Journal of Big Data Analytic and Artificial Intelligence
Volume 8(2), Desember 2025, pp. 54-61

Analisis Konseptual Fusi Multimodal Wajah dan Visual Speech untuk Autentikasi Biometrik Non-Vokal terhadap Deepfake

Ahmad Roihan[1], Rosmawati Dwi[2], Erna Astriyani[3]

[1*]Sistem Komputer, Universitas Raharja, Kota Tangerang, Banten
[2,3] Sistem Informasi, Universitas Raharja, Kota Tangerang, Banten
Email: [1*]ahmad.roihan@raharja.info, [2]rosmawati.dwi@raharja.info, [3]erna.astriyani@raharja.info

Abstrak

Sistem autentikasi biometrik masih menghadapi keterbatasan mendasar, khususnya pada pendekatan unimodal yang rentan terhadap variasi lingkungan dan serangan pemalsuan visual. Untuk mengatasi permasalahan tersebut, biometrik multimodal yang mengintegrasikan ciri fisiologis dan perilaku menjadi pendekatan yang semakin relevan. Penelitian ini menyajikan kajian analitis terhadap riset-riset terkini di bidang biometrik multimodal visual dengan fokus pada strategi fusi level skor serta integrasi pengenalan wajah statis dan gerakan bibir dinamis sebagai mekanisme autentikasi non-vokal. Hasil sintesis literatur menunjukkan bahwa fusi level skor merupakan pendekatan yang paling fleksibel dan stabil dalam menggabungkan modalitas biometrik yang heterogen, terutama pada kombinasi ciri spasial statis dan pola temporal dinamis. Selain itu, arsitektur *deep learning* berbasis Transformer diidentifikasi memiliki potensi signifikan dalam memodelkan dependensi temporal pada gerakan bibir. Kajian ini juga menyoroti tantangan keamanan utama, khususnya serangan presentasi dan *visual-only deepfake*, serta menekankan pentingnya deteksi keaslian berbasis dinamika visual sebagai bagian integral dari sistem autentikasi. Berdasarkan analisis tersebut, penelitian ini merumuskan kerangka konseptual autentikasi biometrik visual multimodal yang mengintegrasikan verifikasi identitas dan *liveness detection* dalam satu alur proses, sekaligus mengidentifikasi peluang riset lanjutan meliputi pemanfaatan *self-supervised learning*, optimasi model untuk perangkat dengan sumber daya terbatas, dan perancangan frasa sandi visual yang lebih diskriminatif.

Kata Kunci: biometrik multimodal, pengenalan wajah, fusi level skor, *deepfake*, *liveness detection*.

A Conceptual Analysis of Face and Visual Speech Multimodal Fusion for Non-Vocal Biometric Authentication

Abstract

Biometric authentication systems still face fundamental limitations, particularly in unimodal approaches that are vulnerable to environmental variations and visual spoofing attacks. To address these challenges, multimodal biometrics integrating physiological and behavioral traits have become an increasingly relevant approach. This study presents an analytical review of recent research in visual multimodal biometrics, with a focus on score-level fusion strategies and the integration of static face recognition and dynamic lip movement analysis as a non-vocal authentication mechanism. The literature synthesis indicates that score-level fusion is the most flexible and stable approach for combining heterogeneous biometric modalities, especially when integrating static spatial features and dynamic temporal patterns. Furthermore, Transformer-based deep learning architectures are identified as having significant potential for modeling the temporal dependencies of lip movements. This study also highlights key security challenges, particularly presentation attacks and visual-only deepfakes, and emphasizes the importance of visual dynamics–based liveness detection as an integral component of biometric authentication systems. Based on these findings, the study formulates a conceptual framework for visual multimodal biometric authentication that integrates identity verification and liveness detection within a unified process, while also identifying future research opportunities, including self-supervised learning, model optimization for resource-constrained devices, and the design of more discriminative visual passphrases.

Keywords: *multimodal biometrics, face recognition, score-level fusion, deepfake, liveness detection.*

I.PENDAHULUAN

Perkembangan sistem autentikasi biometrik modern menunjukkan pergeseran signifikan dari pendekatan unimodal menuju biometrik multimodal sebagai upaya meningkatkan keandalan dan keamanan sistem identitas digital seperti autentikasi, sistem pintu otomatis [1] dan sistem keamanan digital lainnya. Biometrik multimodal didefinisikan sebagai metode autentikasi yang mengombinasikan dua atau lebih ciri biometrik, baik bersifat fisiologis maupun perilaku untuk melakukan proses verifikasi atau identifikasi individu secara lebih akurat dan *robust* [2]. Pendekatan ini muncul sebagai respon terhadap berbagai kelemahan mendasar yang melekat pada sistem biometrik unimodal.

Sistem biometrik unimodal, yang hanya mengandalkan satu modalitas seperti wajah atau sidik jari, memiliki kerentanan tinggi terhadap kondisi operasional yang tidak ideal. Beberapa permasalahan utama meliputi *noise* pada data sensor, variasi intra-kelas akibat perubahan usia, ekspresi, atau kondisi lingkungan, kurangnya universalitas, serta tingginya risiko presentation *attack* seperti penggunaan foto, video, atau artefak sintetis [3], [4]. Ketika satu-satunya modalitas gagal memberikan hasil yang valid, misalnya akibat pencahayaan yang buruk pada pengenalan wajah, maka sistem tidak memiliki mekanisme cadangan untuk memastikan keberlangsungan proses autentikasi.

