

Rancang Bangun Sistem *Computer Vision* Untuk Analisis Perilaku Multimodal Menggunakan Arsitektur *Hybrid YOLO* Dan *FACE-API.JS* Dengan *Tracking Manhattan Distance*

Rico Mesias Tamba, Dewi Yanti Liliana, Mera Kartika Delimayanti, Euis Oktavianti, Iik Muhammad Malik Matin

Teknik Informatika, Teknik Informatika dan Komputer, Politeknik Negeri Jakarta Depok, Jawa Barat
rico.mesias.tamba.tik22@mhs.w.pnj.ac.id, dewiyanti.liliana@tik.pnj.ac.id, mera.kartika@tik.pnj.ac.id,
euis.oktavianti@tik.pnj.ac.id, iik.muhamad.malik.matin@tik.pnj.ac.id

Abstract - Manual analysis of human behavior is susceptible to bias, subjectivity, and fatigue. This research aims to develop a web-based computer vision system to monitor multimodal behavior (actions, attention, and emotions) in real-time. This system implements a hybrid architecture that efficiently distributes the computational load: the YOLOv11 model is executed on the server for physical action detection, while Face-API.js runs on the client side for facial feature extraction. Furthermore, the Manhattan Distance algorithm is implemented for inter-frame object tracking with a low computational load. Based on the evaluation results, the YOLOv11 model achieved an mAP@50 accuracy of 0.607, and Face-API.js achieved a Macro Average F1-score of 58.93%. The computational load distribution proved optimal, with an average processing speed of 18 FPS. The network transmission quality (QoS) over WebSocket was very stable with 0% packet loss, a Round-Trip Time (RTT) of 679 ms, and 12.45 ms jitter. User Acceptance Testing (UAT) results yielded an effectiveness rate of 88.9%, efficiency of 97.8%, and satisfaction rating of 100%. In conclusion, this hybrid architecture system successfully operated stably and responsively, making it highly suitable for use as an efficient multimodal behavior monitoring solution.

Keywords: Multimodal Analysis, Computer Vision, YOLOv11, Face-API.js, Object Tracking.

Abstrak - Analisis perilaku manusia secara manual rentan terhadap bias subjektivitas dan kelelahan. Penelitian ini bertujuan mengembangkan sistem *Computer Vision* berbasis web untuk memantau perilaku multimodal (aksi, atensi, dan emosi) secara real-time. Sistem ini menerapkan arsitektur hybrid yang membagi beban komputasi secara efisien: model YOLOv11 dieksekusi pada server untuk deteksi aksi fisik, sedangkan Face-API.js dijalankan pada sisi klien untuk ekstraksi fitur wajah. Selain itu, algoritma Manhattan Distance diimplementasikan untuk pelacakan objek (object tracking) antar-frame dengan beban komputasi yang ringan. Berdasarkan hasil evaluasi, model YOLOv11 mencapai akurasi mAP@50 sebesar 0,607 dan Face-API.js memperoleh Macro Average F1-score 58,93%. Distribusi beban komputasi terbukti optimal dengan rata-rata kecepatan pemrosesan mencapai 18 FPS. Kualitas transmisi jaringan (QoS) melalui WebSocket berjalan sangat stabil dengan packet loss 0%, Round-Trip Time (RTT) 679 ms, dan jitter 12,45 ms. Pengujian User Acceptance Test (UAT) menghasilkan tingkat efektivitas 88,9%, efisiensi 97,8%, dan kepuasan 100%. Kesimpulannya, sistem arsitektur hibrida ini berhasil beroperasi secara stabil, responsif, dan sangat layak digunakan sebagai solusi pemantauan perilaku multimodal yang efisien.

Kata kunci: Analisis Multimodal, Computer Vision, YOLOv11, Face-API.js, Pelacakan Objek.

I. PENDAHULUAN

Perkembangan *Computer Vision* telah mentransformasi cara sistem memahami manusia, bergeser dari pengenalan objek visual menuju *Human Behavior Analysis* kemampuan mengintegrasikan berbagai modalitas data untuk menafsirkan tindakan dan kondisi emosional manusia secara otomatis. Kemampuan ini

menjadikan analisis perilaku instrumen vital di berbagai sektor industri untuk menciptakan sistem yang responsif dan mampu memitigasi risiko secara real-time.

