Journal of Computer and Information Systems Ampera
Vol. 6, No. 3, September 2025 e-ISSN: 2775-2496
https://journal-computing.org/index.php/journal-cisa/index

Klasifikasi Aksi Peserta Didik di Dalam Ruang Kelas Menggunakan Video Mask Auto Encoder

Meutia Jasmine Annisa Herawan[1], Rani Megasari[2], Yaya WIhardi[3]

[1,2,3]Program Studi Ilmu Komputer, Universitas Pendidikan Indonesia, Kota Bandung, Indonesia
Email: [1]meutia.ja@upi.edu, [2]megasari@upi.edu, [3]yayawihardi@upi.edu

Abstract

Dalam dunia pendidikan, evaluasi terhadap peserta didik di ruang kelas adalah komponen penting yang perlu dilakukan secara sistematis dan terencana sebagai indikator keberhasilan pembelajaran. Penelitian ini mengusulkan pendekatan inovatif dalam mengevaluasi peserta didik melalui klasifikasi aksi di dalam kelas menggunakan model VideoMAE. Dataset yang digunakan terdiri dari 403 video dan disusun langsung oleh penulis karena belum tersedia dataset standar yang sesuai untuk tugas ini saat penelitian dilakukan. Metode yang digunakan adalah fine-tuning model VideoMAE dan TimeSformer, di mana TimeSformer berperan sebagai model pembanding. Hasil evaluasi menunjukkan bahwa VideoMAE mencapai akurasi 68,08%, F1-score 65,58%, recall 68,08%, dan precision 69,98% pada data uji. Pendekatan ini menunjukkan potensi dalam memahami perilaku peserta didik, serta dapat mendukung pengembangan strategi pembelajaran dan sistem evaluasi kelas yang lebih efektif.

Keywords: Aksi Peserta Didik, Klasifikasi Aksi, Pengenalan Aksi Manusia (Human Action Recognition), Time-Series Transformers (TimeSformer), Video Masked Auto Encoder (VideoMAE)

1. PENDAHULUAN

Aksi manusia adalah gerakan yang dilakukan secara sengaja, dipengaruhi oleh kondisi mental internal, dan sering kali mencerminkan keterlibatan, perhatian, atau niat dalam konteks tertentu, termasuk dalam bidang pendidikan [5]. Di dalam kelas, perilaku peserta didik dapat menjadi indikator yang berharga untuk mengevaluasi keterlibatan dan hasil belajar [1]. Oleh karena itu, mengamati aksi peserta didik dianggap sebagai langkah penting untuk memahami perilaku belajar secara individu.

Secara tradisional, evaluasi aksi di dalam kelas dilakukan secara manual oleh pengamat atau guru, yang mengandalkan penilaian subjektif manusia dan membutuhkan waktu serta tenaga yang cukup besar [6]. Metode ini menjadi kurang praktis terutama dalam situasi kelas yang besar. Untuk mengatasi keterbatasan tersebut, pengenalan aksi manusia (Human Action Recognition/HAR) secara otomatis berbasis video dan deep learning muncul sebagai solusi yang menjanjikan.

Deep learning telah menjadi pendekatan mutakhir dalam berbagai tugas penglihatan komputer (computer vision), termasuk klasifikasi citra, deteksi objek, estimasi pose, dan khususnya pengenalan aksi manusia [11]. HAR bertujuan untuk mengenali dan mengklasifikasikan rangkaian gerakan manusia dalam video [8]. Islam dan Iqbal menggunakan mekanisme perhatian (attention mechanism) dan arsitektur LSTM untuk mengklasifikasikan aksi dalam dataset UTD-MHAD dan UT-Kinect dengan akurasi tinggi [7]. Popescu dan Mocanu [9] menerapkan 3D CNN pada empat dataset benchmark, dengan akurasi hingga 98,43% pada MSRDailyActivity3D dan 94,38% pada dataset kustom mereka. Meskipun sukses, pendekatan-pendekatan ini umumnya bergantung pada dataset berlabel besar dan kesulitan diterapkan dalam domain dunia nyata yang lebih terbatas seperti lingkungan kelas.