Penelitian terdahulu menunjukkan bahwa pola gerakan bibir bersifat unik antarindividu, bahkan ketika frasa yang diucapkan sama, sehingga dapat digunakan sebagai ciri biometrik perilaku yang diskriminatif [5]. Dibandingkan sistem autentikasi berbasis suara, pendekatan visual non-vokal ini memiliki keunggulan signifikan karena tidak terpengaruh oleh kebisingan lingkungan, gema, maupun pembatasan privasi di ruang publik.

Penelitian ini melakukan eksperimen awal untuk mengevaluasi kerangka konseptual autentikasi biometrik visual multimodal yang mengintegrasikan pengenalan wajah sebagai biometrik fisiologis dan analisis gerakan bibir sebagai biometrik perilaku dinamis. Eksperimen dilakukan menggunakan dataset visual terkontrol berupa citra wajah dan sekuens video pendek gerakan bibir, yang diproses melalui tahapan ekstraksi fitur, normalisasi, dan verifikasi identitas. Metodologi yang diterapkan menggunakan fusi level skor, di mana skor kecocokan dari masing-masing modalitas digabungkan untuk menghasilkan keputusan autentikasi akhir. Hasil eksperimen awal menunjukkan bahwa integrasi gerakan bibir berpotensi meningkatkan ketahanan sistem terhadap variasi lingkungan dan serangan presentasi berbasis visual dibandingkan autentikasi berbasis wajah tunggal. Selain itu, penelitian ini mengidentifikasi tantangan utama yang masih terbuka, meliputi ancaman *deepfake visual*, keterbatasan data berlabel berskala besar, serta kebutuhan implementasi *real-time* pada perangkat dengan sumber daya terbatas, sekaligus menyoroti peluang riset lanjutan seperti penerapan *self-supervised learning*, optimasi arsitektur deep learning untuk komputasi tepi, dan perancangan frasa sandi visual yang lebih diskriminatif.

II. METODOLOGI PENELITIAN

Penelitian ini menggunakan pendekatan *Systematic Analytical Review* dan Eksperimental untuk menganalisis secara kritis riset-riset terkini di bidang biometrik multimodal, dengan fokus utama pada integrasi pengenalan wajah sebagai biometrik fisiologis statis dan *visual speech* atau gerakan bibir dinamis sebagai biometrik perilaku. Selain meninjau aspek akurasi dan strategi integrasi multimodal, kajian ini secara khusus menempatkan ketahanan terhadap *presentation attack*, termasuk serangan *visual-only deepfake*, sebagai dimensi evaluasi utama. Metode ini dirancang untuk mengidentifikasi pola pendekatan yang dominan, keterbatasan teknis yang masih bertahan, serta peluang riset yang belum dieksplorasi secara optimal dalam literatur biometrik visual.

Desain penelitian yang digunakan adalah studi literatur terstruktur dengan analisis komparatif lintas penelitian, dengan melibatkan eksperimen pada dataset yang ada. Tujuan utama dari pendekatan ini adalah untuk (1) mengkaji berbagai pendekatan autentikasi biometrik multimodal yang telah diusulkan, (2) membandingkan metode ekstraksi ciri, strategi fusi multimodal, serta mekanisme keamanan yang digunakan, dan (3) mengidentifikasi gap penelitian dan peluang inovasi algoritmik, khususnya untuk sistem autentikasi non-vokal berbasis visual. Pendekatan ini sejalan dengan karakteristik review paper analitis, di mana kontribusi ilmiah difokuskan pada sintesis konseptual dan pembentukan kerangka evaluasi, bukan pada pengujian performa numerik baru.

Penelusuran literatur dilakukan secara sistematis melalui basis data ilmiah bereputasi, yaitu IEEE Xplore, ScienceDirect (Elsevier), SpringerLink, dan ACM Digital Library. Proses pencarian menggunakan kombinasi kata kunci yang relevan dengan topik penelitian, meliputi *multimodal biometrics*, *face recognition*, *visual speech recognition*, *lip movement biometrics* [6], *score-level fusion*, *presentation attack detection*, dan *visual deepfake*. Artikel yang dipertimbangkan harus memenuhi kriteria inklusi, yaitu merupakan publikasi jurnal atau prosiding internasional bereputasi, diterbitkan dalam sepuluh tahun terakhir, memiliki DOI dan dapat diakses secara publik, serta membahas setidaknya satu aspek utama dari ekstraksi ciri wajah, ekstraksi ciri gerakan bibir, fusi multimodal, atau *presentation attack detection*. Studi *non-peer reviewed*, penelitian unimodal tanpa pembahasan keamanan, serta makalah yang berfokus murni pada aspek linguistik tanpa konteks biometrik dikeluarkan dari analisis.

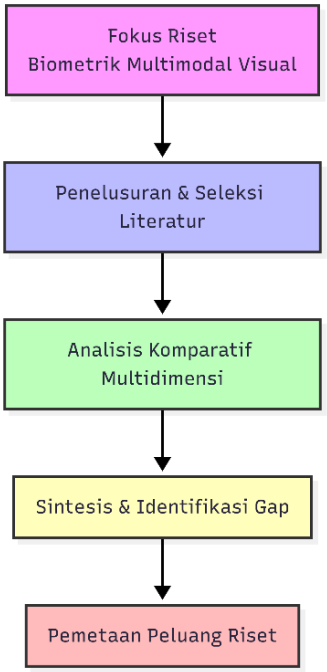
Untuk menjamin konsistensi dan objektivitas, penelitian ini menyusun kerangka analisis multidimensi yang diterapkan secara seragam pada setiap studi terpilih. Dimensi analisis meliputi jenis modalitas biometrik (unimodal atau multimodal, fisiologis statis atau perilaku dinamis), metode ekstraksi ciri yang digunakan (CNN konvensional, kombinasi CNN–RNN/LSTM, atau arsitektur Transformer berbasis self-attention), strategi fusi multimodal (level ciri, level skor, atau hibrida), serta mekanisme keamanan dan *liveness detection* yang diterapkan. Selain itu, metrik evaluasi yang digunakan dalam setiap penelitian, seperti *Equal Error Rate* (EER), *False Acceptance Rate* (FAR), *True Accept Rate* (TAR), dan