Urgensi teknologi ini terlihat jelas pada lingkungan ritel dan pusat perbelanjaan, mengingat respons bawah sadar konsumen sering kali lebih mendominasi pengambilan keputusan dibandingkan

respons sadar [1]. Implementasinya kini tidak lagi terbatas pada keamanan, tetapi juga mampu mengukur fiksasi visual pada titik pembelian untuk memprediksi sikap konsumen baik perilaku mendekat (approach) maupun menghindari (withdrawal) terhadap suatu produk [1].

Teknologi ini telah didiversifikasi ke berbagai domain krusial: Human Activity Recognition, keamanan publik, kesehatan mental, mobilitas dan transportasi, produktivitas kerja, pendidikan, hingga pemasaran dan perilaku konsumen [7], sebagaimana ditampilkan pada Gambar 1.



Gambar 1 Human Behaviour Analytics Applications

Penelitian ini berfokus pada pengembangan sistem yang menganalisis tiga atribut perilaku secara simultan Aksi (gerakan tubuh), Atensi (arah pandangan), dan Emosi (ekspresi wajah) melalui hybrid architecture yang mengombinasikan YOLO dan Face-api.js. YOLO mendeteksi objek tubuh manusia dan mengklasifikasikan aksi secara real-time, sementara Face-api.js mengekstraksi fitur wajah untuk menentukan emosi dan atensi. Kontinuitas identitas antar-frame dijaga melalui algoritma Manhattan Distance yang menghitung jarak perpindahan objek antar-frame.

Studi "YOLO-Act: Unified Spatiotemporal Detection of Human Actions Across Multi-Frame Sequences" [2], mengembangkan metode deteksi aksi presisi menggunakan modifikasi YOLOv11 dan strategi three-keyframes. Namun, penelitian tersebut bersifat single-modality pada aksi fisik tanpa mempertimbangkan ekspresi wajah dan atensi, serta memiliki beban komputasi tinggi akibat pemrosesan korelasi fitur antar-frame secara volumetrik, sehingga berpotensi meningkatkan latensi pada perangkat berspesifikasi terbatas.

Merespons keterbatasan tersebut, penelitian ini mengusulkan pendekatan multimodal yang mengintegrasikan analisis Aksi, Atensi, dan Emosi

secara bersamaan, dengan Manhattan Distance (L1 Norm) sebagai solusi mereduksi kompleksitas komputasi. Efisiensi matematisnya yang hanya menggunakan operasi penjumlahan selisih absolut memungkinkan sistem mempertahankan frame rate (FPS) stabil untuk analisis real-time tanpa mengorbankan kontinuitas pelacakan objek.

II. METODOLOGI PENELITIAN

Penelitian ini berfokus pada pengembangan sistem pengenalan perilaku multimodal secara real-time yang mengintegrasikan analisis afektif, atensi visual, dan aktivitas fisik. Pendekatan yang digunakan bersifat kuantitatif, dengan validasi keandalan sistem berdasarkan dua metrik evaluasi utama, yaitu akurasi dan frame rate (FPS). Penelitian dilaksanakan melalui lima tahapan sistematis yang mengacu pada AI Development Life Cycle, sebagaimana ditampilkan pada Gambar 2.



Gambar 2 Development Phase Pipeline

A. Identifikasi Kebutuhan

Tahap awal berfokus pada pendefinisian parameter atribut yang harus dikenali sistem, yang dikategorikan menjadi tiga dimensi.

1. Pengenalan Aksi Manusia (*Human Action Recognition*): mendeteksi 11 jenis aktivitas fisik, Drinking, Fall-Detected, Fall_down, Lying_down, Nearly_fall, Sit Down, Sitting, Standing, Walking, Walking_on_Stairs, dan Crawling.
2. Pengenalan Ekspresi Emosi (*Facial Expression Recognition*): mengidentifikasi 7 kelas emosi dasar, Happy, Sad, Angry, Disgusted, Surprised, Neutral, dan Fear.
3. Estimasi Atensi Visual (*Head Pose Estimation*): memantau orientasi kepala pada dua sumbu rotasi Yaw (menoleh kiri/kanan) dan Pitch (menunduk/mendongak).

B. Pengumpulan Data

Dataset dikumpulkan dari sumber publik: data aktivitas manusia dari Roboflow (sudah berannotasi bounding box), serta data citra wajah dan label emosi dari Kaggle FER2013.

C. Pemrosesan Data

Data melalui tahap pre-processing meliputi resizing agar sesuai layer input, normalisasi piksel

untuk mempercepat komputasi, serta splitting menjadi data train dan test.

