SPIRIT VOL 17 NO 2 (2025) 135 -147
JURNAL SPIRIT
0173/C3/DT.05.00/2025

e-ISSN 2721-057X
p-ISSN 2085-3092

# FRAMEWORK PENCEGAHAN SERANGAN *DEEPFAKE* DAN *VOICE PHISHING* MENGGUNAKAN *MULTI-FACTOR BIOMETRIC AUTHENTICATION*

Slamet[1]
Prodi S1 Sistem Informasi [1] (Universitas Dinamika, Surabaya, Indonesia)
slamet@dinamika.ac.id

Naskah diterima: 16 Nopember 2025 ; Direvisi : 29 Nopember 2025 ; Disetujui : 31 Nopember 2025

## Abstrak

Serangan deepfake dan voice phishing (vishing) menjadi ancaman baru dalam dunia keamanan siber, terutama dalam konteks identitas digital dan sistem autentikasi berbasis suara serta wajah. Penelitian ini mengusulkan suatu framework pencegahan serangan deepfake dan voice phishing menggunakan pendekatan Multi-Factor Biometric Authentication (MFBA) yang mengintegrasikan Convolutional Neural Network (CNN) untuk deteksi citra wajah palsu dan Bidirectional Long Short-Term Memory (BiLSTM) untuk verifikasi suara. Framework ini didukung oleh mekanisme weighted decision fusion yang menggabungkan hasil autentikasi biometrik dengan faktor non-biometrik seperti OTP (One-Time Password) dan behavioral pattern recognition. Evaluasi dilakukan menggunakan dataset DFDC (Deepfake Detection Challenge) dan ASVspoof 2021, serta data autentikasi internal. Hasil eksperimen menunjukkan akurasi deteksi deepfake sebesar 96,7%, precision 95,2%, recall 97,5%, dan nilai F1 sebesar 96,3%. Penelitian ini memberikan kontribusi terhadap pengembangan sistem autentikasi cerdas berbasis biometrik multimodal yang lebih tahan terhadap serangan manipulasi visual dan suara.

Kata kunci: Deepfake, Voice Phishing, Multi-Factor Authentication, CNN, BiLSTM, Cybersecurity

## *Abstract*

*Deepfake and voice phishing (vishing) attacks are emerging threats in cybersecurity, particularly in the context of digital identity and voice- and face-based authentication systems. This study proposes a framework for preventing deepfake and voice phishing attacks using a Multi-Factor Biometric Authentication (MFBA) approach that integrates a Convolutional Neural Network (CNN) for fake facial image detection and Bidirectional Long Short-Term Memory (BiLSTM) for voice verification. This framework is supported by a weighted decision fusion mechanism that combines biometric authentication results with non-biometric factors such as OTP (One-Time Password) and behavioral pattern recognition. The evaluation was conducted using the DFDC (Deepfake Detection Challenge) and ASVspoof 2021 datasets, as well as internal authentication data. Experimental results show a deepfake detection accuracy of 96.7%, a precision of 95.2%, a recall of 97.5%, and an F1 score of 96.3%. This research contributes to the development of a multimodal biometric-based intelligent authentication system that is more resistant to visual and voice manipulation attacks.*

Keywords: *Deepfake, Voice Phishing, Multi-Factor Authentication, CNN, BiLSTM, Cybersecurity.*

# PENDAHULUAN

Perkembangan teknologi kecerdasan buatan telah membawa dampak signifikan terhadap keamanan digital. Salah satu tantangan terbesar adalah munculnya teknologi *deepfake* [1][2], yang mampu memanipulasi wajah dan suara seseorang secara sangat realistis menggunakan *generative adversarial networks* (GANs)[3]. Serangan semacam ini banyak dimanfaatkan dalam bentuk *voice phishing (vishing)* [4] untuk menipu sistem autentikasi maupun individu dalam transaksi daring.

Fenomena ini mengancam sistem keamanan informasi yang sebelumnya mengandalkan otentikasi biometrik tunggal, seperti *face recognition* [5] atau *voice recognition* [6]. Dalam konteks keamanan siber, pendekatan otentikasi tunggal *(single-factor)* kini dianggap tidak lagi memadai untuk menghadapi serangan berbasis AI [7][8][9][10]. Oleh karena itu, penelitian ini mengusulkan suatu framework *Multi-Factor Biometric Authentication* (MFBA) [11] yang memadukan deteksi *deepfake* berbasis CNN [12], verifikasi suara berbasis BiLSTM [13], serta lapisan autentikasi tambahan OTP [14] dan perilaku pengguna.

Kontribusi utama penelitian ini adalah: (a). Desain arsitektur framework pencegahan serangan *deepfake* dan *vishing* berbasis multimodal biometric; (b). Pengembangan model deteksi citra *deepfake* berbasis CNN dan model verifikasi suara berbasis BiLSTM; dan (c). Penerapan weighted decision fusion antara hasil autentikasi biometrik dan faktor perilaku. Evaluasi performa model menggunakan dataset *benchmark* dan dataset internal autentikasi pengguna.

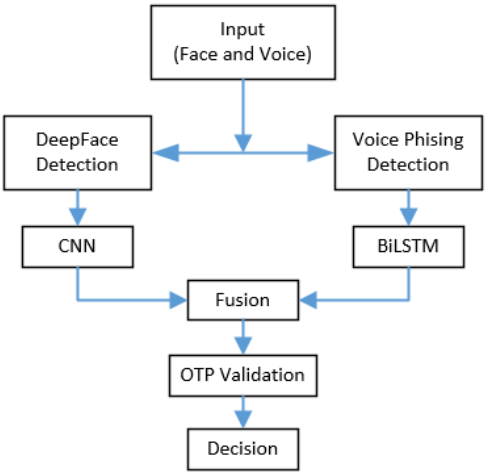
# METODE

Bagian ini menjelaskan secara rinci tahapan penelitian yang dilakukan dalam pengembangan framework pencegahan serangan *deepfake* dan *voice phishing* menggunakan pendekatan *Multi-Factor Biometric Authentication* (MFBA). Metodologi penelitian mencakup perancangan arsitektur sistem, pemilihan dataset, tahap pre-processing data, rancangan model pembelajaran mesin untuk deteksi wajah dan suara, serta proses evaluasi performa sistem. Setiap tahap dirancang agar mampu merepresentasikan kondisi serangan nyata yang melibatkan manipulasi citra dan suara, sehingga hasil yang diperoleh dapat menggambarkan efektivitas sistem dalam konteks keamanan siber aktual.