Journal of Big Data Analytic and Artificial Intelligence
Volume 8(2), Desember 2025, pp. 54-61

akurasi sekuens visual, serta kesiapan implementasi dari sisi kompleksitas komputasi dan potensi penerapan pada perangkat tepi turut dianalisis untuk memberikan perspektif praktis.

Sintesis hasil dilakukan menggunakan pendekatan *comparative–interpretative analysis*, di mana temuan dari berbagai penelitian tidak hanya dirangkum, tetapi juga dipetakan terhadap keterbatasan teknis, relevansi terhadap ancaman keamanan aktual, dan kesesuaiannya untuk skenario autentikasi non-vokal di lingkungan nyata. Gap penelitian diidentifikasi dengan mengamati area yang menunjukkan performa teoritis tinggi tetapi sulit diimplementasikan, aspek keamanan, khususnya terhadap *visual-only deepfake*, yang belum ditangani secara memadai, serta ketimpangan antara kemajuan algoritma dan kesiapan penerapan dunia nyata. Hasil sintesis ini kemudian digunakan untuk merumuskan peluang riset kombinasi pengenalan wajah dan analisis pergerakan bibir secara bersamaan, terutama pada integrasi biometrik perilaku dinamis dan deteksi keaslian visual dalam satu proses autentikasi terpadu.

Sebagai keluaran utama sebagaimana dituangkan dalam Gambar 1, penelitian ini memetakan peluang riset ke dalam domain strategis, meliputi optimalisasi fusi level skor untuk modalitas heterogen, pemanfaatan arsitektur Transformer dalam pemodelan gerakan bibir dinamis, integrasi verifikasi identitas dan *liveness detection visual*, penerapan *self-supervised learning* untuk efisiensi data biometrik, serta optimasi model untuk *edge computing* guna mendukung arsitektur keamanan terdesentralisasi. Dengan demikian, kontribusi penelitian ini tidak terletak pada klaim peningkatan kinerja numerik, melainkan pada penyediaan kerangka analitis dan peta peluang riset yang dapat menjadi landasan pengembangan sistem autentikasi biometrik visual generasi berikutnya.



Gambar 1. Metode Penelitian

III. HASIL DAN PEMBAHASAN

A. Fokus Riset Biometrik Multimodal Visual

Berbagai literatur terkini menegaskan bahwa integrasi lebih dari satu modalitas biometrik merupakan strategi efektif untuk mengatasi keterbatasan sistem unimodal, terutama ketika menghadapi variasi kondisi lingkungan dan potensi ancaman keamanan. Sistem unimodal, seperti pengenalan wajah saja, rentan terhadap kegagalan jika kualitas citra buruk atau pengguna mengalami perubahan penampilan. Dalam konteks ini, kombinasi pengenalan wajah statis dan gerakan bibir dinamis dipandang sebagai kerangka konseptual yang menjanjikan karena mampu memanfaatkan keunggulan masing-masing modalitas. Modalitas wajah menyediakan informasi fisiologis yang stabil, sementara gerakan bibir menawarkan pola perilaku dinamis yang sulit dipalsukan. Integrasi kedua modalitas ini memungkinkan verifikasi identitas sekaligus *liveness detection* berbasis visual, tanpa menambah kompleksitas sistem secara signifikan. Pendekatan ini menempatkan sistem pada posisi yang lebih *robust* terhadap serangan *visual-only*, seperti *deepfake*, sekaligus menjaga kenyamanan pengguna karena proses autentikasi tetap alami. Dalam konteks integrasi biometrik wajah (fisiologis statis) dan gerakan bibir (perilaku dinamis), studi-studi yang ditinjau mengindikasikan bahwa pemilihan tingkat fusi memiliki implikasi langsung terhadap kompleksitas sistem dan potensi peningkatan kinerja.

Fusi pada level ciri (*feature-level fusion*) secara teoritis menyediakan representasi informasi paling kaya karena menggabungkan seluruh fitur sebelum tahap pencocokan. Namun, sebagian besar penelitian melaporkan bahwa pendekatan ini sulit diterapkan secara praktis ketika modalitas yang digabungkan bersifat heterogen, seperti ciri wajah statis dan urutan gerakan bibir spatiotemporal. Tantangan utama meliputi penyelarasan dimensi fitur, perbedaan skala temporal, serta peningkatan beban komputasi [7].

Sebaliknya, fusi level skor (*score-level fusion*) dilaporkan sebagai strategi yang paling umum dan stabil dalam sistem biometrik multimodal heterogen. Pendekatan ini memungkinkan setiap modalitas diproses secara independen hingga menghasilkan skor kecocokan, yang kemudian digabungkan menggunakan metode penjumlahan tertimbang atau klasifikasi pasca-fusi. Sintesis literatur menunjukkan bahwa fusi level skor menawarkan keseimbangan yang lebih baik antara kompleksitas implementasi dan peningkatan akurasi, sehingga sering direkomendasikan dalam sistem autentikasi multimodal dunia nyata.