D. Pengembangan dan Pelatihan Model

Tahap ini berfokus pada pembangunan arsitektur Deep Learning, khususnya model pengenalan aksi, melalui hyperparameter tuning (learning rate, batch size, epoch) untuk mencapai konvergensi optimal dengan akurasi tinggi dan loss rendah.

1. Computer Vision

Computer Vision adalah domain Deep Learning yang memproses data visual sebagai array intensitas piksel [4], mencakup dua tugas utama: Image Classification (memberi label pada gambar) dan Object Detection (melokalisasi sekaligus mengklasifikasikan objek) [3].

2. You Only Look Once (YOLO)

YOLO adalah arsitektur Object Detection yang memproses seluruh gambar dalam satu forward pass untuk memprediksi bounding box dan kelas secara simultan, sehingga efisien untuk deteksi real-time [15].

3. Face-api.js

Face-api.js adalah pustaka JavaScript untuk analisis wajah client-side, dibangun di atas TensorFlow.js dengan akselerasi GPU melalui WebGL.

E. Implementasi Sistem

Pelacakan objek bertujuan mempertahankan identitas target secara konsisten antar-frame, dengan asosiasi data sebagai tahap paling kritis, mencocokkan prediksi posisi objek dari frame sebelumnya dengan deteksi baru [13]. Metrik yang digunakan adalah Manhattan Distance (L1 Norm), yang mengukur jarak dua titik berdasarkan jumlah selisih absolut koordinatnya, dirumuskan sebagai berikut (1)

$$C(P, D) = \sum_{i=1}^n |P_i - D_i| \quad (1)$$

Di mana P adalah vektor prediksi posisi objek yang dilacak, D adalah vektor deteksi posisi objek baru, dan n adalah jumlah dimensi fitur spasial yang dibandingkan. Metrik ini dipilih karena hanya melibatkan operasi penjumlahan dan pengurangan sederhana tanpa akar kuadrat atau perkalian kompleks, sehingga lebih ringan secara komputasi

dibanding Euclidean Distance dan cocok untuk pemrosesan cepat.

Setelah model dikembangkan, sistem mengimplementasikan Manhattan Distance untuk dua keperluan: Head Tracking (mengasosiasikan deteksi wajah antar-frame untuk memantau atensi dan ekspresi) dan Action Tracking (mengasosiasikan deteksi aktivitas tubuh), guna menjaga proses matching yang efisien dengan latensi rendah.

F. Pengujian dan Evaluasi

Tahap akhir penelitian adalah validasi kinerja sistem menggunakan dua pendekatan evaluasi kuantitatif:

1. Evaluasi Akurasi: menggunakan Confusion Matrix untuk mengukur Accuracy, Precision, Recall, dan F1-score.
2. Evaluasi Kinerja Sistem: mengukur rata-rata FPS dan waktu latensi (inference time) saat sistem berjalan real-time.

III. HASIL DAN PEMBAHASAN

A. Hasil Pelatihan Model

Model Artificial Intelligence yang dirancang telah berhasil diimplementasikan dan dilatih menggunakan dataset yang dipersiapkan, mencakup pengembangan model deteksi dan klasifikasi aksi manusia menggunakan arsitektur YOLOv11, serta integrasi analisis wajah dan pengenalan emosi menggunakan Face-api.js.

1. Pelatihan Model YOLOv11

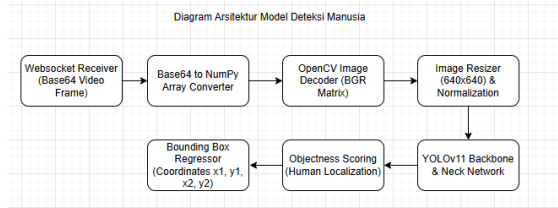
Tahap ini merupakan implementasi pelatihan model deteksi aktivitas manusia menggunakan arsitektur YOLOv11 varian Nano (yolo11n.pt), dipilih karena bobotnya ringan dan mampu melakukan inferensi cepat tanpa mengorbankan performa deteksi. Model memanfaatkan bobot awal (pre-trained weights) untuk mempercepat konvergensi dan meningkatkan generalisasi terhadap kelas aktivitas manusia seperti walking, standing, sitting, falling, dan lying down.