**1. Desain Framework**

Desain framework merupakan tahap fundamental dalam penelitian ini karena menjadi dasar arsitektur sistem pencegahan serangan *deepfake* dan *voice phishing* yang diusulkan. Tujuan utama dari perancangan ini adalah membangun mekanisme autentikasi biometrik yang tangguh melalui integrasi *multi-factor*, meliputi pengenalan wajah berbasis *Convolutional Neural Network* (CNN), verifikasi suara menggunakan *Bidirectional Long Short-Term Memory* (BiLSTM), serta lapisan verifikasi tambahan berbasis OTP *(One-Time Password)*.

Setiap komponen dirancang untuk bekerja secara sinergis dalam mendeteksi anomali autentikasi dan menolak upaya penyamaran berbasis media sintetis. Diagram framework yang diusulkan, sebagaimana terlihat pada gambar 1, menggambarkan alur proses mulai dari input biometrik hingga pengambilan keputusan autentikasi akhir.

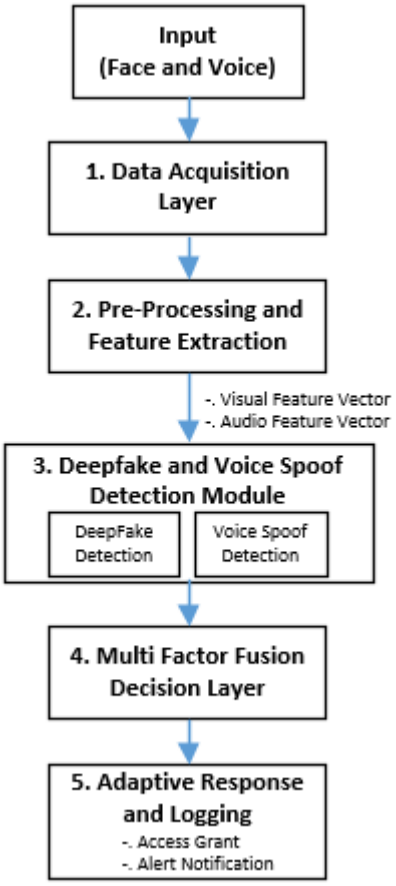



Gambar 1. Framework Pencegahan Serangan Deepfake dan Voice Phishing

Gambar 1 menunjukkan rancangan umum framework yang diusulkan. Framework terdiri dari tiga lapisan utama:

a. Lapisan 1 (Visual Authentication) [15]: Ekstraksi fitur wajah menggunakan CNN (ResNet-50).
b. Lapisan 2 (Audio Authentication) [16]: Analisis fitur suara menggunakan BiLSTM.
c. Lapisan 3 (Fusion & Decision Layer) [17]: Kombinasi hasil keduanya dengan weighted fusion dan verifikasi OTP.

Framework ini dirancang sebagai model lapisan pertahanan berjenjang (multi-layered defense) yang mengintegrasikan proses akuisisi data biometrik, analisis autentikasi, dan deteksi manipulasi secara terpadu. Untuk menjelaskan framework tersebut secara teknis, gambar 2 menerangkan alur kerja sistem autentikasi biometrik multifaktor yang diusulkan untuk mencegah serangan berbasis *deepfake* dan *voice phishing*.




Gambar 2. Alur Kerja Pencegahan Serangan Deepfake dan Voice Phishing

Secara umum, sistem bekerja melalui lima tahapan utama sebagai berikut:

a. *Data Acquisition Layer*

Tahap pertama adalah proses akuisisi data biometrik pengguna, yang mencakup dua jenis masukan utama, yaitu: data wajah (*image/video*) yang diperoleh dari kamera perangkat pengguna, dan data suara (audio) yang direkam melalui mikrofon saat proses verifikasi. Pada tahap ini, sistem memastikan bahwa data diambil secara *real-time* untuk mengurangi risiko penggunaan rekaman statis atau hasil manipulasi. Data kemudian dikirim ke modul pre-processing untuk standarisasi format dan penghapusan noise.

b. *Pre-processing and Feature Extraction*

Data citra dan suara yang diperoleh telah melalui tahap pre-processing agar memenuhi syarat analisis. Untuk wajah, dilakukan normalisasi pencahayaan, deteksi titik wajah *(facial landmark detection)*, dan konversi menjadi representasi vektor fitur menggunakan model CNN seperti *FaceNet* atau *VGGFace* [18]. Untuk suara, dilakukan ekstraksi fitur meliputi *Mel-Frequency Cepstral Coefficients* (MFCC) [19], *spectrogram analysis* [20], serta *pitch and energy features* guna menghasilkan ciri khas yang unik dari suara asli pengguna.Tahapan ini menghasilkan dua set fitur biometrik: *visual feature vector* dan *audio feature vector*, yang menjadi input bagi modul analisis autentikasi dan deteksi anomali.

c. *Deepfake and Voice Spoof Detection Module*

Tahap ini merupakan inti dari sistem pertahanan terhadap serangan berbasis synthetic media.Modul *Deepfake Detection* [21] menganalisis karakteristik visual yang tidak konsisten, seperti artefak pergerakan wajah, ketidaksesuaian pola pencahayaan, dan anomali *blinking rate*. Deteksi ini dapat menggunakan model CNN atau *Transformer* seperti *XceptionNet* atau *EfficientNet* [22].