Arsitektur berbasis transformer seperti Vision Transformer (ViT) baru-baru ini menunjukkan performa unggul dalam tugas pengenalan citra dan video [4]. TimeSformer, yang diperkenalkan oleh Bertasius et al. [2], mengembangkan ViT untuk analisis video dengan memodelkan perhatian spasial-temporal, dan terbukti kompetitif pada benchmark seperti Kinetics-400. Lebih baru lagi, VideoMAE (Video Masked Autoencoder) yang diajukan oleh Tong et al. [10] memanfaatkan pembelajaran mandiri (self-supervised learning) melalui prediksi frame yang dimasking, memungkinkan pemahaman spasial-temporal yang lebih baik dengan jumlah label yang lebih sedikit dan biaya pelatihan yang lebih rendah.

Meskipun TimeSformer dan VideoMAE telah menunjukkan performa yang kuat pada benchmark pengenalan aksi publik, penerapannya dalam konteks pendidikan masih terbatas. Selain itu, penelitian sebelumnya cenderung berfokus pada dataset berskala besar dan aksi manusia secara umum (seperti olahraga, gestur), yang mungkin tidak mencerminkan nuansa perilaku di dalam kelas.

Untuk menjawab kesenjangan ini, penelitian ini mengusulkan sistem klasifikasi aksi menggunakan model VideoMAE dan TimeSformer pada dataset video kelas skala kecil (403 video). Kedua model di-fine-tune agar sesuai dengan domain pendidikan dan diuji kemampuannya dalam mengklasifikasikan aksi peserta didik secara otomatis. Pendekatan ini memberikan kontribusi pada literatur dengan menunjukkan kelayakan model berbasis transformer dalam skenario pendidikan dengan data terbatas, serta meningkatkan evaluasi kelas dengan solusi yang lebih objektif dan skalabel.

2. METODE

2.1. Data

Dataset yang digunakan dalam penelitian ini terdiri dari video aksi peserta didik di dalam kelas, yang direkam secara langsung dalam dua sesi pengambilan data di Gedung B Fakultas Pendidikan Matematika dan Ilmu Pengetahuan Alam, Universitas Pendidikan Indonesia. Aksi yang diklasifikasikan mencakup lima kategori: mengangguk, mengangkat tangan, menggunakan ponsel, menopang kepala, dan menunduk. Proses perekaman dilakukan dengan tiga perangkat kamera dari sudut berbeda (CCTV, Handycam, dan ponsel), guna memperoleh variasi perspektif ruang kelas.

2.2.1. Praproses Data

Sebelum digunakan dalam pelatihan model, video mentah melalui tahap pra-pemrosesan berupa:

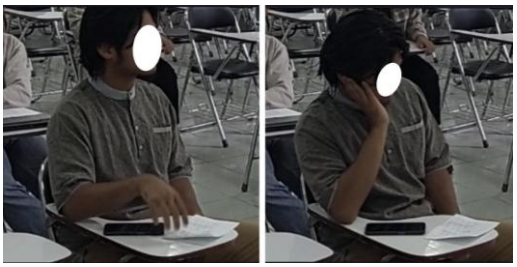
1. Ekstraksi klip video: dilakukan dengan memotong video hanya pada bagian yang mengandung satu aksi spesifik, kemudian melakukan cropping untuk memfokuskan pada individu pelaku aksi.
2. Pelabelan: setiap klip diberi label kelas sesuai aksi yang direkam, disimpan dalam folder kelas masing-masing, dengan format penamaan standar "[kode kelas]_[nomor klip].mp4" guna menjaga konsistensi struktur data.

Contoh frame video dan jumlah data yang berhasil melewati tahap preproses terlampir pada Tabel 1.