B. Penelusuran dan Seleksi Literatur

Penelusuran literatur dilakukan secara sistematis melalui beberapa basis data ilmiah bereputasi, yaitu IEEE Xplore, ScienceDirect, SpringerLink, dan ACM Digital Library, dengan fokus pada publikasi internasional terkini. Kata kunci yang digunakan mencakup multimodal biometrics, *face recognition*, *visual speech recognition*, *lip movement biometrics*, *score-level fusion*, *presentation attack detection*, dan *visual deepfake*, sehingga mencakup topik integrasi modalitas, evaluasi keamanan, dan teknik fusi. Kriteria inklusi

meliputi jurnal atau prosiding internasional yang diterbitkan dalam sepuluh tahun terakhir, memiliki DOI, dapat diakses secara publik, serta membahas setidaknya satu aspek dari ekstraksi ciri wajah, ekstraksi ciri gerakan bibir, fusi multimodal, atau mekanisme anti-spoofing. Sebaliknya, studi *non-peer reviewed*, penelitian unimodal, dan makalah murni linguistik dikeluarkan dari analisis untuk menjaga konsistensi dan kualitas tinjauan. Pendekatan ini memastikan bahwa literatur yang dianalisis relevan dengan konteks biometrik multimodal visual dan memberikan landasan konseptual yang valid untuk eksperimen *proof-of-concept*.

C. Analisis Komparatif Multidimensi

Hasil analisis komparatif menunjukkan bahwa strategi fusi modalitas memiliki implikasi langsung terhadap kompleksitas sistem dan potensi peningkatan akurasi. *Feature-level fusion*, yang menggabungkan seluruh fitur sebelum tahap pencocokan, memberikan representasi informasi paling kaya, namun sulit diterapkan pada modalitas heterogen karena perbedaan dimensi fitur, skala temporal, dan beban komputasi yang tinggi. Sebaliknya, *score-level fusion* lebih stabil dan umum digunakan dalam sistem multimodal heterogen. Pendekatan ini memungkinkan setiap modalitas diproses secara independen hingga menghasilkan skor yang kemudian digabungkan menggunakan metode berbobot atau klasifikasi pasca-fusi. Berdasarkan perbandingan lintas studi, fusi level skor paling sesuai untuk sistem autentikasi yang menggabungkan ciri wajah statis dan gerakan bibir dinamis. Karakteristik data kombinasi pengenalan wajah dan analisis pergerakan bibir secara bersamaan, mengintegrasikan informasi spasial statis dan urutan temporal panjang, menjadikan fusi skor sebagai pendekatan yang mampu menghindari permasalahan penyelarasan fitur mentah.

Studi-studi multimodal serupa, khususnya pada kombinasi wajah dan suara, melaporkan penurunan *Equal Error Rate* (EER) yang signifikan dibandingkan sistem unimodal. Temuan ini menunjukkan bahwa integrasi dua sumber biometrik yang saling melengkapi dapat meningkatkan ketahanan sistem terhadap kegagalan satu modalitas. Dalam konteks kombinasi pengenalan wajah dan analisis pergerakan bibir secara bersamaan, hasil-hasil tersebut tidak ditafsirkan sebagai capaian langsung, melainkan sebagai *benchmark* konseptual yang menunjukkan potensi peningkatan kinerja melalui fusi multimodal berbasis visual.

Dengan demikian, kombinasi pengenalan wajah dan analisis pergerakan bibir secara bersamaan diposisikan sebagai kerangka konseptual yang memanfaatkan keunggulan fusi level skor untuk menggabungkan verifikasi identitas fisiologis dan perilaku dinamis, tanpa meningkatkan kompleksitas sistem secara berlebihan.

Literatur menegaskan bahwa evaluasi sistem biometrik tidak dapat bergantung pada satu metrik tunggal. *Equal Error Rate* (EER) tetap menjadi indikator utama keseimbangan antara keamanan dan kenyamanan pengguna, sementara *False Acceptance Rate* (FAR) mencerminkan risiko keamanan yang paling kritis, terutama dalam konteks serangan presentasi. *True Accept Rate* (TAR) digunakan untuk mengukur tingkat penerimaan pengguna sah.

Namun, karena kombinasi pengenalan wajah dan analisis pergerakan bibir secara bersamaan memanfaatkan gerakan bibir sebagai biometrik perilaku dinamis, evaluasi kinerjanya juga berkaitan erat dengan domain *Visual Speech Recognition* (VSR). Oleh karena itu, beberapa penelitian merekomendasikan penggunaan metrik tambahan seperti akurasi sekuens atau *Word Error Rate* (WER) untuk menilai konsistensi pola artikulasi visual. Integrasi metrik biometrik tradisional dan metrik VSR memberikan gambaran yang lebih komprehensif tentang kemampuan sistem dalam membedakan identitas berbasis pola dinamis, bukan hanya kesamaan visual statis.

Salah satu temuan utama dari analisis literatur adalah bahwa *presentation attack*, khususnya *visual-only deepfake*, merupakan ancaman paling signifikan bagi sistem autentikasi biometrik berbasis visual. Tidak digunakannya komponen audio dalam kombinasi pengenalan wajah dan analisis pergerakan bibir secara bersamaan meningkatkan privasi dan ketahanan terhadap kebisingan, tetapi sekaligus menempatkan seluruh beban keamanan pada mekanisme analisis visual. Hal ini dapat dilihatdalam Tabel 1.

