

Peningkatan Performa Model Pengenalan Hitungan Jari (1–5) Berbasis Konteks Budaya Menggunakan Lokalisasi YOLO, *Transfer Learning*, dan Grad-CAM untuk *Explainable AI*

Khawaritzmi Abdallah Ahmad¹, Rahmat Syam², Mariani³, Ahmad Talib⁴

Departemen Matematika, Fakultas Matematika dan Ilmu Pengetahuan Alam, Universitas Negeri Makassar, Makassar, Indonesia^{1,2,3,4}

khawaritzmi.abdallah@unm.ac.id¹, rahmat.syam@unm.ac.id², mariani@unm.ac.id³, ahmadtalibunm@gmail.com⁴

Corresponding Author: Khawaritzmi Abdallah Ahmad, E-mail: khawaritzmi.abdallah@unm.ac.id

ABSTRACT

Pengenalan gestur hitungan jari (1–5) atau daktilonomi memiliki peran fundamental dalam pembelajaran matematika dasar sebagai representasi kognitif yang terwujud (*embodied numerosity*). Namun, model klasifikasi pada *Natural Hand Digit Dataset* masih menunjukkan akurasi terbatas (sekitar 66%) akibat gangguan latar belakang, variasi posisi, dan pencahayaan. Penelitian ini bertujuan untuk meningkatkan performa pengenalan gestur angka melalui lokalisasi tangan berbasis YOLOv10x yang dilatih pada dataset HaGRID sebagai tahap pra-pemrosesan, serta mengintegrasikan Grad-CAM untuk *explainable artificial intelligence* (XAI). Metode yang digunakan terdiri atas deteksi dan pemotongan wilayah tangan menggunakan YOLOv10x, kemudian klasifikasi jumlah jari menggunakan *transfer learning* ResNet50 dengan *fine-tuning* dan *callback*. Hasil eksperimen menunjukkan bahwa penggunaan citra hasil lokalisasi tangan (*cropped*) meningkatkan akurasi hingga 73,98% dengan *macro F1-score* 72,44%, jauh lebih baik dibandingkan citra asli (akurasi 41,66%). Visualisasi Grad-CAM mengonfirmasi bahwa model benar-benar fokus pada konfigurasi jari yang relevan pada prediksi benar, sementara pada prediksi salah, peta aktivasi mengungkap penyebab kesalahan (misalnya angka "2" dikenali sebagai "1" karena fokus pada satu jari). Penelitian ini menyimpulkan bahwa pendekatan lokalisasi tangan sebelum klasifikasi efektif meningkatkan performa dan ketahanan sistem, sementara Grad-CAM memberikan transparansi dan interpretabilitas yang penting untuk membangun kepercayaan dalam aplikasi pembelajaran matematika interaktif.

KEYWORDS

Finger Counting, Lokalisasi Tangan, YOLOv10x, *Transfer Learning*, *Explainable AI*

ARTICLE DOI:

<https://doi.org/10.53696/venn.v5i3.493>

A. Pendahuluan

Perhitungan jari atau daktilonomi merupakan salah satu manifestasi paling fundamental dari kognisi numerik manusia yang memfasilitasi pemahaman konsep angka abstrak melalui interaksi sensorimotor (di Luca & Mauro, 2011; Pika et al., 2009). Dalam pendidikan matematika dasar, praktik ini memainkan peran fungsional sebagai representasi kognitif yang terwujud (*embodied numerosity*) untuk membangun prinsip kardinalitas dan ordinalitas pada anak-anak (di Luca & Mauro, 2011). Seiring dengan pesatnya transformasi digital, perkembangan teknologi *Computer Vision* dan pembelajaran mendalam (*deep learning*) kini memungkinkan gestur jari diadopsi sebagai modalitas input visual yang natural dalam sistem *Human-Computer Interaction* (HCI) (Ahmad et al., 2025; Foteinos et al., 2026). Integrasi ini membuka peluang luas bagi pengembangan media pembelajaran matematika interaktif yang mampu merespons aktivitas motorik siswa secara *real-time*, sehingga menciptakan pengalaman belajar yang lebih konkret dan intuitif.

Upaya untuk mengotomatisasi pengenalan daktilonomi sebagai alat bantu pembelajaran telah diinisiasi melalui pembangunan pilot dataset yang dikenal sebagai *Natural Hand Digit Dataset*

(Ahmad et al., 2025). Penelitian terdahulu tersebut berfokus pada konstruksi basis data gestur angka satu hingga lima yang dikontekstualisasikan secara budaya guna menghindari bias universal dalam pengenalan *pose* (Ahmad et al., 2025). Proses ekstraksi fitur pada tahap rintisan ini mengandalkan lokalisasi fitur tangan dengan koordinat titik tangan (*hand landmarks*) yang diperoleh melalui *framework MediaPipe Hands*, yang telah terbukti efisien untuk pemrosesan *real-time* (Ahmad et al., 2023). Keberadaan dataset rintisan tersebut memberikan kontribusi krusial sebagai fondasi awal bagi pengembangan sistem pengenalan angka yang lebih inklusif dan adaptif dalam ekosistem pendidikan matematika digital (Ahmad et al., 2025).

Meskipun penelitian rintisan telah memberikan landasan yang kuat, evaluasi performa pada *Natural Hand Digit Dataset* menunjukkan bahwa akurasi klasifikasi yang dicapai masih terbatas pada kisaran 0,66 (Ahmad et al., 2025). Rendahnya akurasi tersebut menunjukkan bahwa model masih memiliki keterbatasan dalam mengenali gestur angka secara konsisten pada dataset yang terkontekstualisasi secara budaya, khususnya ketika dihadapkan pada variasi kondisi lingkungan pengambilan citra (Ahmad et al., 2025). Tantangan utama dalam akurasi klasifikasi sering kali bersumber dari variabel pengganggu (*gesture-irrelevant factors*) seperti latar belakang yang kompleks (*cluttered background*), variasi posisi tangan, orientasi, perubahan skala akibat jarak, hingga fluktuasi pencahayaan. Fenomena ini menunjukkan bahwa interferensi elemen visual di sekitar tangan dapat mengaburkan fitur-fitur halus jari, sehingga diperlukan pendekatan yang mampu memfokuskan proses klasifikasi secara eksklusif pada wilayah tangan sebagai objek utama. Sebagai solusi atas permasalahan tersebut, penelitian ini mengusulkan penerapan lokalisasi tangan menggunakan *pre-trained model* dari *Hand Gesture Recognition Image Dataset* (HaGRID) yang selanjutnya disebut “*cropped*” atau “*crop*” sebagai tahap pra-pemrosesan sebelum klasifikasi akhir. HaGRID memiliki keunggulan berupa anotasi *bounding box* yang presisi pada ribuan sampel gambar dalam kondisi lingkungan yang beragam (Kapitanov et al., 2024; Nuzhdin et al., 2024).