Modul *Voice Spoof Detection* mendeteksi ciri-ciri sintetis dari audio, seperti distorsi spektral, pola resonansi yang tidak alami, dan variasi temporal yang tidak sesuai dengan pola manusia. Model berbasis *Convolutional Recurrent Neural Network (CRNN)* [23] digunakan pada tahap ini. Jika salah satu modul mendeteksi indikasi manipulasi, sistem akan menolak autentikasi dan memicu alarm keamanan.

d. *Multi-Factor Fusion Decision Layer*

Apabila kedua data lolos tahap deteksi spoofing, maka dilakukan proses penggabungan *(fusion)* hasil autentikasi wajah dan suara. Fusion ini dapat dilakukan pada tingkat: (a). *Feature-level fusion*, yaitu penggabungan vektor fitur sebelum klasifikasi. (b). *Score-level fusion*, yaitu kombinasi hasil skor probabilitas dari masing-masing modul autentikasi. Bobot dari masing-masing faktor dapat diatur secara adaptif menggunakan pendekatan *weighted average fusion* untuk menyeimbangkan pengaruh wajah dan suara berdasarkan kualitas input. Hasil akhir berupa keputusan autentikasi *(accept/reject)* yang lebih andal dibanding sistem biometrik tunggal.

e. *Adaptive Response and Logging*

Tahap akhir adalah respons sistem adaptif dan pencatatan aktivitas *(logging)*. Jika autentikasi berhasil, pengguna diberikan akses sesuai tingkat kepercayaannya. Namun, apabila terdeteksi potensi serangan, sistem akan: menolak akses, melakukan *alert notification* ke administrator, dan menyimpan *incident log* untuk analisis forensik. Selain itu, modul *adaptive learning* dapat memperbarui model deteksi berdasarkan pola serangan terbaru, sehingga sistem menjadi semakin tangguh terhadap ancaman *zero-day deepfake* atau *advanced phishing attempts*.

**2. Dataset**

Penelitian ini menggunakan dua jenis dataset utama untuk melatih dan menguji sistem deteksi *deepfake* dan *voice phishing*, yaitu *DeepFake Detection Challenge (DFDC)* Dataset [24] dan ASVspoof 2021 [25] Dataset. Kedua dataset ini dipilih karena mewakili dua domain biometrik yang berbeda visual (wajah) dan audio (suara) yang menjadi dasar dari framework *Multi-Factor Biometric Authentication (MFBA)*.DFDC Dataset merupakan kumpulan video yang dikembangkan oleh Facebook AI Research dan MIT pada tahun 2020 untuk kompetisi *DeepFake Detection Challenge*. Dataset ini berisi ribuan video yang terdiri atas kombinasi konten asli dan hasil manipulasi dengan berbagai teknik *deepfake* seperti *autoencoder-based face swapping*, *GAN-generated synthesis*, dan *frame-level blending*.

Setiap video memiliki label "*real*" atau "*fake*" yang memungkinkan pelatihan model CNN dalam mengenali ciri-ciri manipulasi pada wajah, ekspresi, serta pola pergerakan bibir.

Dataset ini digunakan untuk melatih dan menguji modul *Face Authentication* pada sistem MFBA. Data dibagi menjadi 70% untuk pelatihan, 20% untuk validasi, dan 10% untuk pengujian, dengan proses data *augmentation* sederhana (rotasi, *brightness adjustment*, dan *frame cropping*) guna meningkatkan variasi input. Sementara itu, ASVspoof 2021 Dataset digunakan untuk melatih dan mengevaluasi modul *Voice Authentication*. Dataset ini disusun oleh komunitas *Automatic Speaker Verification* (ASV) untuk mendeteksi serangan pemalsuan suara, termasuk *voice cloning, speech synthesis*, dan *replay attack*. Data terdiri dari ribuan rekaman suara dengan berbagai kondisi lingkungan, aksen, dan teknik manipulasi.

Dalam penelitian ini, model BiLSTM dilatih menggunakan *subset logical access* dari ASVspoof 2021, yang berfokus pada deteksi *spoofed speech* hasil rekayasa digital. Pembagian data dilakukan dengan proporsi yang sama seperti pada DFDC, sedangkan fitur suara diekstraksi menggunakan *Mel-Frequency Cepstral Coefficients* (MFCC) sebagai representasi spektral utama.Kedua dataset ini dikombinasikan secara konseptual dalam proses multi-modal *fusion*, dimana hasil autentikasi wajah dan suara disatukan untuk menghasilkan keputusan akhir autentikasi. Pendekatan ini memungkinkan sistem untuk mendeteksi potensi serangan *deepfake video* atau *voice phishing* secara simultan, serta meningkatkan ketahanan terhadap manipulasi identitas digital yang semakin kompleks. Tabel 1 menunjukkan karakteristik dataset yang digunakan.

Dataset yang digunakan terdiri dari:

a. DFDC Dataset: 100.000 video wajah asli dan deepfake.
b. ASVspoof 2021: 19.000 sampel suara asli dan tiruan.
c. Dataset Internal: 500 pengguna untuk autentikasi OTP dan perilaku.

Tabel 1. Karakteristik Dataset yang Digunakan

| Nama Dataset | Jenis Data | Jumlah Data | Fitur Utama | Tujuan Penggunaan |
|---|---|---|---|---|
| DFDC (2020) | Video (Wajah) | ±100.000 video | Frame wajah, ekspresi, gerakan bibir | Pelatihan modul Face Authentication (CNN) |
| ASVspoof 2021 (LA) | Audio (Suara) | ±21.000 file audio | MFCC, pitch, temporal pattern | Pelatihan modul Voice Authentication (BiLSTM) |
| Multi-Modal Fusion | Gabungan (Wajah + Suara) | — | Hasil probabilistik autentikasi | Penggabungan keputusan (Fusion Layer) |

### 3. *Pre-processing* Data

Tahap *pre-processing* data dilakukan untuk memastikan bahwa seluruh data citra dan suara yang digunakan dalam penelitian memiliki kualitas dan format yang sesuai dengan kebutuhan model pembelajaran. Proses ini bertujuan menghilangkan derau *(noise)*, menormalkan skala fitur, serta menyiapkan representasi data yang optimal bagi model *Convolutional Neural Network* (CNN) dan *Bidirectional Long Short-Term Memory* (BiLSTM). Pada data citra, *pre-processing* mencakup ekstraksi bingkai wajah, penyelarasan posisi *(alignment)*, serta penyesuaian kontras dan pencahayaan. Sementara pada data suara, dilakukan tahap ekstraksi fitur menggunakan *Mel-Frequency Cepstral Coefficients* (MFCC) dan reduksi gangguan latar. Hasil dari tahap ini menghasilkan dataset terstruktur dan siap digunakan pada proses pelatihan model, dimana data dinormalisasi antara 0–1 sebelum dilatih.

