2025). SISTEM DETEKSI GAMBAR DEEPFAKE MENGGUNAKAN CNN DENSENET-121 DENGAN WATERMARKING LEAST SIGNIFICANT BIT (LSB). *Jurnal Rekayasa Sistem Informasi Dan Teknologi*, *2*(3).

Karandikar, A. (2020). Deepfake Video Detection Using Convolutional Neural Network. *International Journal of Advanced Trends in Computer Science and Engineering*, *9*(2), 1311–1315. https://doi.org/10.30534/ijatcse/2020/62922020

Kholifatullah, B. A. H., & Prihanto, A. (2023). Penerapan Metode Long Short Term Memory Untuk Klasifikasi Pada Hate Speech. *Journal of Informatics and Computer Science*, *04*(3), 292–297.

Kularkar, T., Jikar, T., Rewaskar, V., Dhawale, K., Thomas, A., & Madankar, M. (2023). Deepfake Detection Using LSTM and ResNext. *INTERNATIONAL JOURNAL OF CREATIVE RESEARCH THOUGHTS*, *11*(11), 78–86. www.ijcrt.org

Maksutov, A. A., Morozov, V. O., Lavrenov, A. A., & Smirnov, A. S. (2020, January 27). Methods

of Deepfake Detection Based on Machine Learning. *2020 IEEE Conference of Russian Young Researchers in Electrical and Electronic Engineering (EIConRus)*.

Nguyen, T. T., Nguyen, Q. V. H., Nguyen, D. T., Nguyen, D. T., Huynh-The, T., Nahavandi, S., Nguyen, T. T., Pham, Q.-V., & Nguyen, C. M. (2022). Deep Learning for Deepfakes Creation and Detection: A Survey. *Computer Vision and Image Understanding*, *223*, 1–19. https://doi.org/10.1016/j.cviu.2022.103525

Patil, K., Kale, S., Dhokey, J., & Gulhane, A. (2023). DEEPFAKE DETECTION USING BIOLOGICAL FEATURES: A SURVEY. *ArXiv*, 1–18. http://arxiv.org/abs/2301.05819

Rossler, A., Cozzolino, D., Verdoliva, L., Riess, C., Thies, J., & Niessner, M. (2019). FaceForensics++: Learning to Detect Manipulated Facial Images. *Proceedings of the IEEE International Conference on Computer Vision*, 1–11. https://doi.org/10.1109/ICCV.2019.00009

Rosyd, A., Irma Purnamasari, A., & Ali, I. (2024). PENERAPAN METODE LONG SHORT TERM MEMORY (LSTM) DALAM MEMPREDIKSI HARGA SAHAM PT BANK CENTRAL ASIA. *Jurnal Mahasiswa Teknik Informatika*, *8*(1).

Son, S., Lee, J., Min, K., & Kim, W. (2023). Enhancing Deepfake Detection: Spatial-Temporal Preprocessing and Self-Attention ResI3D Model. *Proceedings of the 2023 6th Artificial Intelligence and Cloud Computing Conference*, 27–35. https://doi.org/10.1145/3639592.3639597

Vakili, M., Ghamsari, M., & Rezaei, M. (n.d.). *Performance Analysis and Comparison of Machine and Deep Learning Algorithms for IoT Data Classification*.

Wang, H., Li, M., & Yue, X. (2021). IncLSTM: Incremental Ensemble LSTM Model towards Time Series Data. *Computers and Electrical Engineering*, *92*, 107156. https://doi.org/https://doi.org/10.1016/j.compeleceng.2021.107156

Yeh, J., Chan, H. T., & Hsia, C. H. (2021, November). ResNeXt with Cutout for Finger Vein Analysis. *ISPACS 2021 - International Symposium on Intelligent Signal Processing and Communication Systems: 5G Dream to Reality, Proceeding*. https://doi.org/10.1109/ISPACS51563.2021.9650921