Dengan mengintegrasikan model lokalisasi ini, sistem dapat mendeteksi dan memotong wilayah tangan secara akurat, sehingga input klasifikasi menjadi lebih terfokus pada gestur jari dan meminimalkan pengaruh *noise* visual dari latar belakang. Kebaruan (*novelty*) penelitian ini terletak pada evaluasi penerapan strategi lokalisasi berbasis HaGRID sebagai tahap isolasi objek guna meningkatkan ketahanan dan akurasi sistem pengenalan gestur angka dalam mendukung media pembelajaran matematika interaktif.

Sebagai pelengkap untuk meningkatkan transparansi dan kepercayaan terhadap model, penelitian ini juga mengintegrasikan metode Grad-CAM (*Gradient-weighted Class Activation Mapping*). Grad-CAM menghasilkan peta aktivasi (*heatmap*) kelas yang menyoroti wilayah-wilayah penting dalam gambar yang paling berkontribusi terhadap keputusan klasifikasi model (Selvaraju et al., 2019). Dengan visualisasi ini, peneliti dan pengguna dapat memahami apakah model benar-benar fokus pada konfigurasi jari yang relevan atau justru terpengaruh oleh artefak latar belakang. Penerapan Grad-CAM dalam konteks pengenalan hitungan jari (1–5) berbasis konteks budaya ini bertujuan untuk memberikan kemampuan *explainable AI*, sehingga sistem pembelajaran interaktif yang dikembangkan tidak hanya akurat, tetapi juga dapat dijelaskan dan dipercaya oleh pendidik maupun peserta didik.

Berdasarkan paparan permasalahan di atas, pertanyaan penelitian yang diajukan dalam studi ini adalah:

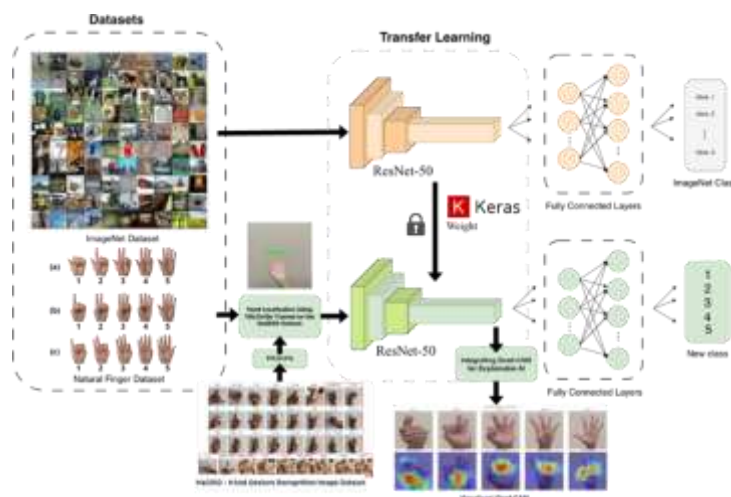
- (1) Bagaimana pengaruh lokalisasi tangan berbasis HaGRID sebelum klasifikasi terhadap performa pengenalan gestur angka berbasis *finger counting* dalam pembelajaran matematika interaktif?

- (2) Bagaimana Grad-CAM dapat memvisualisasikan wilayah keputusan model untuk mendukung *explainable* AI pada sistem pengenalan gestur angka tersebut?

Sejalan dengan pertanyaan tersebut, tujuan dari penelitian ini adalah untuk mengevaluasi performa sistem pengenalan gestur angka berbasis *computer vision* melalui penerapan lokalisasi tangan sebelum klasifikasi, serta mengimplementasikan Grad-CAM guna memberikan kemampuan *explainability* yang menunjukkan fokus perhatian model dalam mengenali setiap gestur jari. Melalui pendekatan ini, diharapkan dapat dihasilkan sistem yang lebih akurat, *robust*, dapat dijelaskan (*explainable*), dan relevan dengan kebutuhan pembelajaran matematika dasar, sehingga mampu memberikan respons yang konsisten dan transparan dalam berbagai skenario penggunaan di lingkungan instruksional yang nyata.

B. Metode Penelitian

Metode yang digunakan dalam penelitian ini terdiri atas tahap lokalisasi tangan (*cropped image*) dan klasifikasi jumlah jari. Citra gestur tangan terlebih dahulu diproses menggunakan *pre-trained model* YOLOv10x untuk mendeteksi area tangan. Model YOLOv10x yang digunakan merupakan model deteksi tangan yang dilatih pada dataset HaGRID. Hasil deteksi berupa *bounding box* kemudian digunakan untuk memotong area tangan dari citra asli. Tahap ini bertujuan untuk mengurangi informasi latar belakang yang tidak relevan dan memfokuskan proses klasifikasi pada objek utama, yaitu tangan dan jari.

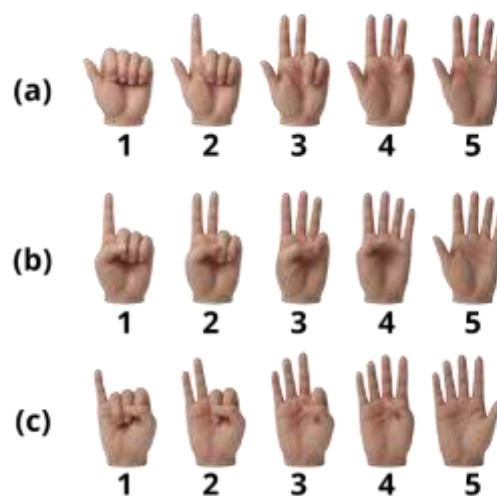


Gambar 1. Arsitektur model *transfer learning* berbasis YOLOv10x dan ResNet50 beserta visualisasi *heatmap* Grad-CAM.