Tabel 1. Analisis Riset Biometrik Multimodal & *Score-Level Fusion*

Tema Penelitian	Modalitas / Fusion Level	Temuan Kinerja / Fokus
Score level fusion menggunakan triangular *norms* [8]	*Multimodal biometrics, score level*	Menunjukkan *score level fusion* efektif dalam menggabungkan modalitas dengan norma-norma t-norm untuk menangani ketidakpastian
Weighted score level fusion (iris + wajah) [9], [10]	Multimodal (iris, wajah), *weighted score fusion*	Integrasi skor dengan bobot untuk meningkatkan akurasi dan meminimalkan EER
Evaluasi fusion skor pada multimodal [11]	Multimodal 3 karakteristik, score fusion	Evaluasi skor sum rule & SVM fusion; multimodal outperform unimodal
Optimal score level fusion untuk tiga modalitas [12]	Multimodal (iris, finger vein, fingerprint), optimal score fusion	Model optimal meningkatkan ketahanan sistem dengan resolusi konflik skor
Fusion skor untuk 3D *face* & 3D *ear recognition* [13]	Multimodal (3D *face*, 3D	*Score-level fusion* menggabungkan skor dua modalitas

Journal of Big Data Analytic and Artificial Intelligence
Volume 8(2), Desember 2025, pp. 54-61

Tema Penelitian	Modalitas / Fusion Level	Temuan Kinerja / Fokus
	ear), *score level*	3D dengan akurasi pengenalan tinggi
Survei fusion level dalam biometrik [14]	Multimodal, berbagai level fusion	Survey teknik fusion termasuk score-level untuk memahami *trade-off* dan performa
Multimodal score level fusion (wajah + iris SVM) [15]	Multimodal, score fusion (SVM)	Menggabungkan skor wajah & iris dengan SVM untuk autentikasi lebih baik dari unimodal

Literatur terkini menunjukkan bahwa pendekatan *liveness detection* berbasis dinamika visual merupakan strategi yang paling menjanjikan untuk menghadapi ancaman ini. Analisis kedalaman 3D, tekstur kulit, refleksi cahaya, serta *micro-expressions* dilaporkan sebagai indikator keaslian yang sulit direplikasi secara konsisten oleh media palsu atau video sintetis. Dalam konteks kombinasi pengenalan wajah dan analisis pergerakan bibir secara bersamaan, keharusan pengguna untuk mengartikulasikan kata sandi secara diam menciptakan stimulus dinamis alami yang dapat dimanfaatkan secara simultan untuk verifikasi identitas dan deteksi keaslian.

D. Sintesis dan Identifikasi Gap

Salah satu temuan utama dari kajian literatur adalah bahwa *presentation attack*, khususnya berbasis *visual-only deepfake*, merupakan ancaman paling signifikan bagi sistem autentikasi visual. Karena pendekatan ini tidak menggunakan komponen audio, seluruh beban keamanan bertumpu pada analisis visual. Strategi *liveness detection* berbasis dinamika visual dianggap paling menjanjikan, dengan indikator keaslian seperti kedalaman 3D, tekstur kulit, refleksi cahaya, dan *micro-expressions* yang sulit direplikasi secara konsisten. Berdasarkan sintesis, beberapa peluang riset dapat diidentifikasi, antara lain: optimalisasi fusi level skor untuk modalitas heterogen, pemanfaatan arsitektur Transformer untuk memodelkan dependensi spatiotemporal gerakan bibir, penggunaan *self-supervised learning* untuk efisiensi data biometrik, optimasi model untuk perangkat *edge computing*, serta perancangan sandi visual yang lebih diskriminatif agar pola dinamis eksekusi autentikasi lebih unik dan sulit ditiru. Semua temuan ini menegaskan kebutuhan pengembangan sistem autentikasi biometrik visual yang lebih adaptif dan aman, sekaligus membuka jalur riset lanjutan dalam integrasi modalitas perilaku dan fisiologis. Pendekatan ini dipandang sebagai keunggulan konseptual kombinasi pengenalan wajah dan analisis pergerakan bibir secara bersamaan, karena mengintegrasikan autentikasi dan *liveness detection* dalam satu alur proses tanpa menambah beban interaksi pengguna.

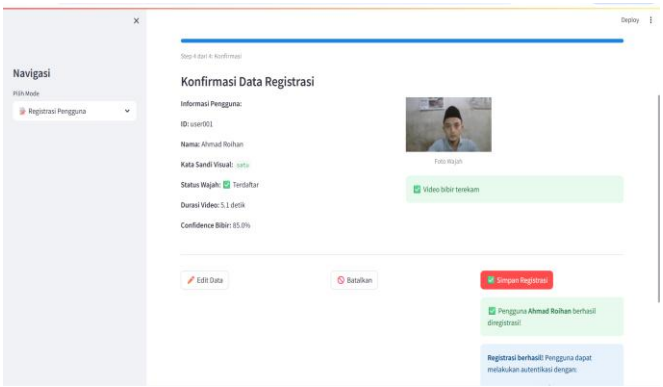
Analisis komparatif menunjukkan bahwa peluang riset kombinasi pengenalan wajah dan analisis pergerakan bibir secara bersamaan terletak pada irisan antara biometrik multimodal, *visual speech*, dan keamanan *anti-spoofing* [16]. Beberapa arah riset yang secara konsisten muncul dalam literatur meliputi pemanfaatan arsitektur Transformer untuk pemodelan dependensi *spatiotemporal* jangka panjang, penggunaan *self-supervised learning* untuk mengurangi ketergantungan pada data berlabel sensitif, serta optimasi model untuk implementasi pada perangkat tepi.