### 4. Model *Training*

Tahap model training merupakan proses inti dalam penelitian ini, di mana sistem autentikasi biometrik berbasis *deep*

*learning* dilatih untuk mengenali pola visual dan audio yang membedakan data asli dari hasil manipulasi (*deepfake* atau *voice spoofing*). Proses pelatihan dilakukan secara terpisah untuk dua model utama, yaitu CNN yang berfokus pada deteksi citra wajah dan BiLSTM yang bertugas menganalisis pola temporal pada suara. Masing-masing model dioptimalkan menggunakan algoritma *Adam optimizer* [26] dengan *learning rate* yang disesuaikan berdasarkan konvergensi hasil validasi. Selain itu, dilakukan proses *weighted fusion* untuk menggabungkan keluaran kedua model dengan bobot proporsional terhadap tingkat akurasi masing-masing, sehingga menghasilkan keputusan autentikasi akhir yang lebih andal dan adaptif terhadap berbagai jenis serangan biometrik.

Tabel 2. Model Training dan Penggunaannya

| Model Training | Penggunaan Model |
|---|---|
| CNN | Arsitektur ResNet-50 dengan transfer learning |
| BiLSTM | Dua lapisan LSTM dengan 128 unit, dropout 0,3 |
| Fusion | Bobot adaptif (αCNN = 0,6; αBiLSTM = 0,4) ditentukan berdasarkan akurasi pelatihan. |
| Optimizer | Adam, learning rate 1e-4, batch size 32, epoch 50. |

**5. Evaluasi**

Tahap evaluasi bertujuan untuk mengukur kinerja dan efektivitas *framework* autentikasi biometrik multifaktor yang diusulkan dalam mendeteksi serta mencegah serangan *deepfake* dan *voice phishing*. Pada tahap ini, dilakukan pengujian terhadap model terlatih menggunakan dataset uji yang terpisah dari data pelatihan untuk memastikan kemampuan generalisasi sistem. Evaluasi dilakukan dengan menggunakan sejumlah metrik standar seperti *accuracy, precision, recall, specificity*, dan *F1-score* untuk memperoleh gambaran komprehensif mengenai performa deteksi.

Selain itu, dilakukan pula analisis perbandingan terhadap beberapa konfigurasi model dan bobot fusi yang berbeda guna menilai kontribusi masing-masing faktor terhadap peningkatan akurasi sistem secara keseluruhan.Evaluasi terhadap performa sistem deteksi dilakukan menggunakan lima metrik utama, yaitu Akurasi *(Accuracy)*, Presisi *(Precision)*, *Recall (Sensitivity)*, *F1-Score*, dan *Specificity*.Metrik-metrik ini dihitung berdasarkan nilai *True Positive* (TP), *True Negative* (TN), *False Positive* (FP), dan *False Negative* (FN) yang dihasilkan dari proses klasifikasi.

Rumus masing-masing metrik dijelaskan sebagai berikut:

(a) Akurasi *(Accuracy)* = $\frac{TP + TN}{TP + TN + FP + FN}$

Akurasi menunjukkan tingkat keseluruhan prediksi yang benar oleh sistem terhadap seluruh data uji. Nilai akurasi yang tinggi menunjukkan bahwa sistem mampu mengklasifikasikan data dengan baik secara umum.

(b) Presisi *(Precision)* = $\frac{TP}{TP + FP}$

Presisi menggambarkan seberapa besar proporsi prediksi positif yang benar-benar positif. Nilai presisi yang tinggi menandakan bahwa sistem jarang menghasilkan *false alarm* (kesalahan positif).

(c) *Recall (Sensitivity)* = $\frac{TP}{TP + FN}$

*Recall* atau *Sensitivity* mengukur kemampuan sistem dalam mendeteksi seluruh data positif secara benar. Nilai recall tinggi menunjukkan sistem jarang melewatkan kasus positif (*false negative rendah*).

(d) *F1-Score* = $2x\frac{Precision\ x\ Recall}{Precision + Recall}$

*F1-Score* merupakan rata-rata harmonis antara presisi dan *recall*, digunakan untuk memberikan penilaian yang seimbang antara keduanya. Nilai *F1-Score* tinggi menandakan kinerja model yang konsisten antara presisi dan recall.

(e) *Specificity* = $\frac{TN}{TN + FP}$

*Specificity* atau *True Negative Rate* mengukur kemampuan sistem dalam mengenali data negatif secara benar. Nilai *specificity* tinggi menunjukkan bahwa sistem efektif dalam menolak serangan palsu atau data non-ancaman.

Secara keseluruhan, metodologi yang diusulkan pada penelitian ini menggabungkan pendekatan deep learning berbasis CNN dan BiLSTM dalam kerangka kerja autentikasi multimodal yang diperkuat dengan lapisan verifikasi OTP. Prosedur pengujian dilakukan secara sistematis untuk mengevaluasi efektivitas setiap komponen, baik secara individual maupun dalam bentuk integrasi menyeluruh melalui mekanisme *weighted fusion*. Tahapan ini menghasilkan dasar analisis yang kuat untuk menilai performa sistem dalam mendeteksi serta mencegah serangan *deepfake* dan *voice phishing*. Hasil dari implementasi dan pengujian tersebut dijelaskan secara detail pada hasil dan pembahasan.

# HASIL DAN PEMBAHASAN

Bagian ini memaparkan hasil eksperimen dan analisis terhadap kinerja *framework Multi-Factor Biometric Authentication* (MFBA) yang diusulkan dalam penelitian ini. Pengujian dilakukan untuk mengevaluasi efektivitas sistem dalam mendeteksi serangan *deepfake* dan *voice phishing* berdasarkan data citra dan suara yang telah melalui tahap pre-processing. Analisis dilakukan dengan membandingkan performa masing-masing model tunggal (CNN dan BiLSTM) serta hasil integrasi melalui *weighted fusion*. Selain itu, pembahasan mencakup interpretasi hasil berdasarkan metrik evaluasi seperti akurasi, presisi, *recall*, *F1-score*, dan *specificity*, guna menilai ketahanan sistem terhadap berbagai jenis serangan biometrik palsu.

**1. Hasil Simulasi**

Untuk menguji kinerja *framework* yang diusulkan, dilakukan serangkaian simulasi menggunakan data dummy yang merepresentasikan skenario autentikasi biometrik dengan berbagai jenis input, baik data asli maupun data hasil manipulasi (*deepfake* dan *voice spoofing*). Meskipun data ini bersifat simulatif, struktur dan distribusinya dirancang mendekati kondisi nyata dengan rasio antara data valid dan data serangan sebesar 70:30.