Dataset yang telah dianotasi dibagi menjadi tiga subset (train, validation, test) dan dilatih menggunakan framework Ultralytics YOLO pada Google Colaboratory dengan dukungan GPU Tesla T4. Proses pelatihan mencakup:

- a. Pemuatan dataset aktivitas manusia yang telah dianotasi

- b. Konfigurasi dan *fine-tuning* model YOLOv11 menggunakan parameter pelatihan tertentu
 - c. Evaluasi performa model menggunakan metrik *Precision*, *Recall*, *F1-score*, dan *mean Average Precision (mAP)*
2. Model Deteksi Manusia

Tahap awal pemrosesan sisi server berfokus pada object localization untuk mendeteksi keberadaan manusia dalam frame video stream, sebagaimana ditunjukkan pada Gambar 3.

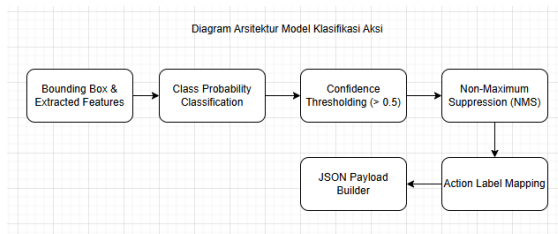


Gambar 3 Arsitektur Model Deteksi Manusia

Alur data bermula dari Input Layer, di mana peladen WebSocket menerima data gambar berformat Base64 dari klien, dikonversi menjadi array numerik pada Decoding Layer, lalu diterjemahkan OpenCV Image Decoder menjadi matriks citra. Citra kemudian melewati Preprocessing Layer untuk disesuaikan menjadi 640×640 piksel dan dinormalisasi, sebelum diekstraksi fiturnya melalui struktur Backbone dan Neck YOLOv11. Hasil ekstraksi divalidasi pada Objectness Scoring, dan diakhiri Bounding Box Regressor yang menghasilkan koordinat kotak pembatas absolut (x1, y1, x2, y2) sebagai dasar klasifikasi aks.

3. Model Klasifikasi Aksi

Setelah lokasi objek manusia dipetakan, pipeline dilanjutkan dengan klasifikasi aksi berdasarkan fitur visual di dalam bounding box, sebagaimana diilustrasikan pada Gambar 4.



Gambar 4 Arsitektur Model Klasifikasi Aksi

Koordinat bounding box dan fitur hasil ekstraksi diteruskan ke Prediction Head untuk menghitung probabilitas kelas aksi. Hasil prediksi

disaring melalui Confidence Thresholding (hanya prediksi >50% dipertahankan), dilanjutkan Non-Maximum Suppression (NMS) untuk mengeliminasi kotak prediksi ganda. Prediksi akhir diterjemahkan menjadi label aksi tekstual (misalnya "Berdiri", "Duduk", "Terjatuh"), lalu dikemas oleh JSON Payload Builder untuk dikirim ke klien melalui WebSocket.

4. Inisialisasi Model Face-API.js

Implementasi dimulai dengan menginisialisasi Face-API.js di sisi klien, memuat model Deep Learning (deteksi wajah, landmark, klasifikasi emosi) secara asinkron via Promise.all agar loading berjalan paralel. Bersamaan dengan itu, sistem mengakses kamera klien melalui WebRTC (`navigator.mediaDevices.getUserMedia`) untuk menangkap video stream. Model yang digunakan dirangkum pada Tabel I.

TABEL I. Model Face-API.js

Model Face-API.js	Fungsi
TinyFaceDetector	Mendeteksi area wajah pada frame video secara real-time dengan beban komputasi ringan.
FaceLandmark68Net	Memetakan 68 titik landmark anatomi wajah (mata, hidung, mulut, rahang).
FaceExpressionNet	Mengklasifikasikan probabilitas ekspresi emosi wajah pengguna.

TinyFaceDetector dipilih sebagai pendeteksi utama karena ringan dan cepat, menjaga stabilitas frame rate. FaceLandmark68Net menyuplai koordinat titik wajah untuk analisis head pose, sementara FaceExpressionNet beroperasi paralel mengidentifikasi kondisi emosional dominan pengguna (netral, senang, sedih, marah).

5. Ekstraksi *Face landmarks*, Arah Wajah, dan Emosi

Setelah inisialisasi selesai, sistem mengeksekusi ekstraksi fitur wajah secara real-time dari aliran video frame-by-frame, mencakup lokalisasi wajah, ekstraksi koordinat landmark, klasifikasi emosi, dan estimasi arah kepala. Koordinat wajah digunakan untuk menghitung arah hadap (Yaw dan Pitch) berdasarkan rasio jarak antar titik (mata, hidung, dagu), serta menganalisis perubahan bentuk fitur wajah untuk klasifikasi emosi. Fitur yang diekstraksi dirangkum pada Tabel II.