Selain itu, tantangan non-universalitas biometrik perilaku membuka peluang penelitian pada perancangan frasa sandi visual yang lebih diskriminatif, di mana fokus autentikasi bergeser dari sekadar konten artikulasi menuju pola dinamis eksekusinya.

E. Pemetaan Peluang Riset

Sebagai implementasi awal, penelitian ini mengembangkan sistem *proof-of-concept* yang mensimulasikan integrasi pengenalan wajah dan gerakan bibir. Selain untuk pemetaan peluang riset, eksperimen ini juga dirancang untuk mengatasi keterbatasan sistem autentikasi unimodal dengan memanfaatkan pola unik gerakan bibir ketika pengguna mengucapkan kata sandi visual tanpa suara. Integrasi dua modalitas tersebut berfungsi sebagai mekanisme *passive liveness detection*, di mana gerakan bibir alami menjadi bukti keaslian pengguna yang sulit untuk dipalsukan menggunakan media statis maupun rekaman video. *Adapun library* yang digunakan dalam sistem eksperimen ini antara lain:

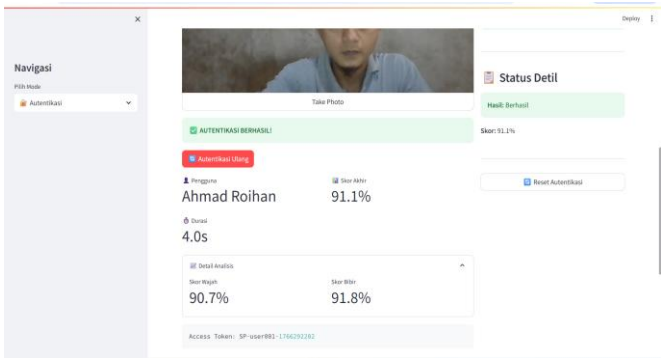
1. streamlit==1.28.1
2. numpy==1.24.3
3. opencv-python==4.8.1.78
4. torch==2.1.0
5. torchvision==0.16.0
6. facenet-pytorch==2.5.3
7. mediapipe==0.10.8
8. scikit-learn==1.3.0
9. matplotlib==3.7.2
10. pandas==2.0.3
11. seaborn==0.12.2
12. dlib==19.24.1
13. Pillow==10.0.1
14. gradio==3.50.2
15. python-multipart==0.0.6
16. gunicorn==21.2.0



Gambar 2. Registrasi Pengguna

Hasil eksperimen divisualisasikan melalui antarmuka web, registrasi pengguna (Gambar 2). Penting dicatat bahwa sistem ini bersifat simulasi terbatas, bukan evaluasi performa eksperimental komprehensif. Implementasi sistem menggunakan arsitektur *neural network* bercabang dua, yang memproses fitur wajah dan fitur gerakan bibir secara terpisah sebelum digabungkan. Cabang wajah memanfaatkan *embedding* berdimensi 512, sedangkan cabang gerakan bibir menggunakan vektor fitur berdimensi 600. Masing-masing cabang diproses melalui transformasi linear dengan fungsi aktivasi ReLU, kemudian hasilnya digabungkan (*concatenation*) dan dilewatkan ke *fusion layer* untuk menghasilkan representasi multimodal yang lebih ringkas dan informatif.

Proses verifikasi identitas dilakukan dengan menghitung tingkat kemiripan antara fitur autentikasi dan template yang tersimpan menggunakan *cosine similarity*. Skor autentikasi akhir diperoleh dari kombinasi berbobot antara skor kecocokan wajah dan skor konsistensi gerakan bibir, dengan bobot yang lebih besar diberikan pada modalitas wajah. Strategi ini memastikan bahwa sistem tetap mengandalkan identitas visual utama, namun tetap mempertimbangkan dinamika gerakan bibir sebagai faktor pendukung keamanan. Visualisasi hasil deteksi dan skor autentikasi ditampilkan pada Gambar 3.



Gambar 3. Skor Autentikasi Pengguna

Eksperimen dilakukan menggunakan *framework* PyTorch dengan arsitektur yang relatif sederhana namun efektif, terdiri dari dua *fully connected layer* pada setiap cabang, fusion layer, serta *dropout regularization* untuk mengurangi risiko *overfitting*. Meskipun masih bersifat dasar, hasil penelitian menunjukkan potensi pengembangan lebih lanjut, seperti penambahan *convolutional layers* untuk ekstraksi fitur visual yang lebih kuat dan penerapan pemodelan temporal guna meningkatkan akurasi analisis urutan gerakan bibir.

Perlu ditegaskan bahwa penelitian ini bersifat kajian awal (*preliminary analysis*) yang dipadukan dengan eksperimen *proof-of-concept* simulasi. Sistem yang dikembangkan masih dalam skala terbatas dan belum diuji secara eksperimental komprehensif di lingkungan nyata. Oleh karena itu, temuan yang disajikan lebih bersifat landasan konseptual dan peta peluang riset, bukan sebagai klaim peningkatan performa numerik final. Penelitian ini berfokus pada validasi konsep dasar autentikasi multimodal berbasis wajah dan gerakan bibir, dengan implementasi yang disengaja disederhanakan untuk mengisolasi dan menganalisis prinsip-prinsip fundamental sebelum pengembangan lebih kompleks.