Tabel 1. Data setelah Praproses

Label Kelas	Contoh Sequence Frame dari Klip	Jumlah

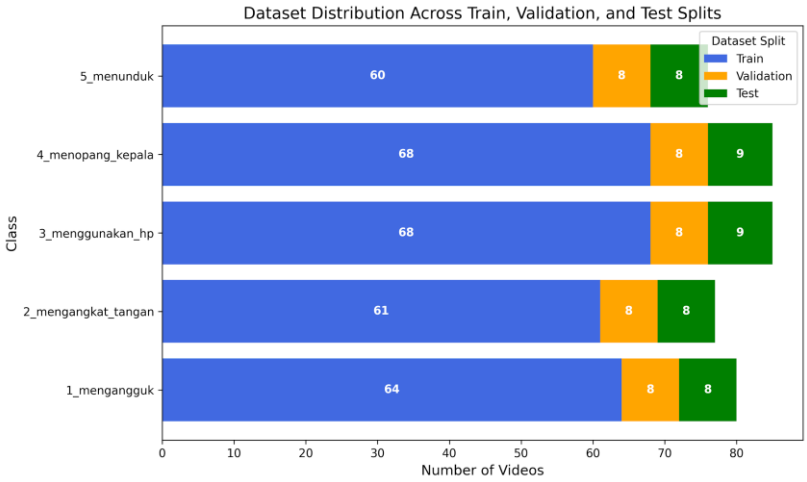
Vol. 3, No. 3, September 2022 e-ISSN: 2775-2496
https://journal-computing.org/index.php/journal-cisa/index

		Klip
1_menganggguk		80
2_mengangkat_tangan		77
3_menggunakan_hp		85
4_menopang_kepala		85

5_menunduk		76
TOTAL		403

2.2.1. Pembentukan Dataset

Dataset akhir terdiri dari 403 klip video, yang kemudian dibagi menjadi tiga subset: training (80%), validation (10%), dan testing (10%) secara acak namun proporsional antar kelas. Pembagian ini dilakukan untuk memastikan model dapat dilatih dan dievaluasi secara objektif, dengan distribusi yang merata di setiap subset. Distribusi pembagian dataset per label aksi divisualisasikan dengan diagram batang pada Gambar ….



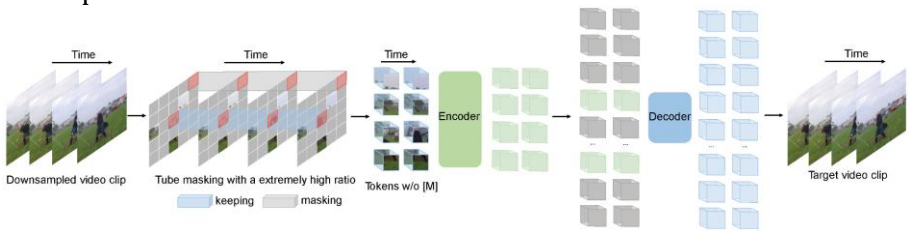
Gambar 1. Distribusi Dataset

2.2. Model

Penelitian ini menggunakan dua model pralatih berbasis Vision Transformer (ViT) yang dikembangkan untuk memproses data video.