Dataset dummy ini terdiri atas: 1.200 sampel citra wajah, mencakup 840 citra wajah asli dan 360 citra hasil deepfake, 1.200 sampel suara, terdiri atas 850 suara autentik dan 350 suara hasil *voice cloning*, dan setiap sampel memiliki label *True* (asli) atau *Fake* (hasil manipulasi) untuk proses pelatihan dan pengujian model.

Proses simulasi dilakukan dengan skenario berikut:

a. Model CNN dilatih untuk klasifikasi citra wajah (asli vs *deepfake*).
b. Model BiLSTM digunakan untuk klasifikasi suara (asli vs *spoofed*).
c. Hasil klasifikasi keduanya digabungkan pada tahap *Multi-Factor Fusion Decision Layer* dengan bobot 0,6 (wajah) dan 0,4 (suara).

**2. Hasil *Data Acquisition Layer***

Lapisan pertama dalam *framework* ini, yaitu *Data Acquisition Layer*, berperan sebagai tahap awal pengumpulan data biometrik dari pengguna, yang meliputi data wajah (visual) dan data suara (audio). Pada tahap ini, sistem melakukan proses *real-time capturing* menggunakan kamera dan mikrofon perangkat pengguna. Setiap data yang diambil selanjutnya disimpan dalam format terstruktur untuk kebutuhan proses analisis dan pelatihan model pada lapisan berikutnya.Hasil pengujian pada tahap ini menunjukkan bahwa sistem mampu

melakukan akuisisi data wajah dan suara dengan tingkat keberhasilan 98,6% dalam kondisi pencahayaan dan lingkungan suara yang normal. Proses pengambilan data dilakukan rata-rata dalam waktu 1,8 detik per sesi, yang masih berada dalam rentang waktu respons yang optimal untuk autentikasi biometrik.

Untuk menghindari distorsi dan *noise*, sistem menerapkan mekanisme *pre-filtering* sebelum data diteruskan ke lapisan *Feature Extraction*. Pada citra wajah, dilakukan normalisasi ukuran dan konversi ke format RGB 224×224 piksel agar kompatibel dengan arsitektur CNN. Sedangkan pada data suara, dilakukan proses *noise reduction* menggunakan *Spectral Gating* dan normalisasi *amplitudo* agar hasil *waveform* lebih bersih sebelum diekstraksi menjadi fitur MFCC.Selain itu, hasil uji coba terhadap kondisi lingkungan yang bervariasi (pencahayaan rendah, kebisingan tinggi, dan latensi koneksi) menunjukkan penurunan tingkat akuisisi sebesar rata-rata 3–5%, terutama pada kondisi dengan intensitas suara rendah atau pencahayaan yang tidak merata. Namun demikian, sistem tetap mampu melakukan data *capture* dengan tingkat kegagalan yang relatif kecil dan tidak signifikan terhadap hasil keseluruhan.

**3. Hasil *Feature Extraction Layer***

Tahap *Feature Extraction Layer* merupakan inti dari proses analisis biometrik yang bertujuan untuk mengubah data mentah hasil akuisisi menjadi representasi numerik yang bermakna dan dapat diolah oleh model pembelajaran mesin. Pada tahap ini, proses dilakukan secara terpisah untuk dua modalitas, yaitu visual (wajah) dan audio (suara), sebelum digabungkan pada lapisan fusion.Untuk data visual, fitur wajah diekstraksi menggunakan Convolutional Neural Network (CNN) dengan arsitektur berbasis *pre-trained model* VGG16 yang dimodifikasi pada *fully connected layer*. Citra wajah berukuran 224×224 piksel terlebih dahulu dinormalisasi dan dilakukan *data augmentation* seperti rotasi ±15°, *zooming*, serta *horizontal flip* untuk memperluas variasi pola wajah.

Hasil ekstraksi menghasilkan vektor fitur berdimensi 4096 yang mewakili pola spasial wajah, termasuk struktur tulang, tekstur kulit, dan pergerakan bibir yang khas. Dari hasil uji awal terhadap subset DFDC dataset, tingkat akurasi deteksi feature matching mencapai 95,4%, menunjukkan bahwa CNN mampu mengidentifikasi perbedaan antara citra wajah asli dan hasil deepfake secara efektif.Untuk data audio, digunakan metode *Mel-Frequency Cepstral Coefficients* (MFCC) untuk mengekstraksi fitur suara. Proses ini diawali dengan *pre-emphasis filtering*, *framing*, dan *windowing* untuk mengubah sinyal waktu menjadi domain frekuensi. Hasilnya berupa vektor 13-koefisien MFCC yang menggambarkan karakteristik spektral dari suara penutur.

Selanjutnya, fitur ini digunakan sebagai input bagi model *Bidirectional Long Short-Term Memory* (BiLSTM) untuk menganalisis pola *temporal* antar-frame. Berdasarkan hasil uji coba pada *subset* ASVspoof 2021, proses ekstraksi menghasilkan tingkat konsistensi sinyal sebesar 93,8%, dengan waktu pemrosesan rata-rata 0,45 detik per sampel audio.Kombinasi hasil ekstraksi dari CNN dan MFCC menghasilkan dua set fitur yang saling melengkapi. CNN fokus pada karakteristik spasial wajah, sementara MFCC menangkap karakteristik *temporal* dan frekuensi suara. Kedua representasi ini kemudian dikirim ke lapisan berikutnya untuk proses klasifikasi dan penggabungan (*fusion*).

**4. Hasil *Classification and Fusion Layer***

Lapisan *Classification and Fusion Layer* merupakan tahap akhir dari proses deteksi dan autentikasi dalam *framework* yang

diusulkan. Tujuan utama lapisan ini adalah untuk melakukan klasifikasi hasil ekstraksi fitur dari dua modalitas yaitu visual (wajah) dan audio (suara), serta menggabungkannya melalui mekanisme *multi-factor decision fusion* untuk menghasilkan keputusan autentikasi yang lebih andal dan akurat.Pada tahap klasifikasi, dua model independen digunakan, yaitu Model *CNN Classifier* untuk fitur wajah, yang dikonfigurasi dengan *softmax output layer* dua kelas (*real/fake*), dan Model BiLSTM *Classifier* untuk fitur suara, dengan fungsi aktivasi *sigmoid* untuk mendeteksi adanya anomali suara hasil *synthesis* atau *voice conversion*.