Temuan eksperimen ini membuka beberapa arah pengembangan riset masa depan yang potensial untuk dikembangkan lebih lanjut:

1. Pertama, implementasi sistem nyata dengan pengujian eksperimental berskala besar menjadi langkah kritis berikutnya.

Penelitian lanjutan perlu mengembangkan dataset yang lebih representatif dengan variasi demografis, kondisi pencahayaan, dan lingkungan pengambilan data yang lebih beragam. Pengujian terhadap ribuan sampel dengan variasi etnis, usia, dan kondisi fisiologis yang berbeda akan memberikan validasi eksternal yang lebih kuat. Selain itu, pengujian di lingkungan *real-world* dengan kondisi dinamis akan mengungkap tantangan praktis yang tidak terlihat dalam setting laboratorium terkontrol.

2. Eksplorasi arsitektur Transformer dan *attention mechanisms*

Hal ini untuk pemodelan gerakan bibir dan integrasi fusi multimodal menjanjikan peningkatan signifikan dalam akurasi sistem. Transformer, dengan kemampuannya menangkap dependensi jangka panjang dan konteks temporal, sangat sesuai untuk memodelkan urutan gerakan bibir yang bersifat dinamis. Mekanisme attention dapat digunakan untuk mengidentifikasi frame-frame kritis dalam urutan gerakan bibir yang paling informatif untuk autentikasi, serta untuk mengoptimalkan fusi multimodal dengan memberikan bobot adaptif berdasarkan kualitas dan relevansi masing-masing modalitas.

3. Pemanfaatan *self-supervised learning* (SSL)

Hal ini berguna untuk efisiensi penggunaan data biometrik menjadi peluang riset yang strategis mengingat keterbatasan dataset biometrik berlabel besar. Teknik SSL seperti *contrastive learning* dan *masked autoencoding* dapat memanfaatkan data biometrik tak berlabel yang lebih melimpah untuk mempelajari representasi yang robust sebelum fine-tuning dengan data berlabel terbatas. Pendekatan ini sangat relevan dalam konteks Indonesia yang memerlukan dataset biometrik spesifik populasi lokal.

4. Optimasi model pada perangkat *edge computing*

Optimasi model berguna untuk mendukung implementasi terdesentralisasi dan *privacy-preserving authentication*. Penelitian dapat mengarah pada pengembangan model yang lebih ringan melalui teknik seperti *quantization*, *pruning*, *knowledge distillation*, dan arsitektur *neural network* yang efisien (misalnya MobileNet, EfficientNet). Implementasi di perangkat *edge* tidak hanya mengurangi ketergantungan pada infrastruktur cloud tetapi juga meningkatkan privasi dengan memproses data biometrik secara lokal tanpa perlu transmisi ke server pusat.

5. Pengembangan sandi visual yang lebih diskriminatif dengan memanfaatkan pola gerakan bibir unik pengguna

Cara ini berguna untuk meningkatkan keamanan sistem. Penelitian dapat mengeksplorasi kombinasi fonem-fonem tertentu yang menghasilkan pola gerakan bibir yang paling unik dan sulit ditiru, atau mengembangkan mekanisme *personalized password phrases* yang diadaptasi berdasarkan karakteristik fisiologis individu. Pendekatan ini mengintegrasikan aspek *something you know* (kata sandi) dengan *something you are* (pola biologis) secara lebih intrinsik.

Selain kelima arah tersebut, penelitian lanjutan juga dapat mengembangkan mekanisme adaptif yang dapat menyesuaikan bobot modalitas berdasarkan kondisi lingkungan, kualitas sensor, atau tingkat ancaman keamanan. Integrasi dengan modalitas biometrik tambahan seperti pola iris, suara, atau perilaku (*behavioral biometrics*) juga dapat dieksplorasi untuk sistem autentikasi yang lebih *robust* dan *multi-factor*.

Dengan demikian, penelitian ini tidak dimaksudkan sebagai solusi final, melainkan sebagai pijakan awal yang menyediakan kerangka konseptual, analisis literatur sistematis, dan demonstrasi *proof-of-concept* sebagai dasar untuk pengembangan sistem autentikasi biometrik visual generasi berikutnya. Temuan dan rekomendasi yang dihasilkan diharapkan dapat memandu penelitian-penelitian lanjutan di bidang ini, khususnya dalam konteks kebutuhan keamanan siber Indonesia yang semakin kompleks dan mendesak.

IV. KESIMPULAN

Penelitian ini menegaskan bahwa sistem autentikasi biometrik unimodal masih memiliki keterbatasan mendasar, khususnya dalam menghadapi variasi kondisi lingkungan dan ancaman pemalsuan visual. Kajian literatur sistematis yang dilakukan menunjukkan bahwa biometrik visual multimodal, yang mengintegrasikan pengenalan wajah statis dan gerakan bibir dinamis, merupakan pendekatan yang lebih *robust* dan menjanjikan untuk meningkatkan keamanan dan ketahanan sistem autentikasi.

Analisis komparatif lintas studi mengindikasikan bahwa fusi level skor merupakan strategi paling fleksibel untuk menggabungkan modalitas biometrik heterogen, karena menyeimbangkan kompleksitas implementasi dengan potensi peningkatan akurasi. Selain itu, arsitektur deep learning berbasis Transformer memiliki potensi signifikan dalam memodelkan pola gerakan bibir yang bersifat temporal, mendukung pemrosesan urutan dinamis secara efisien. Kajian ini juga menekankan pentingnya deteksi keaslian berbasis dinamika visual, seperti analisis tekstur, kedalaman 3D, refleksi cahaya, dan *micro-expressions*, untuk menghadapi ancaman presentasi, termasuk serangan *visual-only deepfake*.