Setelah tahap lokalisasi, citra tangan hasil *crop* digunakan sebagai input pada model klasifikasi berbasis *transfer learning*. Arsitektur yang digunakan adalah ResNet50 dengan bobot pra-latih dari ImageNet. ResNet50 berfungsi sebagai ekstraktor fitur visual, sedangkan lapisan klasifikasi baru ditambahkan untuk menyesuaikan model terhadap lima kelas jumlah jari. Proses pelatihan dilakukan dengan membekukan lapisan dasar pada tahap awal, kemudian dilakukan penyesuaian pada sebagian lapisan akhir melalui *fine-tuning*. Output model berupa probabilitas kelas, dan kelas dengan probabilitas tertinggi dipilih sebagai prediksi jumlah jari. Dengan alur ini, sistem dapat memanfaatkan kemampuan deteksi tangan dari YOLOv10x dan kemampuan ekstraksi fitur visual dari ResNet50 untuk mengklasifikasikan gestur *finger counting* secara lebih efektif.

1. Dataset

Dataset yang digunakan dalam penelitian ini dikoleksi berdasarkan *Natural Hand Digit Dataset* ([Ahmad et al., 2025](#)) yang dikembangkan pada penelitian sebelumnya, yang dirancang untuk merepresentasikan gestur angka "1" sampai "5" dengan mempertimbangkan variasi budaya dalam praktik berhitung menggunakan jari. Secara konseptual, variasi tersebut mengacu pada kajian Bender dan Beller (2012) yang menunjukkan bahwa *finger counting* bukan sekadar aktivitas motorik yang bersifat universal, melainkan sistem representasi numerik yang dikodekan secara kultural (Bender & Beller, 2012). Dalam konteks penelitian ini, variasi budaya dibatasi pada tiga pola dasar *finger counting*, yaitu pola (a), (b), dan (c). Pola (a) merepresentasikan kebiasaan yang banyak ditemukan di Eropa, yaitu memulai perhitungan dengan tangan kiri dan ibu jari. Pola (b) merepresentasikan varian yang umum ditemukan di Cina dan juga luas digunakan di Amerika Serikat, yaitu memulai perhitungan dari jari telunjuk. Sementara itu, pola (c) merepresentasikan varian yang dilaporkan pada konteks Iran, yaitu memulai perhitungan dari tangan kanan dan jari kelingking.

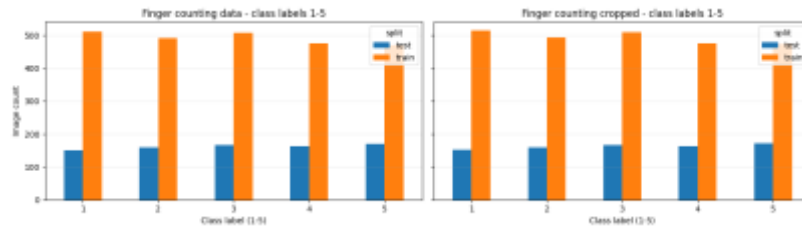


Gambar 2. Contoh sampel citra untuk setiap kelas pada dataset *Natural Hand Digit*

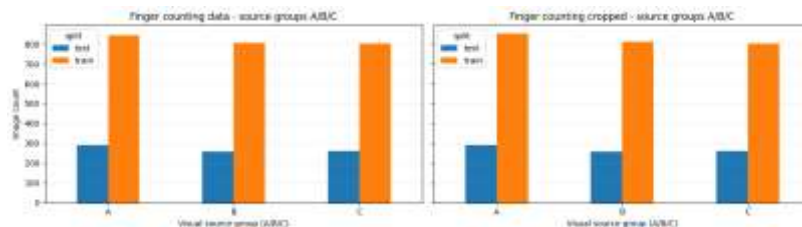
Pemilihan ketiga pola tersebut didasarkan pada pertimbangan bahwa ketiganya masih berada dalam sistem dasar lima jari, memiliki struktur visual yang relatif sebanding, tetapi berbeda secara kultural pada aspek tangan awal dan jari awal yang digunakan. Dengan demikian, ketiga pola tersebut memungkinkan pengujian klasifikasi gestur angka "1" sampai "5" secara lebih terkontrol tanpa mencampurkan sistem *finger counting* lain yang lebih kompleks, seperti sistem bahasa isyarat, sistem berbasis lipatan jari, atau sistem yang melibatkan bagian tubuh lain. Batasan ini juga ditetapkan agar proses akuisisi, pelabelan, dan evaluasi model tetap konsisten dengan fokus penelitian, yaitu pengenalan gestur angka "1" sampai "5" berbasis konfigurasi jari tangan. Proses pengumpulan data melibatkan partisipasi dari lingkungan mahasiswa, dengan akuisisi berupa video yang kemudian diekstraksi menjadi *frame* citra statis. Pengambilan data dilakukan menggunakan smartphone dengan latar belakang yang lebih beragam daripada penelitian sebelumnya ([Ahmad et al., 2025](#)), dan setiap data diberi label secara manual berdasarkan kelas angka "1" sampai "5". Dengan demikian, dataset ini tidak hanya berorientasi pada bentuk visual jari, tetapi juga mempertimbangkan keberagaman representasi budaya dalam *finger counting*.