2.3.1. VideoMAE

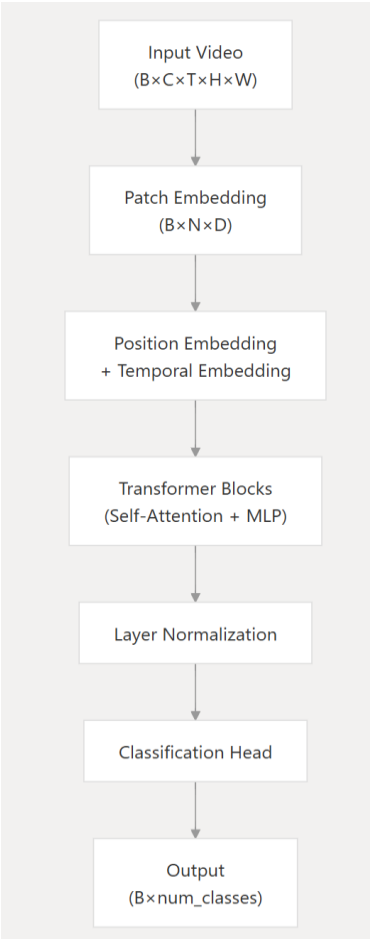
Arsitektur VideoMAE terdiri dari encoder-decoder asimetris). *Encoder* menerima sebagian kecil *patch* video yang tidak di*masking*, sedangkan *decoder* yang lebih ringan bertugas merekonstruksi seluruh video dari *token* yang terlihat dan *token mask* yang dapat dipelajari. *Patch* video di*masking* dalam bentuk *tube* (spasial-temporal) untuk mendorong pemodelan dinamika gerak secara efektif. Arsitektur VideoMAE dapat dilihat pada Gambar 2.



Gambar 2. Arsitektur VideoMAE

2.3.2. TimeSformer

TimeSformer (Time-Space Transformer) adalah arsitektur video berbasis Vision Transformer (ViT) yang menggantikan sepenuhnya konvolusi dengan self-attention untuk memahami informasi spatio-temporal dalam video. Setiap frame video dipecah menjadi patch 16×16 piksel, yang kemudian diproyeksikan menjadi token embedding. Token ini diperkaya dengan posisi dan waktu, lalu diproses oleh encoder transformer. Untuk efisiensi, TimeSformer menggunakan *divided attention*, yaitu memisahkan proses *attention* antara dimensi spasial (dalam satu frame) dan temporal (antar frame). Arsitektur TimeSformer dapat dilihat pada Gambar 3.



Gambar 3. Arsitektur TimeSformer

2.3.3. Konfigurasi Model

Penelitian ini menggunakan dua model pralatih yang diimplementasikan pada dataset Kinetics 400 *Human Action Recognition* [12]. Model pralatih dapat diakses publik pada *platform* HuggingFace. Perbandingan Konfigurasi Model dirangkum pada Tabel 2.

Tabel 2. Perbandingan Konfigurasi Model

Parameter	VideoMAE	TimeSformer
image_size	224	224
initializer_range	0.02	0.02
intermediate_size	3072	3072
num_attention_heads	12	12
num_channels	3	3
num_frames	16	8
num_hidden_layers	12	12
patch_size	16	16
tubelet_size	2	2
hidden_act	gelu	gelu
decoder_hidden_size	384	-
decoder_intermediate_size	1536	-
decoder_num_attention_heads	6	-
decoder_num_hidden_layers	4	-
use_mean_pooling	True	-
Trainable Parameters	**94.2M**	**121M**

2.3. Hyperparameter Fine-tune

Fine-tuning adalah proses mengatur hyperparameter sebelum pelatihan untuk mengoptimalkan performa model. Tabel 3 menunjukkan kombinasi hyperparameter yang digunakan dalam pelatihan.

Tabel 3. Kombinasi Hyperparameter

No.	Model	Epoch	Batch Size	Learning Rate
1	VideoMAE	50	8	1e-3
2	VideoMAE	20	4	5e-5
3	TimeSformer	15	4	5e-3
4	TimeSformer	20	4	5e-5

2.4. Matriks Evaluasi

Accuracy, Precision, Recall, dan F1-Score merupakan metrik evaluasi performa model yang digunakan dalam penelitian ini [3]. Accuracy mengukur proporsi prediksi yang benar dari total seluruh prediksi yang dilakukan.

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

Presisi menunjukkan seberapa banyak prediksi positif yang benar dibandingkan dengan seluruh prediksi positif yang dibuat.

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

Recall mengukur seberapa banyak data positif yang berhasil dikenali dengan benar oleh model.