Namun, perlu ditegaskan bahwa penelitian ini bersifat kajian awal (*preliminary analysis*) yang dipadukan dengan eksperimen *proof-of-concept* simulasi. Sistem yang dikembangkan masih dalam skala terbatas dan belum diuji secara eksperimental komprehensif di lingkungan nyata. Oleh karena itu, temuan yang disajikan lebih bersifat landasan konseptual dan peta peluang riset, bukan sebagai klaim peningkatan performa numerik final.

Temuan ini membuka beberapa arah pengembangan riset masa depan, antara lain:

1. Implementasi sistem nyata dengan pengujian eksperimental berskala besar.
2. Eksplorasi arsitektur Transformer untuk pemodelan gerakan bibir dan integrasi fusi multimodal.
3. Pemanfaatan *self-supervised learning* untuk efisiensi penggunaan data biometrik.
4. Optimasi model pada perangkat *edge* untuk mendukung implementasi terdesentralisasi.
5. Pengembangan sandi visual yang lebih diskriminatif, memanfaatkan pola gerakan bibir unik pengguna untuk meningkatkan keamanan.

Dengan demikian, penelitian ini tidak dimaksudkan sebagai solusi final, melainkan sebagai pijakan awal yang menyediakan kerangka konseptual, analisis literatur, dan demonstrasi *proof-of-concept* sebagai dasar untuk pengembangan sistem autentikasi biometrik visual generasi berikutnya.

REFERENSI

[1] A. Roihan et al., “Perancangan Purwarupa Sistem Keamanan Kunci Pintu Berbasis Pengenalan Wajah”, *Journal of Innovation And Future Technology (IFTECH)*, vol. 6, no. 2, pp. 234–242, Aug. 2024, doi: 10.47080/iftech.v6i2.3415.

[2] A. Ross and A. K. Jain, “Multimodal biometrics: An overview,” *Proc. 12th European Signal Processing Conference*, 2004. doi: 10.1109/EUSIPCO.2004.7075379.

[3] F. Schroff, D. Kalenichenko, and J. Philbin, “FaceNet: A unified embedding for face recognition and clustering,” *Proc. IEEE Conf. Computer Vision and Pattern Recognition (CVPR)*, 2015, pp. 815–823. doi: 10.1109/CVPR.2015.7298682.

[4] R. Raghavendra and C. Busch, “Presentation attack detection methods for face recognition systems: A comprehensive survey,” *IEEE Access*, vol. 7, pp. 100–132, 2019. doi: 10.1109/ACCESS.2019.2929433.

[5] J. S. Chung, A. Senior, O. Vinyals, and A. Zisserman, “Lip reading sentences in the wild,” *Proc. IEEE Conf. Computer Vision and Pattern Recognition (CVPR)*, 2017, pp. 3444–3453. doi: 10.1109/CVPR.2017.367.

[6] T. Afouras, J. S. Chung, and A. Zisserman, “Deep lip reading: A comparison of models and an online application,” *IEEE/ACM Trans. Audio, Speech, and Language Processing*, vol. 26, no. 12, pp. 2246–2260, 2018. doi: 10.1109/TASLP.2018.2810139.

[7] J. Yang, A. Waibel, and A. Jain, “Biometric recognition: Challenges and opportunities,” *IEEE Computer*, vol. 44, no. 1, pp. 74–80, 2011. doi: 10.1109/MC.2010.384.

[8] P. K. Ratha, K. Ricanek, and M. Savvides, “Score level fusion of multimodal biometrics using triangular norms,”

Pattern Recognition Letters, vol. 32, no. 14, pp. 1843–1850, 2011. doi:10.1016/j.patrec.2011.06.029.

[9] O. N., Kadhim, M. H., Abdulameer, Y. M. H., Al-Mayali, "A multimodal biometric system for iris and face traits based on hybrid approaches and score level fusion," *In BIO Web of Conferences*, vol. 97, pp. 00016, 2024. doi: 10.1051/bioconf/20249700016.

[10] S. R. B. Kisku, J. S. Chang, and A. Kumar, "Multimodal biometrics: Weighted score level fusion based on non-ideal iris and face images," *Expert Systems with Applications*, vol. 41, no. 11, pp. 5390–5404, 2014. doi:10.1016/j.eswa.2014.02.051.

[11] M. He et al., "Performance evaluation of score level fusion in multimodal biometric systems," *Pattern Recognition*, vol. 43, no. 5, pp. 1789–1800, 2010. doi:10.1016/j.patcog.2009.11.018.

[12] A. Naseem et al., "Robust multimodal biometric system based on optimal score level fusion model," *Expert Systems with Applications*, vol. 116, pp. 364–376, 2019. doi:10.1016/j.eswa.2018.08.036.

[13] S. Tharewal et al., "Score-Level Fusion of 3D Face and 3D Ear for Multimodal Biometric Human Recognition," *Computational Intelligence and Neuroscience*, 2022, Art. no. 3019194. doi:10.1155/2022/3019194.

[14] S. N. Garg, R. Vig, and S. Gupta, "A Survey on Different Levels of Fusion in Multimodal Biometrics," *Indian Journal of Science and Technology*, vol. 10, no. 44, pp. 1–11, 2017. doi:10.17485/ijst/2017/v10i44/120575.

[15] F. Wang and J. Han, "Multimodal biometric authentication based on score level fusion using support vector machine," *Opto-Electronics Review*, vol. 17, no. 1, pp. 59–64, 2009. doi:10.2478/s11772-008-0054-8.

[16] F., Shafizadegan, A. R., Naghsh-Nilchi, E. Shabaninia, "Multimodal vision-based human action recognition using deep learning: a review," *Artificial Intelligence Review*, vol. 57, 178, 2024. doi:10.1007/s10462-024-10730-5