TABEL II. Fitur Analisis Wajah

Fitur Analisis Wajah	Fungsi
Face Detection	Melokalisasi dan mendeteksi keberadaan wajah pada setiap frame (bounding box) sebagai langkah awal.
68-Point Face Landmarks	Mengekstraksi 68 titik koordinat anatomi wajah (mata, hidung, mulut, garis rahang).
Head Pose Estimation	Menghitung rasio Yaw dan Pitch untuk menentukan status tingkat atensi.
Emotion Recognition	Mengidentifikasi dan mengklasifikasikan probabilitas ekspresi emosi secara real-time.

Berdasarkan Tabel II, Face Detection dijalankan sebagai gerbang utama sebelum wajah tervalidasi diproses lebih lanjut melalui 68-Point Face Landmarks. Koordinat landmark tersebut digunakan untuk dua fungsi: (1) Head Pose Estimation, menentukan arah hadap (fokus, menoleh, menunduk, mendongak) sebagai indikator atensi kognitif; dan (2) Emotion Recognition, mengklasifikasikan ekspresi ke dalam kategori emosi dasar (neutral, happy, sad, angry, surprised).

6. Rendering Visualisasi dan Overlay Aktif

Tahap akhir pemrosesan sisi klien adalah rendering overlay interaktif menggunakan elemen HTML5 <canvas>, yang menyinkronkan payload deteksi aksi (YOLOv11) dari peladen dan hasil analitik wajah (Face-API.js) yang diproses lokal, sehingga seluruh informasi, bounding box, landmark, arah pandangan, probabilitas emosi, hingga Tracking ID, divisualisasikan serentak. Sistem menerapkan color coding berbeda pada bounding box dan label untuk menghindari cognitive overload, dieksekusi berulang per frame agar overlay tetap presisi. Komponen visualisasi dirangkum pada Tabel III.

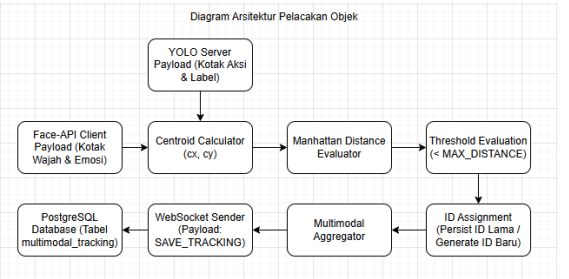
TABEL III. Komponen Visualisasi

Komponen Visualisasi	Fungsi
Bounding Box YOLOv11	Menampilkan area deteksi aksi manusia hasil inferensi YOLOv11.
Bounding Box Wajah	Menampilkan area deteksi wajah hasil analisis Face-API.js.
Face Landmarks	Menampilkan titik-titik landmark wajah pengguna.
Tracking ID	Menampilkan identitas unik wajah yang dipantau.
Label Emosi	Menampilkan hasil klasifikasi ekspresi emosi.
Status Fokus	Menampilkan informasi fokus berdasarkan arah wajah.
Overlay Canvas	Media visualisasi seluruh hasil analisis secara real-time.

Bounding Box YOLOv11 dan Bounding Box Wajah berfungsi sebagai indikator spasial untuk melacak pergerakan subjek, dengan Face Landmarks digambar dinamis di dalam area wajah. Pada header bounding box, sistem merender Tracking ID, Label Emosi, dan Status Fokus sebagai indikator kondisi kognitif subjek. Seluruh elemen ini dilukis pada satu Overlay Canvas transparan, menghasilkan dasbor pemantauan multimodal yang informatif, interaktif, dan ringan dijalankan di peramban.

B. Hasil Implementasi Pelacakan Objek

Untuk menjaga persistensi identitas subjek antar-frame, sistem mengimplementasikan Object Tracking menggunakan Manhattan Distance, sebagaimana diilustrasikan pada Gambar 5.