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

F1 Score merupakan rata-rata harmonik dari presisi dan recall, digunakan untuk menyeimbangkan keduanya terutama ketika distribusi kelas tidak seimbang.

$$F1\ Score\ = \frac{2 \times TP}{2 \times TP + FP + FN} \quad (4)$$

Keterangan:
TP (True Positive): jumlah prediksi positif yang benar (model memprediksi positif dan kenyataannya memang positif),

TN (True Negative): jumlah prediksi negatif yang benar (model memprediksi negatif dan kenyataannya memang negatif),

FP (False Positive): jumlah prediksi positif yang salah (model memprediksi positif tetapi sebenarnya negatif),

FN (False Negative): jumlah prediksi negatif yang salah (model memprediksi negatif tetapi sebenarnya positif).

3. HASIL DAN PEMBAHASAN

Setelah model dilatih dengan parameter pada Tabel 1, evaluasi dilakukan pada data test menggunakan metrik yang telah ditentukan. Hasil evaluasi ditampilkan pada Tabel 4.

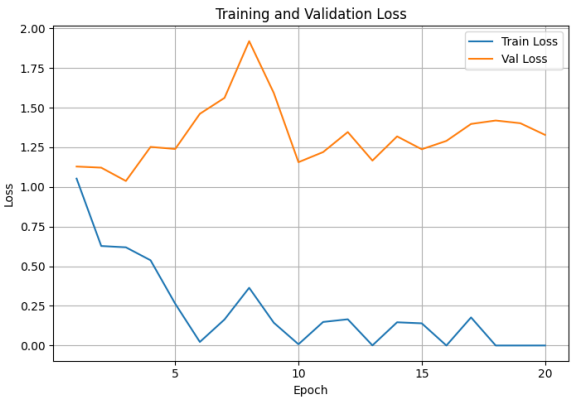
Tabel 4. Hasil Evaluasi Matriks Model

No.	Model	Epoch, Batch Size, Learning Rate	Accuracy (%)	Precision (%)	Recall (%)	F1 (%)
1	VideoMAE	50, 8, 1e-3	20.21	12.99	20.21	09.69
2	VideoMAE	20, 4, 5e-5	68.08	65.58	68.08	69.98
3	TimeSformer	15, 4, 5e-3	22.11	08.77	22.11	05.47
4	TimeSformer	20, 4, 5e-5	**78.84**	**79.30**	**78.84**	**80.34**

Hasil evaluasi menunjukkan bahwa performa kedua model, VideoMAE dan TimeSformer, sangat dipengaruhi oleh konfigurasi pelatihan yang digunakan. VideoMAE pada Eksperimen 1 menunjukkan performa rendah dengan akurasi hanya 20.21%, sedangkan pada Eksperimen 2 terjadi

peningkatan signifikan hingga 68.08%, dengan F1-score, recall, dan precision yang seimbang. Hal serupa terjadi pada TimeSformer, di mana Eksperimen 3 menghasilkan performa rendah (akurasi 22.11%), namun meningkat drastis pada Eksperimen 4 hingga mencapai akurasi 78.84% dan F1-score 79.30%. Hasil terbaik diperoleh oleh TimeSformer pada Eksperimen 4, yang menunjukkan performa lebih unggul dibandingkan VideoMAE, baik dari segi akurasi maupun keseimbangan metrik lainnya. Temuan ini menegaskan bahwa efektivitas model sangat bergantung pada pemilihan parameter dan strategi pelatihan yang tepat.

Namun, berdasarkan grafik loss pada data train dan val, model dengan performa terbaik mengalami overfitting, dimana model terlalu menghafal data latih sehingga kurang mampu mengenali data yang belum pernah dilihat. Kemungkinan overfitting ini terjadi karena dataset dikumpulkan menggunakan tiga perangkat secara bersamaan, sehingga satu aksi yang dilakukan oleh individu yang sama terekam dari beberapa sudut berbeda. Akibatnya, meskipun data tersebut terbagi ke dalam set pelatihan dan pengujian, model yang mengalami overfitting tetap dapat mengenali aksi dalam data uji karena kemiripan visual dengan data latih. Grafik loss untuk eksperimen ke-4 ditampilkan pada Gambar 4.