Secara spesifik, dataset yang digunakan merupakan kumpulan citra gestur tangan untuk pengenalan jumlah jari yang terdiri atas lima kelas, yaitu angka "1", "2", "3", "4", dan "5" yang distribusinya divisualisasikan pada Gambar 3, yang masing-masing merepresentasikan jumlah jari yang ditampilkan pada citra. Dataset ini dibagi menjadi data pelatihan dan data pengujian, dengan total keseluruhan 3.264 citra, yang terdiri dari 2.455 citra sebagai data pelatihan (*training set*) dan

809 citra sebagai data pengujian (*test set*). Setiap kelas juga dikelompokkan ke dalam tiga kelompok visual, yaitu (a), (b), dan (c) yang distribusinya divisualisasikan pada Gambar 4, untuk merepresentasikan variasi tampilan gestur tangan. Distribusi data pada setiap kelas relatif seimbang, sehingga dataset ini sangat sesuai digunakan untuk pengembangan dan evaluasi sistem klasifikasi gestur tangan berbasis citra.



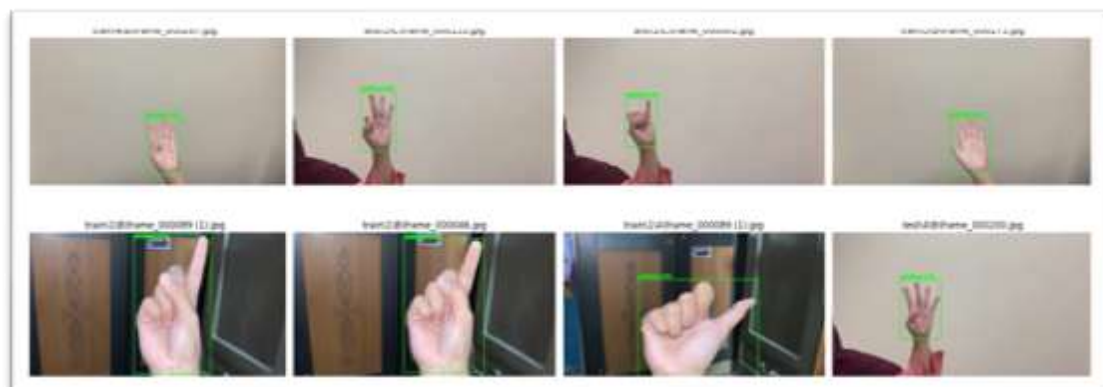
Gambar 3. Distribusi dari *data train* dan *data test* berdasarkan kelas (1-5)



Gambar 4. Distribusi dari *data train* dan *data test* berdasarkan kelas (a,b, dan c)

2. Lokalisasi Tangan Menggunakan YOLOv10x

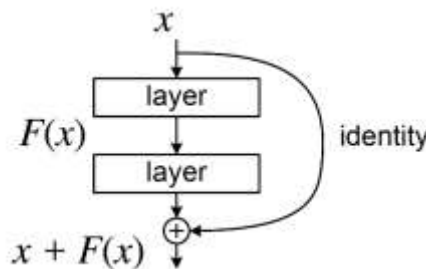
Tahap lokalisasi tangan dilakukan menggunakan *pre-trained model* YOLOv10x *Hands* yang berasal dari model deteksi tangan terlatih pada dataset HaGRID ([Kapitanov et al., 2024](#); [Nuzhdin et al., 2024](#)). YOLOv10x digunakan untuk mendeteksi posisi tangan pada citra dan menghasilkan koordinat *bounding box* sebagai dasar pemotongan area tangan. Pada penelitian ini, YOLOv10x tidak digunakan untuk mengklasifikasikan jumlah jari, melainkan hanya sebagai tahap pra-pemrosesan untuk memfokuskan citra pada region tangan. Proses deteksi dilakukan dengan *confidence threshold* sebesar 0,25 dan ukuran input 640 piksel. Setelah *bounding box* diperoleh, area tangan dipotong dengan tambahan padding sebesar 8% agar bentuk tangan dan jari tetap tercakup secara utuh. Proses ini menghasilkan 3.279 citra tangan hasil *crop* dari 3.264 citra asli, dengan hanya satu citra yang tidak terdeteksi. Penggunaan lokalisasi tangan ini bertujuan untuk mengurangi pengaruh latar belakang dan meningkatkan relevansi fitur visual yang digunakan pada tahap klasifikasi berikutnya.



Gambar 5. Contoh hasil lokalisasi tangan menggunakan YOLOv10x.

3. Transfer Learning dengan ResNet-50

Setelah citra tangan berhasil dilokalisasi dan dipotong, tahap berikutnya adalah mengklasifikasikan jumlah jari menggunakan pendekatan *transfer learning*. Pendekatan ini dipilih karena jumlah dataset relatif terbatas untuk melatih model *deep learning* dari awal, sehingga penelitian ini memanfaatkan model pralatih yang telah belajar dari dataset berskala besar, dalam hal ini ImageNet, dan kemudian menyesuaikannya dengan tugas klasifikasi gestur jari. Model dasar yang digunakan adalah ResNet50, yaitu arsitektur *convolutional neural network* dengan 50 lapisan yang menerapkan konsep *residual connection*, sebagaimana ditunjukkan pada Gambar 6 (He et al., 2015).



Gambar 6. *Residual Connection* (He et al., 2015)

Melalui mekanisme ini, informasi dari lapisan sebelumnya dapat diteruskan secara langsung ke lapisan berikutnya melalui jalur pintas, sehingga model lebih stabil dalam mempelajari representasi visual pada jaringan yang dalam. Mekanisme *residual connection* juga membantu mengurangi kendala *vanishing gradient* yang sering muncul pada pelatihan jaringan berlapis banyak (He et al., 2015). Dalam penelitian ini, ResNet50 digunakan sebagai *feature extractor* untuk mengekstraksi karakteristik visual dari citra tangan, seperti tepi, bentuk, tekstur, pola jari, dan struktur visual tangan. Fitur-fitur tersebut kemudian diproses oleh lapisan klasifikasi baru untuk menentukan kelas jumlah jari. Pemilihan ResNet50 didasarkan pada pertimbangan bahwa ResNet merupakan *backbone CNN* yang mapan, banyak digunakan dalam berbagai tugas *computer vision*, serta sesuai dengan kebutuhan penelitian yang menuntut keseimbangan antara kemampuan ekstraksi fitur, stabilitas pelatihan, dan efisiensi komputasi (Elharrouss et al., 2022).