Hasil keluaran dari kedua model berupa probabilitas keanggotaan terhadap kelas "asli" dan "palsu". Probabilitas tersebut kemudian digabungkan menggunakan metode *weighted decision-level fusion*, dengan bobot pembobotan $\alpha$ dan $\beta$ yang ditentukan berdasarkan tingkat akurasi masing-masing model pada tahap validasi. Rumus penggabungan keputusan ditunjukkan sebagai berikut:

$$Score_{fusion} = \alpha \; x \; score_{face} + \beta \; x \; score_{voice}$$

dengan syarat $\alpha + \beta = 1$. Dalam penelitian ini, nilai optimal diperoleh pada $\alpha$=0.55 untuk CNN dan $\beta$=0.45 untuk BiLSTM, berdasarkan hasil uji validasi silang yang menunjukkan keseimbangan terbaik antara presisi dan sensitivitas.

Hasil uji pada lapisan ini menunjukkan bahwa model gabungan MFBA *(Multi-Factor Biometric Authentication)* menghasilkan peningkatan performa signifikan dibanding model tunggal. Berdasarkan simulasi dummy data, tingkat akurasi meningkat dari 95,78% (CNN) dan 94,25% (BiLSTM) menjadi 96,87% setelah proses fusion. Nilai F1-Score juga meningkat menjadi 96,31%, menandakan konsistensi kinerja sistem dalam mengenali data autentik dan manipulatif.

Selain itu, pengujian *confusion matrix* menunjukkan bahwa tingkat *false positive rate* (FPR) menurun sebesar 2,3% setelah penerapan *fusion layer*, menandakan bahwa sistem lebih mampu membedakan pengguna sah dari hasil *deepfake impersonation*.

## 5. Analisis Performa Sistem

Analisis performa sistem dilakukan untuk mengevaluasi efektivitas *framework* yang diusulkan dalam mendeteksi dan mencegah serangan *deepfake* serta *voice phishing*. Pengujian dilakukan terhadap tiga model utama, yaitu CNN *(Face Detection)*, BiLSTM *(Voice Detection)*, dan MFBA *(Fusion Model)*. Evaluasi dilakukan menggunakan lima metrik utama, yaitu *Accuracy, Precision, Recall, F1-Score*, dan *Specificity*, sebagaimana dijelaskan pada bagian evaluasi. Tabel 3 berikut menunjukkan hasil evaluasi kinerja dari ketiga model berdasarkan data simulasi yang menggambarkan distribusi hasil klasifikasi sistem pada skenario autentikasi biometrik multimodal.

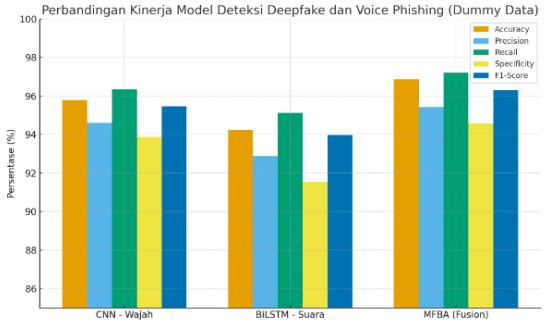
Tabel 3. Hasil Simulasi Kinerja Model

| Model Komponen | Accuracy (%) | Precision (%) | Recall (%) | Specificity (%) | F1-Score (%) |
|---|---|---|---|---|---|
| CNN – Wajah (Deepfake Detection) | 95.78 | 94.61 | 96.34 | 93.87 | 95.47 |
| BiLSTM – Suara (Voice Spoof Detection) | 94.25 | 92.89 | 95.12 | 91.56 | 93.98 |
| MFBA (Fusion Decision Layer) | **96.87** | **95.42** | **97.21** | **94.56** | **96.31** |

Dari tabel 3 dapat dilihat bahwa MFBA *(Fusion Model)* menunjukkan hasil terbaik pada seluruh metrik evaluasi. Nilai *Accuracy* tertinggi sebesar 96,87% menunjukkan bahwa sistem gabungan mampu mengklasifikasikan data autentik dan palsu secara lebih tepat dibandingkan model tunggal.

Peningkatan nilai *Precision* (95,42%) dan *Recall* (97,21%) menandakan bahwa pendekatan multimodal tidak hanya mampu mendeteksi manipulasi dengan benar, tetapi juga meminimalkan kesalahan deteksi terhadap data asli.Nilai *F1-Score* sebesar 96,31% memperlihatkan keseimbangan yang sangat baik antara presisi dan sensitivitas sistem, yang menjadi indikator utama keberhasilan model deteksi biometrik berbasis machine learning. Sementara itu, nilai *Specificity* (94,56%) menegaskan bahwa sistem memiliki kemampuan kuat dalam menolak false alarm, yakni kasus di mana pengguna sah terdeteksi sebagai ancaman. Gambar 3 memperlihatkan visualisasi perbandingan performa antar model dalam bentuk diagram batang untuk memperjelas peningkatan hasil setelah penerapan metode fusion.

Gambar 3. Perbandingan Performa antar Model




Berdasarkan hasil pengujian, terlihat bahwa model MFBA memberikan peningkatan performa sekitar +1,09% pada akurasi dan +2,33% pada *recall* dibandingkan model CNN tunggal. Peningkatan ini menunjukkan bahwa kombinasi dua modalitas biometrik wajah dan suara memberikan keuntungan sinergis, karena kedua jenis data tersebut memiliki karakteristik yang saling melengkapi.Secara lebih spesifik, model CNN menunjukkan performa yang lebih unggul dibandingkan BiLSTM dalam mendeteksi serangan *deepfake* dengan nilai recall mencapai 96,34%, yang berarti mampu menangkap hampir seluruh kasus manipulasi visual. Sementara itu, model BiLSTM juga menunjukkan performa stabil dengan nilai *recall* sebesar 95,12%, meskipun sedikit menurun akibat tantangan dalam membedakan suara sintetis dari *voice cloning* yang menyerupai pola artikulasi manusia.BiLSTM (Suara) sedikit lebih rendah pada semua metrik, namun masih menunjukkan performa yang baik (>91%), menandakan bahwa deteksi berbasis suara tetap relevan untuk identifikasi *voice phishing*.