Gambar 5 Arsitektur Pelacakan Objek

Alur integrasi bermula dari dua sumber data spasial YOLO Server Payload dan Face-API Client Payload yang masuk ke Centroid Calculator untuk menentukan titik pusat (cx, cy) kotak pembatas objek. Titik pusat tersebut diteruskan ke Manhattan Distance Evaluator untuk menghitung selisih jarak

absolut terhadap posisi objek pada frame sebelumnya. Manhattan Distance dipilih karena beban komputasinya lebih ringan dibanding Euclidean Distance, ideal dijalankan di sisi klien. Parameter utama tracking dirangkum pada Tabel IV.

TABEL IV. Parameter Tracking

Parameter Tracking	Fungsi
Tracking ID	Identitas persisten dan unik untuk setiap wajah/subjek yang terdeteksi.
Centroid Wajah	Titik acuan (C_x , C_y) posisi pusat wajah pada frame video.
Manhattan Distance	Nilai kuantitatif perpindahan posisi titik pusat objek antar-frame.
Threshold Tracking	Batas maksimum toleransi perpindahan wajah agar tetap dianggap objek yang sama.
Frame Missing Counter	Menghapus tracking otomatis jika wajah gagal terdeteksi (oklusi) beberapa frame berturut-turut.

Berdasarkan Tabel IV, Tracking ID dikelola pada lapisan penugasan identitas, di mana saat proses matching sistem membandingkan Manhattan Distance antar-titik Centroid dan mempertahankan Tracking ID lama jika jarak perpindahan masih di bawah Threshold Tracking, sementara untuk menangani anomali visibilitas (wajah terhalang atau keluar jangkauan kamera) sistem menerapkan Frame 6 penumpukan data objek tidak aktif; pada tahap final, objek yang identitasnya telah tervalidasi dimasukkan ke Multimodal Aggregator yang menyatukan Tracking ID dengan fitur aksi dan emosi menjadi satu payload, lalu mengirimkannya melalui WebSocket ke basis data PostgreSQL sebagai riwayat analitik perilaku yang utuh.

C. Hasil Pengujian Model

Pada tahap ini dilakukan dua pengujian, yaitu pengujian model YOLO dan pengujian model Face-API.

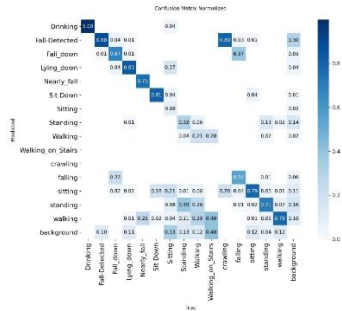
1. Hasil Model YOLO

Pengujian dilakukan menggunakan metrik Precision, Recall, F1-score, dan mAP@50, dengan hasil dirangkum pada Tabel V.

TABEL V. Hasil Evaluasi Pengujian Model YOLOv11

Metrik	Evaluasi
Mean Precision	0.618
Mean Recall	0.611
Mean F1-score	0.614
mAP@50	0.607

Berdasarkan Tabel V, model YOLOv11 memperoleh Mean Precision 0,618 dan Mean Recall 0,611, menunjukkan ketepatan dan kemampuan model yang cukup baik dalam mendeteksi objek manusia beserta aksinya. Mean F1-score 0,614 menunjukkan keseimbangan performa precision dan recall, sementara mAP@50 sebesar 0,607 menunjukkan kemampuan deteksi yang cukup baik pada ambang IoU 50%. Secara keseluruhan, model YOLOv11 menunjukkan performa yang stabil untuk kebutuhan sistem monitoring real-time. Visualisasi hasil pengujian ditunjukkan pada Gambar 6.



Gambar 6 Confusion matrix Model YOLOv11

Gambar 6 menampilkan Normalized Confusion Matrix, di mana nilai diagonal utama menunjukkan prediksi benar sedangkan nilai di luar diagonal menunjukkan kesalahan klasifikasi. Kelas seperti Drinking, Fall-Detected, Lying_down, Sit Down, dan Walking menunjukkan klasifikasi yang cukup baik, namun kesalahan masih terjadi pada aktivitas berpola gerakan serupa (standing, walking, sitting, falling) serta kelas background akibat kemiripan visual dan variasi posisi objek. Secara umum, model YOLOv11 mampu mengenali sebagian besar aktivitas manusia dengan performa cukup baik untuk implementasi sistem real-time monitoring.