4. Visualisasi Interpretabilitas Menggunakan Grad-CAM

Guna menganalisis area citra yang berkontribusi signifikan terhadap keputusan model, penelitian ini mengadopsi metode Gradient-weighted Class Activation Mapping (Grad-CAM) sebagai teknik Explainable AI (XAI) untuk meningkatkan transparansi dan interpretabilitas visual (Balat et al., 2025). Grad-CAM digunakan sebagai metode diagnostik untuk memvalidasi apakah model secara konsisten berfokus pada fitur-fitur morfologi yang relevan, seperti bentuk tangan dan posisi jari, sekaligus untuk mengidentifikasi potensi bias selama proses klasifikasi jumlah jari (Selvaraju et al., 2019). Pada penelitian ini, Grad-CAM diimplementasikan pada arsitektur ResNet50 yang telah dilatih dengan memanfaatkan aktivasi dari lapisan konvolusi terakhir. Penggunaan lapisan terdalam ini didasarkan pada fakta bahwa lapisan tersebut merupakan titik temu paling optimal dalam menangkap konsep semantik tingkat tinggi (seperti ruas jari) sambil tetap mempertahankan informasi spasial yang memadai (Selvaraju et al., 2019).

Secara teknis, proses ini dilakukan dengan menghitung gradien dari skor kelas prediksi terhadap feature map pada lapisan konvolusi tersebut, yang kemudian diproses melalui global average pooling untuk menghasilkan bobot kepentingan neuron. Hasil kombinasi linier terbobot ini dilewatkan melalui fungsi aktivasi ReLU untuk menghasilkan heatmap yang menunjukkan wilayah dengan kontribusi terbesar terhadap keputusan model (Balat et al., 2025; Selvaraju et al., 2019). Heatmap tersebut selanjutnya ditumpangkan (superimposed) pada citra input asli untuk dianalisis secara visual,

sehingga dapat dipastikan bahwa pengambilan keputusan model didasarkan pada fitur tangan yang akurat dan bukan pada gangguan latar belakang.

5. Metode Evaluasi

Evaluasi model dilakukan menggunakan data uji untuk mengukur kemampuan model dalam mengklasifikasikan citra gestur tangan ke dalam lima kelas jumlah jari. Model menghasilkan probabilitas untuk setiap kelas, kemudian kelas dengan probabilitas tertinggi dipilih sebagai prediksi akhir. Metrik evaluasi yang digunakan meliputi confusion matrix (Tabel 1), precision (Persamaan 1), recall (Persamaan 2), accuracy (Persamaan 3), F1-score (Persamaan 4), dan categorical cross-entropy loss (Persamaan 5). Accuracy digunakan untuk mengukur proporsi prediksi benar terhadap seluruh data uji, sedangkan confusion matrix digunakan untuk menganalisis distribusi kesalahan klasifikasi antarkelas. Precision mengukur ketepatan prediksi pada suatu kelas, recall mengukur kemampuan model mengenali seluruh data pada kelas tersebut, dan F1-score digunakan untuk menilai keseimbangan antara precision dan recall. Selain metrik per kelas, digunakan pula macro average (Persamaan 6-8) dan weighted average (Persamaan 9-11) untuk memperoleh ringkasan performa model secara keseluruhan.

TABLE 1. Tabel confusion matrix
Kelas Prediksi

Kelas Aktual	Classes	a	b	c
	a	TP_a	$F_{b a}$	$F_{c a}$
	b	$F_{a b}$	TP_b	$F_{c b}$
	c	$F_{a c}$	$F_{b c}$	TP_c

Sumber: (Ahmad et al., 2023)

$$Precision_a = \frac{TP_a}{TP_a + \sum_X F_{a|X}} \quad (1)$$

$$Recall_a = \frac{TP_a}{TP_a + \sum_X F_{X|a}} \quad (2)$$

$$Accuracy = \frac{\sum_a TP_a}{\sum_a TP_a + \sum_a F_{X|a} + \sum_a F_{a|X}} = \frac{\sum_a TP_a}{N} \quad (3)$$

$$F1 - Score = 2 \cdot \frac{precision \cdot recall}{precision + recall} \quad (4)$$

$$Categorical CE Loss = -\log \left(\frac{e^{s_i}}{\sum_{j=1}^C e^{s_j}} \right) \quad (5)$$

$$Macro Precision = \frac{1}{C} \sum_{a=1}^C Precision_a \quad (6)$$

$$Macro Recall = \frac{1}{C} \sum_{a=1}^C Recall_a \quad (7)$$

$$Macro F1 = \frac{1}{C} \sum_{a=1}^C F1_a \quad (8)$$

$$\text{Weighted Precision} = \sum_{a=1}^c \frac{n_a}{N} \times \text{Precision}_a \quad (9)$$

$$\text{Weighted Recall} = \sum_{a=1}^c \frac{n_a}{N} \times \text{Recall}_a \quad (10)$$

$$\text{Weighted F1} = \sum_{a=1}^c \frac{n_a}{N} \times \text{F1}_a \quad (11)$$

C. Hasil Penelitian dan Pembahasan

Pada tahap awal, model dilatih menggunakan mekanisme *callback*, yaitu *EarlyStopping* dan *ReduceLROnPlateau*. *Callback* digunakan untuk membantu proses pelatihan agar lebih stabil. *EarlyStopping* menghentikan pelatihan ketika performa validasi tidak lagi meningkat, sedangkan *ReduceLROnPlateau* menurunkan nilai *learning rate* ketika nilai *validation loss* tidak mengalami perbaikan. Dengan strategi ini, model dapat menghindari pelatihan berlebihan (*overfitting*) dan mempertahankan bobot terbaik selama proses pelatihan. Hasil pelatihan menggunakan *callback* pada Tabel 2 menunjukkan bahwa dataset hasil lokalisasi tangan atau *Finger counting cropped* memberikan performa lebih baik dibandingkan dengan dataset asli atau *Finger counting data*.