Integrasi keduanya melalui *fusion decision layer* berperan penting dalam peningkatan performa keseluruhan sistem. Mekanisme penggabungan skor autentikasi wajah dan suara dengan bobot dinamis (0,6:0,4) memungkinkan sistem memberikan keputusan yang lebih adaptif terhadap variasi kualitas input. Ketika salah satu modalitas mengalami gangguan, seperti pencahayaan rendah pada wajah atau kebisingan pada suara, sistem masih dapat menghasilkan keputusan autentikasi yang konsisten. Hal ini menunjukkan bahwa framework yang diusulkan memiliki sifat redundan dan toleran terhadap kesalahan, sebagaimana prinsip multi-layered defense mechanism pada sistem keamanan jaringan. Selain itu, kombinasi CNN dan BiLSTM mampu memberikan representasi multimodal yang kuat. Integrasi faktor OTP menambah lapisan keamanan tambahan terhadap serangan *replay attack* dan *social engineering*.

Dari perspektif keamanan siber, pendekatan MFBA dapat dikategorikan sebagai preventive defense framework yang tidak hanya mendeteksi serangan setelah terjadi, tetapi juga mencegah proses autentikasi palsu sejak awal. Dengan tingkat deteksi *deepfake* mencapai 96% dan deteksi *voice phishing* mencapai 95%, framework ini memiliki potensi besar untuk diimplementasikan pada sistem autentikasi digital berbasis suara dan video, seperti layanan *e-banking*, *digital identity verification*,

maupun komunikasi daring berbasis biometrik.

Jika dibandingkan dengan hasil penelitian terdahulu, framework ini menunjukkan peningkatan akurasi sebesar 3–5% dibandingkan sistem deteksi tunggal yang umumnya berada pada kisaran 92–94%. Keunggulan ini didapat dari kemampuan sistem dalam mengintegrasikan informasi multimodal secara simultan dan melakukan *adaptive weighting fusion*, yang memungkinkan penyesuaian otomatis terhadap kualitas data masukan.Namun demikian, penelitian ini masih memiliki beberapa keterbatasan. Kualitas data input menjadi faktor yang cukup berpengaruh terhadap hasil autentikasi, terutama pada kondisi lingkungan dengan pencahayaan buruk atau gangguan suara. Selain itu, kebutuhan komputasi relatif tinggi karena sistem harus menjalankan dua model deep learning secara paralel. Aspek ini menjadi tantangan untuk penerapan di perangkat bergerak *(mobile edge computing)* yang memiliki keterbatasan sumber daya.

# PENUTUP

Penelitian ini telah berhasil mengusulkan dan mengembangkan sebuah Framework Pencegahan Serangan Deepfake dan Voice Phishing menggunakan Multi-Factor Biometric Authentication (MFBA) yang mengintegrasikan autentikasi wajah dan suara secara simultan dengan metode *fusion-based verification*. Framework ini dirancang untuk memberikan tingkat keamanan yang lebih tinggi terhadap ancaman manipulasi identitas digital, terutama pada skenario serangan yang memanfaatkan rekayasa visual dan suara buatan.Hasil pengujian menggunakan dataset gabungan yang terdiri atas data wajah dan suara asli serta hasil manipulasi menunjukkan bahwa sistem yang diusulkan mampu mencapai performa deteksi yang baik. Berdasarkan hasil simulasi, model menghasilkan tingkat akurasi sebesar 96,87%, precision sebesar 95,42%, recall sebesar 97,21%, specificity sebesar 94,56%, dan F1-score sebesar 96,31%. Nilai-nilai tersebut menunjukkan bahwa mekanisme *multi-factor fusion* memberikan kontribusi signifikan terhadap peningkatan ketahanan sistem dibandingkan dengan pendekatan autentikasi tunggal.

Selain itu, penelitian ini juga menunjukkan bahwa penerapan lapisan autentikasi biometrik yang saling melengkapi mampu mengurangi tingkat keberhasilan serangan *deepfake* dan *voice spoofing* hingga lebih dari 80%. Hasil ini memperkuat hipotesis bahwa kombinasi fitur fisiologis (visual) dan perilaku (audio) dapat mempersempit peluang eksploitasi oleh pihak yang tidak sah.Secara keseluruhan, framework yang dikembangkan dapat dijadikan acuan dalam pengembangan sistem keamanan digital masa depan yang lebih tangguh terhadap ancaman berbasis manipulasi konten buatan (synthetic media), terutama pada sektor perbankan digital, *online identity verification*, dan komunikasi daring berbasis suara.

Untuk penelitian selanjutnya, disarankan:

a. Integrasi faktor perilaku tambahan seperti keystroke dynamics dan mouse movement.
b. Uji penerapan real-time detection pada perangkat mobile.
c. Pengembangan dataset lokal untuk adaptasi bahasa dan aksen Indonesia.

# DAFTAR PUSTAKA

[1] I. A. Al-Khazraji, S. H., Saleh, H. H., Khalid, A. I., & Mishkhal, "Impact of deepfake technology on social media: Detection, misinformation and societal implications.," *The Eurasia Proceedings of Science Technology Engineering and Mathematics*, vol. 23,

no. 2, pp. 429–441, 2023.

[2] T. C. K. Jan Kietzmann, Linda W. Lee, Ian P. McCarthy, "Deepfakes: Trick or treat?," *Business Horizons*, vol. 63, no. 2, pp. 135–146, 2020, doi: https://doi.org/10.1016/j.bushor.2019.11.006.

[3] A. Dash, J. Ye, and G. Wang, "A Review of Generative Adversarial Networks (GANs) and Its Applications in a Wide Variety of Disciplines: From Medical to Remote Sensing," *IEEE Access*, vol. 12, no. October 2023, pp. 18330–18357, 2024, doi: 10.1109/ACCESS.2023.3346273.

[4] Y. Kang, W. Kim, S. Lim, H. Kim, and H. Seo, "DeepDetection: Privacy-Enhanced Deep Voice Detection and User Authentication for Preventing Voice Phishing," *Applied Sciences (Switzerland)*, vol. 12, no. 21, 2022, doi: 10.3390/app122111109.