2. Hasil Face-API.js

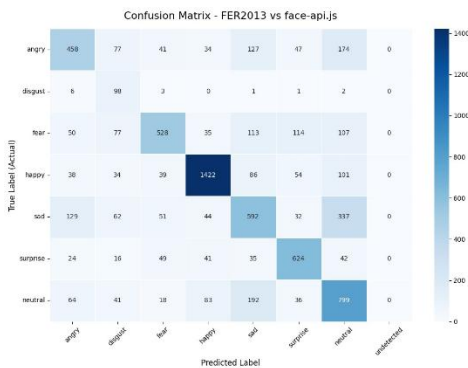
Pengujian dilakukan terhadap model Tiny Face Detector (TinyFD) menggunakan dataset FER2013, dipilih karena memiliki variasi ekspresi, pencahayaan, pose, dan kualitas citra yang

menyerupai kondisi penggunaan nyata. Evaluasi menggunakan metrik precision, recall, dan F1-score pada tujuh kategori emosi (angry, disgust, fear, happy, sad, surprise, neutral), dihitung hanya pada wajah yang berhasil terdeteksi. Hasil pengujian ditunjukkan pada Tabel VI.

TABEL VI. Hasil Pengujian Model TinyFD

Label Emosi	Precision	Recall	F1-Score
Angry	59,56	59,56	59,56
Disgust	24,20	24,20	24,20
Fear	72,43	72,43	72,43
Happy	85,71	85,71	85,71
Sad	51,66	51,66	51,66
Surprise	68,72	68,72	68,72
Neutral	51,15	51,15	51,15
Macro Average	59,06	59,06	59,06

Model TinyFD menunjukkan performa cukup baik dengan macro average precision 59,06%, recall 65,03%, dan F1-score 58,93%. Emosi happy memperoleh performa terbaik (F1-score 82,84%), diikuti surprise (F1-score 71,77%). Sebaliknya, disgust memiliki precision terendah (24,20%) meski recall tinggi (88,29%), mengindikasikan banyak false positive, sementara sad dan neutral juga menunjukkan kesalahan klasifikasi relatif tinggi akibat kemiripan visual dengan emosi lain. Distribusi prediksi divisualisasikan pada Gambar 7.



Gambar 7 Evaluasi Model TinyFD

Confusion matrix menunjukkan sumbu vertikal sebagai label aktual dan sumbu horizontal sebagai hasil prediksi. Emosi happy memiliki jumlah prediksi benar tertinggi, diikuti neutral dan surprise yang juga stabil. Namun, masih terdapat kesalahan klasifikasi antar-kategori, misalnya sad yang sering

diprediksi sebagai neutral, serta angry yang kadang diprediksi sebagai sad atau neutral. Secara keseluruhan, TinyFD dan Face-API.js mampu melakukan deteksi wajah dan pengenalan emosi dengan cukup baik pada dataset FER2013.

IV. KESIMPULAN DAN SARAN

Penelitian ini berhasil mengembangkan sistem multimodal behaviour analytics yang mengintegrasikan deteksi aksi manusia (YOLOv11), analisis wajah dan emosi (Face-API.js), serta pelacakan objek berbasis Manhattan Distance secara real-time. Model YOLOv11 memperoleh mAP@50 sebesar 0,607 dengan performa yang cukup stabil, sementara model TinyFD pada Face-API.js memperoleh macro average F1-score sebesar 58,93% dalam klasifikasi emosi. Kedua modalitas tersebut berhasil diintegrasikan melalui mekanisme tracking berbasis Manhattan Distance yang ringan secara komputasi namun tetap menjaga persistensi identitas objek antar-frame, sehingga sistem mampu menyajikan analisis aksi, atensi, dan emosi secara bersamaan dalam satu dasbor pemantauan interaktif.

Untuk pengembangan selanjutnya, disarankan menambah variasi dan jumlah data latih pada kelas-kelas yang masih menunjukkan tingkat kesalahan klasifikasi tinggi guna meningkatkan akurasi model secara keseluruhan. Selain itu, penelitian lanjutan dapat mengeksplorasi arsitektur atau model alternatif yang lebih ringan namun tetap akurat, serta menguji performa sistem pada kondisi lingkungan yang lebih beragam agar hasil implementasi lebih representatif untuk digunakan pada skenario dunia nyata.