Tabel 2. Performa Model dari Pelatihan dengan *Callback*

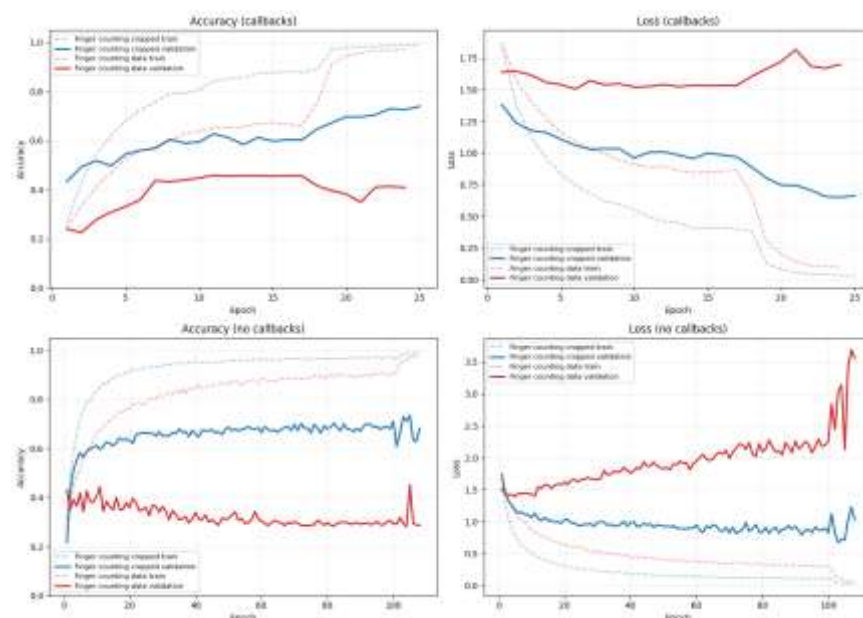
<i>Metrics</i>	<i>Finger counting cropped</i>	<i>Finger counting data</i>
<i>Accuracy</i>	0.739827	0.416564
<i>Macro Avg Precision</i>	0.772152	0.614896
<i>Macro Avg Recall</i>	0.739626	0.405538
<i>Macro Avg F1-Score</i>	0.724359	0.351132
<i>Weighted Avg Precision</i>	0.77	0.62
<i>Weighted Avg Recall</i>	0.739827	0.416564
<i>Weighted Avg F1-Score</i>	0.723508	0.359464

Berdasarkan hasil tersebut, model yang dilatih menggunakan dataset *cropped* memperoleh akurasi sebesar 73,98%, sedangkan model yang menggunakan dataset asli hanya memperoleh akurasi sebesar 41,66%. Nilai *macro F1-score* pada dataset *cropped* juga lebih tinggi, yaitu 72,44%, dibandingkan dengan dataset asli yang hanya mencapai 35,11%. Hal ini menunjukkan bahwa lokalisasi tangan memberikan kontribusi penting dalam meningkatkan performa klasifikasi. Dengan menggunakan citra hasil *crop*, model lebih fokus pada area tangan dan jari, sehingga informasi latar belakang yang tidak relevan dapat dikurangi. Selanjutnya, pelatihan juga dilakukan tanpa menggunakan *callback*. Tujuan eksperimen ini adalah untuk melihat sejauh mana model dapat mengalami konvergensi ketika proses pelatihan berjalan tanpa penghentian otomatis atau penyesuaian *learning rate* secara adaptif. Dengan kata lain, eksperimen tanpa *callback* digunakan untuk mengamati perilaku alami model selama pelatihan penuh.

TABEL 3. Performa Model dari Pelatihan tanpa *Callback*

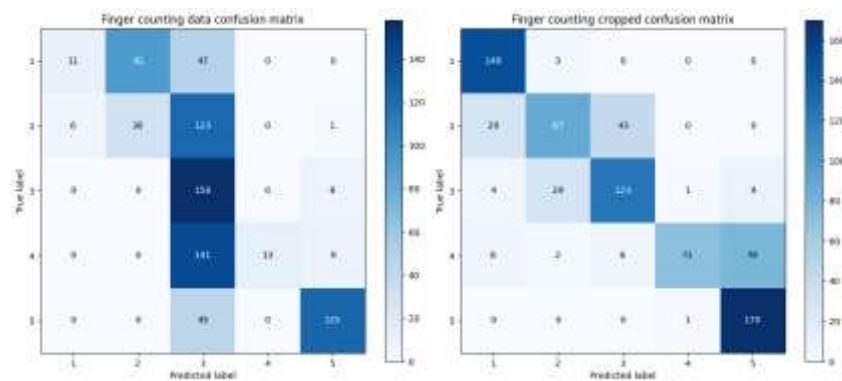
<i>Metric</i>	<i>Finger counting cropped</i>	<i>Finger counting data</i>
<i>Accuracy</i>	0.681874	0.28801
<i>Macro Avg Precision</i>	0.7366	0.466565
<i>Macro Avg Recall</i>	0.678129	0.284059
<i>Macro Avg F1-Score</i>	0.650966	0.171637
<i>Weighted Avg Precision</i>	0.73	0.45
<i>Weighted Avg Recall</i>	0.681874	0.28801
<i>Weighted Avg F1-Score</i>	0.650062	0.172397

Pada skenario tanpa *callback*, performa model mengalami penurunan. Dataset *cropped* memperoleh akurasi sebesar 68,19%, lebih rendah dibandingkan pelatihan dengan *callback* yang mencapai 73,98%. Penurunan juga terlihat pada *macro F1-score*, dari 72,44% menjadi 65,10%. Pada dataset asli, penurunan performa lebih besar, yaitu dari akurasi 41,66% menjadi 28,80%. Hasil ini menunjukkan bahwa *callback* berperan penting dalam menjaga stabilitas pelatihan. Tanpa *callback*, model tetap dapat belajar, terutama pada dataset *cropped*, tetapi performanya tidak sebaik ketika *callback* digunakan. Hal ini mengindikasikan bahwa proses pelatihan tanpa *callback* cenderung kurang optimal, kemungkinan karena model terus dilatih meskipun performa validasi tidak lagi membaik. Kondisi tersebut dapat menyebabkan model mulai menyesuaikan diri secara berlebihan terhadap data pelatihan dan kehilangan kemampuan generalisasi pada data uji. Secara keseluruhan, hasil eksperimen memperlihatkan dua temuan utama. Pertama, dataset hasil crop memberikan performa yang lebih baik dibandingkan dataset asli, baik dengan *callback* maupun tanpa *callback*. Kedua, penggunaan *callback* membantu model mencapai performa yang lebih stabil dan lebih tinggi. Oleh karena itu, kombinasi antara lokalisasi tangan dan strategi pelatihan menggunakan *callback* merupakan pendekatan yang lebih efektif untuk klasifikasi jumlah jari pada penelitian ini.