Gambar 4. Plot Training dan Validation Loss Eksperimen 4

4. KESIMPULAN

Penelitian ini menunjukkan bahwa model VideoMAE berhasil diimplementasikan untuk klasifikasi aksi peserta didik dengan hasil akurasi 68.08%. Dibandingkan dengan model baseline, TimeSformer pada eksperimen keempat memberikan performa lebih tinggi dengan akurasi 78.84%, meskipun menunjukkan indikasi overfitting saat menggunakan hyperparameter 20 untuk Epoch, 4 untuk Batch Size, dan 5e-5 Learning Rate. Hal ini Kemungkinan terjadi karena dataset dikumpulkan menggunakan tiga perangkat secara bersamaan, sehingga satu aksi yang dilakukan oleh individu yang sama terekam dari beberapa sudut berbeda. Akibatnya, meskipun data tersebut terbagi ke dalam set pelatihan dan pengujian, model yang mengalami overfitting tetap dapat mengenali aksi dalam data uji karena kemiripan visual dengan data latih.

REFERENSI

[1] Akmalia, R., Oktapia, D., Hasibuan, E., Hasibuan, I., Azzahrah, N., & Harahap, T. (2023). Pentingnya Evaluasi Peserta Didik dalam Proses Pembelajaran. 5. 4089-4092. 10.31004/jpdk.v5i1.11661.

[2] Bertasius, G., Wang, H., & Torresani, L. (2021, July). Is space-time attention all you need for video understanding?. In Icml (Vol. 2, No. 3, p. 4).

[3] C. Sammut, Encyclopedia of machine learning, G. I. Webb, Ed. Springer, 3 2011. [Online]. Available: https://books.google.com/books/about/Encyclopedia of Machine Learning.html?hl=id&id=i8hQhp1a62UC

[4] Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Gelly, S., Uszkoreit J., & Houlsby, N. (2020). An Image Is Worth 16x16 Words: Transformers for Image Recognition at Scale (Version 2).

[5] Glasscock, J & Tenenbaum, S. (2023, January 11). Action. Stanford Encyclopedia of Philosophy. https://plato.stanford.edu/archives/spr2023/entries/action/

[6] Halim, S., Wahid, R., & Halim, T. (2018). Classroom Observation- a Powerful Tool for Continous Professional Development (CPD). International Journal on Language Research and Education Studies.

[7] Islam, Md Mofijul & Iqbal, Tariq. (2020). HAMLET: A Hierarchical Multimodal Attention-based Human Activity Recognition Algorithm. 10.1109/IROS45743.2020.9340987.

[8] Kong, Yu & Fu, Y. (2022). Human Action Recognition: A Survey: V3. ArXiv.

[9] Popescu, Ana-Cosmina & Mocanu, Irina & Cramariuc, Bogdan. (2020). Fusion Mechanisms for Human Activity Recognition Using Automated Machine Learning. IEEE Access. 8. 1-1. 10.1109/ACCESS.2020.3013406.

[10] Tong, Z., Song, Y., Wang, J., & Wang, L. (2022). Videomae: Masked autoencoders are data-efficient learners for self-supervised video pre-training. Advances in neural information processing systems, 35, 10078-10093.

[11] Voulodimos, A., Doulamis, N.D., Doulamis, A.D., & Protopapadakis, E.E. (2018). Deep Learning for Computer Vision: A Brief Review. Computational Intelligence and Neuroscience, 2018.

[12] Will Kay, Joao Carreira, Karen Simonyan, Brian Zhang, Chloe Hillier, Sudheendra Vijayanarasimhan, Fabio Viola, Tim Green, Trevor Back, Paul Natsev, et al. The kinetics human action video dataset. arXiv preprint arXiv:1705.06950, 2017.