V. REFERENSI

- [1] Alsharif, A.H. et al. (2022) 'Consumer Behaviour to Be Considered in Advertising: A Systematic Analysis and Future Agenda', Behavioral Sciences, 12(12). Available at: <https://doi.org/10.3390/bs12120472>.
- [2] Alzahrani, N., Bchir, O. and Ismail, M.M. Ben (2025) 'YOLO-Act: Unified Spatiotemporal Detection of Human Actions Across Multi-Frame Sequences', Sensors, 25(10), pp. 1–24. Available at: <https://doi.org/10.3390/s25103013>
- [3] Alzubaidi, L. et al. (2021) Review of Deep Learning: concepts , CNN architectures , challenges , applications , future directions, Journal of Big Data. Springer International Publishing. Available at: <https://doi.org/10.1186/s40537-021-00444-8>.
- [4] Chai, J. et al. (2021) 'Deep Learning in Computer Vision: A critical review of emerging techniques and application scenarios', Machine Learning with Applications, 6(August), p. 100134. Available at: <https://doi.org/10.1016/j.mlwa.2021.100134>
- [5] Chun, S. et al. (2024) 'Bidirectional Regression for Monocular 6DoF Head Pose Estimation and Reference

- System Alignment', arXiv preprint arXiv:2407.14136 [Preprint].
- [6] Emanuel, A.W.R., Mudjihartono, P. and Nugraha, J.A.M. (2021) 'Snapshot-Based *Human Action Recognition* using OpenPose and *Deep Learning*', *IAENG International Journal of Computer Science*, 48(4), pp. 2–7.
 - [7] Essahraui (2025) 'Human Behavior Analysis: A Comprehensive Survey on Techniques , Applications , Challenges , and Future Directions', *IEEE Access*, PP, p. 1. Available at: <https://doi.org/10.1109/ACCESS.2025.3589938>.
 - [8] Garcia, B., Gallego, F., & Gortázar, F. (2020). Assessment of QoE for Video and Audio in WebRTC Applications Using Full-Reference Models. *Electronics*, 9(3), 462. MDPI.
 - [9] Goldberg, P. *et al.* (2021) 'Multimodal Engagement Analysis from Facial Videos in the Classroom', *IEEE TRANSACTION ON AFFECTIVE COMPUTING*, 3045(c), pp. 1–16. Available at: <https://doi.org/10.1109/TAFFC.2021.3127692>.
 - [10] Kisa, T., et al. (2023). "Impact of *Packet loss* Rate on Quality of Compressed High Resolution Videos". *Sensors*, 23(5), 2744. MDPI.
 - [11] Liliana, D.Y. *et al.* (2022) 'Emotion Intelligent Application for Independent Wellbeing Management using Certainty Factor', *2021 4th International Conference on Computer and Informatics Engineering (IC2IE)* [Preprint].
 - [12] Pabba, C., Bhardwaj, V. and Kumar, P. (2024) 'A visual intelligent system for students ' behavior classification using body pose and facial features in a smart classroom', *Multimedia Tools and Applications* (2024), pp. 36975–37005. Available at: <https://doi.org/10.1007/s11042-023-16388-5>.
 - [13] Razzok, M. *et al.* (2023) 'Pedestrian Detection and Tracking System Based on Deep-SORT , YOLOv5 , and New Data Association Metrics', *Information* 2023, pp. 1–16.
 - [14] Reddy, P. C. M., Yadav, N. S. K., Goud, P. A. K., Rekha, K. K., & Kumar, A. V. (2025). Real Time Object and Tracking Using *Deep Learning* and Opencv. *International Journal on Science and Technology (IJSAT)*, 16(1), 1-9.
 - [15] Terven, J., Córdova-Esparza, D.M. and Romero-González, J.A. (2023) 'A Comprehensive Review of YOLO Architectures in *Computer Vision*: From YOLOv1 to YOLOv8 and YOLO-NAS', *Machine Learning and Knowledge Extraction*, 5(4), pp. 1680–1716. Available at: <https://doi.org/10.3390/make5040083>.
 - [16] Tu, N. D., Muthanna, A., Khakimov, A., Kochetkova, I., Samouylov, K., Ateya, A. A., & Koucheryavy, A. (2025). SHARP-AODV: An Intelligent Adaptive Routing Protocol for Highly Mobile Autonomous Aerial Vehicle (AAV) Networks. *Sensors*, 25(24), 7522. MDPI.
 - [17] Wang, J., et al. (2023). "Latency Optimization in Edge-Cloud Collaborative Architecture for Real-time Video Analytics". *Sensors*, 23(11), 5120. MDPI.
 - [18] Warsuta, B. *et al.* (2024) 'Sistem Pengendalian Akses Berbasis Face-Recognition dengan Face-API . js dan Algoritma *Manhattan Distance*', *MULTINETICS*, 10(2), pp. 113–120.