Gambar 7. Kurva Pelatihan Model

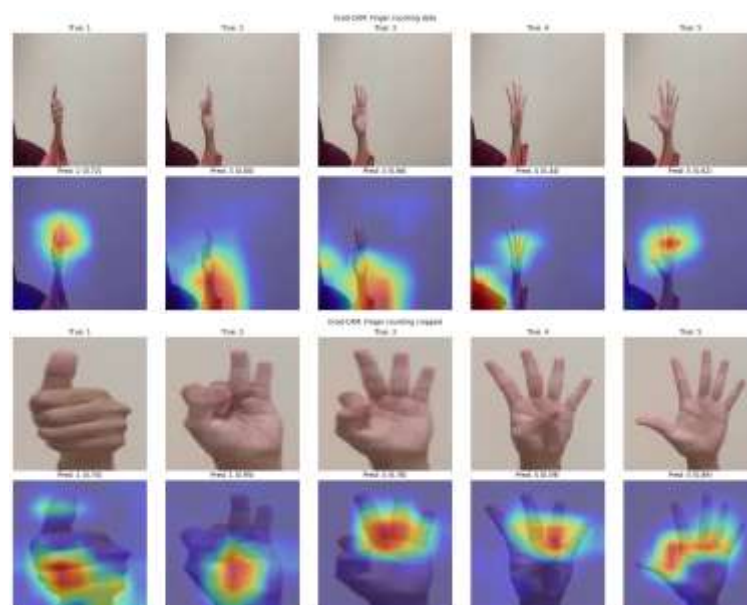
Gambar 7 memperlihatkan kurva pelatihan yang menunjukkan perubahan nilai akurasi dan *loss* selama proses *training* dan validasi pada dua skenario, yaitu model dengan *callbacks* dan tanpa *callbacks*. Garis putus-putus merepresentasikan performa pada data pelatihan, sedangkan garis penuh menunjukkan performa pada data validasi. Warna biru menunjukkan dataset *Finger counting cropped*, sedangkan warna merah menunjukkan dataset *Finger counting data*. Kurva juga memperlihatkan adanya indikasi *overfitting*, terutama pada dataset asli. Hal ini terlihat dari akurasi pelatihan yang terus meningkat hingga sangat tinggi, tetapi akurasi validasi tetap rendah atau bahkan menurun. Pada saat yang sama, nilai *validation loss* pada dataset asli cenderung meningkat, terutama setelah tahap *fine-tuning*. Kondisi ini menunjukkan bahwa model mampu mempelajari pola pada data pelatihan, tetapi generalisasinya terhadap data validasi masih terbatas.



Gambar 8. *Confusion matrix* yang dihasilkan proses pelatihan model

Berdasarkan *confusion matrix* dari Gambar 8, pada metode pelatihan menggunakan *callback*, model yang dilatih menggunakan dataset hasil crop dari proses lokalisasi tangan dengan YOLOv10x menunjukkan performa klasifikasi yang lebih baik dibandingkan dengan model yang dilatih menggunakan citra asli tanpa *treatment* apa pun. Pada dataset tanpa *crop*, model cenderung mengalami bias prediksi terhadap kelas "3", yang ditunjukkan oleh *recall* kelas "3" yang tinggi tetapi *precision* yang rendah. Hal ini mengindikasikan bahwa banyak citra dari kelas lain salah diklasifikasikan sebagai kelas "3". Selain itu, kelas "1" dan "4" memiliki *recall* yang rendah, sehingga sebagian besar sampel pada kelas tersebut belum dikenali dengan baik.

Sebaliknya, pada dataset cropped, *confusion matrix* menunjukkan peningkatan jumlah prediksi benar, terutama pada kelas "1" dan "5". Kelas "1" memperoleh *recall* 0,98 dan kelas "5" memperoleh *recall* 0,99, yang menunjukkan bahwa model mampu mengenali kedua kelas tersebut dengan sangat baik. Meskipun demikian, kesalahan masih terjadi pada kelas "2", "3", dan "4", terutama karena kemiripan bentuk gestur pada jumlah jari menengah. Hasil ini menunjukkan bahwa lokalisasi tangan melalui proses *cropping* dapat mengurangi pengaruh latar belakang dan membantu model lebih fokus pada fitur visual yang relevan, yaitu bentuk tangan dan jumlah jari yang terbuka.



Gambar 9. Hasil prediksi *data test* dan visualisasi *heatmap* dari Grad-CAM

Gambar 9 menampilkan hasil visualisasi Grad-CAM yang digunakan untuk menganalisis fokus perhatian model terhadap citra input. Pada dataset hasil *crop*, area aktivasi Grad-CAM cenderung terpusat pada bagian tangan dan jari, khususnya pada region yang membedakan jumlah jari terbuka. Hal ini menunjukkan bahwa model memanfaatkan fitur visual yang relevan dalam proses klasifikasi. Sebaliknya, pada citra asli, aktivasi Grad-CAM dapat menyebar ke area latar belakang atau bagian non-tangan terutama pada kelas "2", "3", dan "4", sehingga informasi yang dipelajari model menjadi kurang spesifik. Temuan ini mendukung hasil kuantitatif bahwa dataset *cropped* menghasilkan performa klasifikasi yang lebih tinggi dibandingkan dengan dataset asli.