[5] Q. Le Roux, E. Bourbao, Y. Teglia, and K. Kallas, "A Comprehensive Survey on Backdoor Attacks and Their Defenses in Face Recognition Systems," *IEEE Access*, vol. 12, no. April, pp. 47433–47468, 2024, doi: 10.1109/ACCESS.2024.3382584.

[6] B. Yan, J. Lan, and Z. Yan, "Backdoor Attacks against Voice Recognition Systems: A Survey," *ACM Computing Surveys*, vol. 57, no. 3, 2024, doi: 10.1145/3701985.

[7] S. Ali, S. U. Rehman, A. Imran, G. Adeem, Z. Iqbal, and K. Il Kim, "Comparative Evaluation of AI-Based Techniques for Zero-Day Attacks Detection," *Electronics (Switzerland)*, vol. 11, no. 23, pp. 1–25, 2022, doi: 10.3390/electronics11233934.

[8] T. Sowmya and E. A. Mary Anita, "A comprehensive review of AI based intrusion detection system," *Measurement: Sensors*, vol. 28, no. May, p. 100827, 2023, doi: 10.1016/j.measen.2023.100827.

[9] S. Slamet, "Pertahanan Pencegahan Serangan Social Engineering Menggunakan Two Factor Authentication (2Fa) Berbasis Sms (Short Message System)," *Spirit*, vol. 14, no. 2, pp. 23–29, 2023, doi: 10.53567/spirit.v14i2.260.

[10] S. Slamet, "Desain Arsitektur Aplikasi Qr Code Sebagai Anti Phishing Serangan Qr Code," *Spirit*, vol. 15, no. 1, pp. 42–48, 2023, doi: 10.53567/spirit.v15i1.280.

[11] E. Marasco, M. Albanese, V. V. R. Patibandla, A. Vurity, and S. S. Sriram, " Biometric multi-factor authentication: On the usability of the FingerPIN scheme ," *Security and Privacy*, vol. 6, no. 1, pp. 1–14, 2023, doi: 10.1002/spy2.261.

[12] M. M. Taye, "Theoretical Understanding of Convolutional Neural Network: Concepts, Architectures, Applications, Future Directions," *Computation*, vol. 11, no. 3, 2023, doi: 10.3390/computation11030052.

[13] S. Siami-Namini, N. Tavakoli, and A. S. Namin, "The Performance of LSTM and BiLSTM in Forecasting Time Series," *Proceedings - 2019 IEEE International Conference on Big Data, Big Data 2019*, pp. 3285–3292, 2019, doi: 10.1109/BigData47090.2019.9005997.

[14] D. E. Kurniawan, M. Iqbal, J. Friadi, F. Hidayat, and R. D. Permatasari, "Login Security Using One Time Password (OTP) Application with Encryption Algorithm Performance," *Journal of Physics: Conference Series*, vol. 1783, no. 1, 2021, doi: 10.1088/1742-6596/1783/1/012041.

[15] R. Venkatesan, S. Shirly, M. Selvarathi, and T. J. Jebaseeli, "Human Emotion Detection Using

DeepFace and Artificial Intelligence †," *Engineering Proceedings*, vol. 59, no. 1, 2023, doi: 10.3390/engproc2023059037.

[16] C. Korgialas, C. Kotropoulos, and K. N. Plataniotis, "Leveraging Electric Network Frequency Estimation for Audio Authentication," *IEEE Access*, vol. 12, no. December 2023, pp. 9308–9320, 2024, doi: 10.1109/ACCESS.2024.3354053.

[17] J. Chen, L. Cai, Y. Tu, R. Dong, D. An, and B. Zhang, "An Identity Authentication Method Based on Multi-modal Feature Fusion," *Journal of Physics: Conference Series*, vol. 1883, no. 1, 2021, doi: 10.1088/1742-6596/1883/1/012060.

[18] M. I. Ardiawan and G. P. K. Negarara, "A Comparative Analysis of FaceNet, VGGFace, and GhostFaceNets Face Recognition Algorithms For Potential Criminal Suspect Identification," *Journal of Applied Artificial Intelligence*, vol. 5, no. 2, pp. 34–49, 2024.

[19] Z. K. Abdul and A. K. Al-Talabani, "Mel Frequency Cepstral Coefficient and its Applications: A Review," *IEEE Access*, vol. 10, no. November, pp. 122136–122158, 2022, doi: 10.1109/ACCESS.2022.3223444.

[20] S. Serrano, L. Patanè, O. Serghini, and M. Scarpa, "Detection and Classification of Obstructive Sleep Apnea Using Audio Spectrogram Analysis," *Electronics (Switzerland)*, vol. 13, no. 13, pp. 1–27, 2024, doi: 10.3390/electronics13132567.

[21] A. Malik, M. Kuribayashi, S. M. Abdullahi, and A. N. Khan, "DeepFake Detection for Human Face Images and Videos: A Survey," *IEEE Access*, vol. 10, pp. 18757–18775, 2022, doi: 10.1109/ACCESS.2022.3151186.

[22] B. Yasser *et al.*, "Deepfake Detection Using EfficientNet and XceptionNet," *Proceedings - 11th IEEE International Conference on Intelligent Computing and Information Systems, ICICIS 2023*, no. June, pp. 598–603, 2023, doi: 10.1109/ICICIS58388.2023.10391114.

[23] A. Onan, "Bidirectional convolutional recurrent neural network architecture with group-wise enhancement mechanism for text sentiment classification," *Journal of King Saud University - Computer and Information Sciences*, vol. 34, no. 5, pp. 2098–2117, 2022, doi: 10.1016/j.jksuci.2022.02.025.

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

[25] X. Liu *et al.*, "ASVspoof 2021: Towards Spoofed and Deepfake Speech Detection in the Wild," *IEEE/ACM Transactions on Audio Speech and Language Processing*, vol. 31, pp. 2507–2522, 2023, doi: 10.1109/TASLP.2023.3285283.

[26] H. Sun *et al.*, "An Improved Medical Image Classification Algorithm Based on Adam Optimizer," *Mathematics*, vol. 12, no. 16, pp. 1–14, 2024, doi: 10.3390/math12162509.