D. Kesimpulan

Berdasarkan hasil penelitian, lokalisasi tangan berbasis HaGRID yang diimplementasikan melalui YOLOv10x sebelum tahap klasifikasi terbukti memberikan pengaruh positif terhadap performa pengenalan gestur angka berbasis finger counting. Tahap lokalisasi mampu mengisolasi area tangan dari latar belakang, sehingga citra masukan model ResNet50 menjadi lebih terfokus pada bentuk tangan, jumlah jari yang terbuka, dan konfigurasi visual jari yang relevan. Dibandingkan dengan penggunaan citra asli, dataset hasil cropping menghasilkan performa yang lebih stabil dan lebih baik, baik pada skenario pelatihan dengan callback maupun tanpa callback. Dengan demikian, lokalisasi tangan bukan hanya berperan sebagai pra-pemrosesan teknis, tetapi sebagai komponen penting yang meningkatkan kualitas fitur visual yang dipelajari model. Temuan ini menunjukkan bahwa sistem pengenalan gestur angka untuk pembelajaran matematika interaktif lebih layak dikembangkan dengan alur deteksi tangan terlebih dahulu sebelum melakukan klasifikasi angka. Dalam konteks pembelajaran, pendekatan ini penting karena citra yang diperoleh dari kamera pengguna dapat memiliki latar belakang, pencahayaan, jarak, dan orientasi tangan yang beragam. Dengan membatasi perhatian model pada wilayah tangan, sistem menjadi lebih robust dan lebih relevan untuk digunakan sebagai media interaktif yang membutuhkan respons pengenalan gestur secara konsisten.

Grad-CAM dalam penelitian ini berhasil memberikan dukungan explainable AI dengan memvisualisasikan wilayah citra yang berkontribusi terhadap keputusan model. Pada citra hasil lokalisasi, aktivasi Grad-CAM cenderung terpusat pada area tangan dan jari, terutama bagian yang membedakan jumlah jari terbuka. Hal ini menunjukkan bahwa model membuat keputusan berdasarkan fitur visual yang sesuai dengan tujuan klasifikasi. Sebaliknya, pada citra asli, aktivasi dapat menyebar ke area non-tangan atau latar belakang, yang menjelaskan mengapa performa model pada dataset tanpa lokalisasi lebih rendah. Secara keseluruhan, penelitian ini menyimpulkan bahwa kombinasi lokalisasi tangan berbasis HaGRID, transfer learning ResNet50, dan visualisasi Grad-CAM membentuk pendekatan yang kuat untuk pengenalan gestur angka dalam pembelajaran matematika interaktif. Lokalisasi meningkatkan performa dan ketahanan model, sedangkan Grad-CAM memperkuat transparansi keputusan model dengan menunjukkan bahwa sistem benar-benar memanfaatkan area tangan dan jari sebagai dasar pengenalan. Meskipun masih terdapat tantangan pada gestur dengan bentuk jari yang saling mirip, hasil penelitian ini menunjukkan bahwa pendekatan yang diusulkan layak menjadi dasar pengembangan sistem pembelajaran matematika berbasis gestur yang lebih akurat, adaptif, dan dapat dijelaskan.

Daftar Pustaka

Ahmad, K. A., Nurriqhi, A., Fettyana, F., & Fadilah, N. (2025). *Pembangunan Dataset Gestur Angka (1-5) yang Terkontekstualisasi Budaya sebagai inisiasi Basis Data Pengenalan untuk Alat Bantu Pembelajaran Matematika*.

- Ahmad, K. A., Silpani, D. C., & Yoshida, K. (2023). The Impact of Large Sample Datasets on Hand Gesture Recognition by Hand Landmark Classification. *International Journal of Affective Engineering*, 22(3), 253–259. <https://doi.org/10.5057/ijae.ijae-d-22-00022>
- Balat, M., Awaad, R., Zaky, A. B., & Aly, S. A. (2025). *Revolutionizing Communication with Deep Learning and XAI for Enhanced Arabic Sign Language Recognition*. <http://arxiv.org/abs/2501.08169>
- Bender, A., & Beller, S. (2012). Nature and culture of finger counting: Diversity and representational effects of an embodied cognitive tool. *Cognition*, 124(2), 156–182. <https://doi.org/10.1016/j.cognition.2012.05.005>
- di Luca, S., & Mauro, P. (2011). Finger numeral representations: More than just another symbolic code. In *Frontiers in Psychology* (Vol. 2, Number NOV). <https://doi.org/10.3389/fpsyg.2011.00272>
- Elharrouss, O., Akbari, Y., Almaadeed, N., & Al-Maadeed, S. (2022). *Backbones-Review: Feature Extraction Networks for Deep Learning and Deep Reinforcement Learning Approaches*. <https://doi.org/10.1016/j.cosrev.2024.100645>
- Foteinos, K., Linardakis, M., Radoglou-Grammatikis, P., Argyriou, V., Sarigiannidis, P., Varlamis, I., & Papadopoulos, G. Th. (2026). *Visual Hand Gesture Recognition with Deep Learning: A Comprehensive Review of Methods, Datasets, Challenges and Future Research Directions*. <http://arxiv.org/abs/2507.04465>
- He, K., Zhang, X., Ren, S., & Sun, J. (2015). *Deep Residual Learning for Image Recognition*. <http://arxiv.org/abs/1512.03385>
- Kapitanov, A., Kvanchiani, K., Nagaev, A., Kraynov, R., & Makhliarchuk, A. (2024). *HaGRID - HAnd Gesture Recognition Image Dataset*. <https://doi.org/10.1109/WACV57701.2024.00451>
- Nuzhdin, A., Nagaev, A., Sautin, A., Kapitanov, A., & Kvanchiani, K. (2024). *HaGRIDv2: 1M Images for Static and Dynamic Hand Gesture Recognition*. <http://arxiv.org/abs/2412.01508>
- Pika, S., Nicoladis, E., & Marentette, P. (2009). How to order a beer: Cultural differences in the use of conventional gestures for numbers. *Journal of Cross-Cultural Psychology*, 40(1), 70–80. <https://doi.org/10.1177/0022022108326197>
- Selvaraju, R. R., Cogswell, M., Das, A., Vedantam, R., Parikh, D., & Batra, D. (2019). *Grad-CAM: Visual Explanations from Deep Networks via Gradient-based Localization*. <https://doi.org/10.1007/s11263-019-01228-7>