BAB I
PENDAHULUAN

1.1 Latar Belakang

Kehidupan sehari-hari manusia selalu terlibat dalam berbagai aksi. Aksi adalah tindakan disengaja yang muncul dari kondisi mental pelaku dan diarahkan untuk mencapai tujuan tertentu (Glasscock, J. 2023). Sebagai salah satu bentuk interaksi manusia dengan lingkungan, aksi memainkan peran krusial dalam memahami perilaku individu. Aksi-aksi yang dilakukan sering kali mencerminkan keadaan internal seseorang, seperti minat, perhatian, atau keterlibatan dalam suatu aktivitas. Oleh karena itu, pengamatan atau observasi aksi menjadi langkah penting untuk dapat memperoleh wawasan mengenai hubungan antara individu dengan lingkungannya, termasuk dalam konteks pendidikan.

Evaluasi aktivitas peserta didik merupakan bagian penting dalam mengukur pencapaian tujuan pembelajaran (Akmalia dkk., 2023). Pemahaman yang lebih dalam terhadap perilaku atau aksi peserta didik selama proses pembelajaran dapat membantu pendidik, lembaga pendidikan, dan peserta didik sendiri dalam mengidentifikasi kebutuhan individu secara spesifik, menyesuaikan strategi pengajaran, serta menciptakan lingkungan belajar yang lebih efektif. Salah satu indikator penting dalam proses ini adalah keterlibatan peserta didik, yang telah terbukti menjadi prediktor keberhasilan akademik (Carini, R. M., 2006).

Tradisionalnya, pengamatan aksi peserta didik dilakukan secara manual oleh pengamat atau pendidik di dalam kelas selama pembelajaran berlangsung. Pengamat melakukan observasi langsung dengan memperhatikan atau mencatat aksi relevan peserta didik yang dapat menggambarkan performa ketika di kelas. Meskipun metode ini sering digunakan dan mampu memberikan data observasi langsung, prosesnya bergantung pada subjektivitas pengamat. Keberadaan pengamat dapat mempengaruhi perilaku peserta didik karena sadar sedang diamati (Halim dkk., 2018). Metode ini juga memerlukan banyak waktu dan sulit diterapkan dalam kelas dengan jumlah peserta yang besar. Keterbatasan tersebut menunjukkan

perlunya pendekatan yang lebih objektif dan efisien, salah satunya melalui pemanfaatan teknologi *computer vision*.

Perkembangan *deep learning* dalam bidang *computer vision* telah menghasilkan capaian signifikan, khususnya pada empat subdomain utama, yaitu *image classification*, *object detection*, *human action recognition*, dan *pose estimation* (Voulodimos dkk., 2018). *Human action recognition* (HAR), yang berfokus pada pengenalan dan klasifikasi aksi manusia melalui analisis rangkaian gerakan dalam video (Kong & Fu, 2022), menjadi relevan untuk konteks pendidikan. Terdapat beberapa penelitian terdahulu yang mengimplementasikan metode *deep learning* pada lingkup HAR. Islam & Iqbal (2020) berhasil mencapai akurasi tinggi dengan mengimplementasikan mekanisme *Attention* dan arsitektur *Long Short Term Memory* (LTSM) pada dataset HAR publik UTD-MHAD (95.12%) dan UT-Kinect (97.45%). Penelitian lain yang dilakukan oleh Popescu & Mocanu (2024) menggunakan *3D Convolutional Neural Network* (CNN) juga berhasil mencapai akurasi tinggi pada tiga dataset publik dan 1 dataset baru milik penulis, yaitu MSRDailyActivity3D (98.43%), UTD-MHAD (91.41%), NTU RGB+D (90.95%), dan PRECIS HAR (94.38%), dataset buatan peneliti. Meskipun begitu, kinerja model HAR tetap dipengaruhi oleh ukuran data dan kompleksitas tugas (Krizhevsky, A., 2012), sehingga masih diperlukan upaya untuk meningkatkan performa dan efisiensinya.

Dalam beberapa tahun terakhir, *Vision Transformer* (ViT) menjadi alternatif arsitektur baru yang unggul dalam tugas pengenalan visual (Dosovitskiy dkk., 2020). ViT adalah metode berbasis arsitektur *transformer* untuk berbagai tugas pengenalan gambar. ViT menunjukkan performa yang sangat baik, bahkan melampaui CNN sebagai *state-of-the-art* sebelumnya dalam tugas pengenalan gambar (Yao dkk. 2018). Keberhasilan ViT dalam tugas ini mendorong adopsi yang cepat di berbagai tugas lain. Salah satu bidang tersebut adalah pengenalan video, khususnya dalam tugas HAR, dimana informasi spasial dan temporal dibutuhkan untuk menganalisis data. Dalam satu tahun, mulai bermunculan model mengadopsi ViT sebagai backbone model pengenalan video.

Model adaptasi ViT pertama yang dirancang khusus untuk tugas pengenalan video (video transformer) adalah *Time-Space Transformer* atau disebut TimeSformer, yang diperkenalkan oleh Bertasius dkk. (2021). TimeSformer mengadaptasi ViT dengan memperluas mekanisme *attention* dari gambar dua dimensi ke video tiga dimensi, sehingga dapat memproses informasi spasial dan temporal secara bersamaan. Meskipun merupakan desain baru, TimeSformer mampu mencapai hasil terbaik pada beberapa *state-of-the-art* dari dataset *benchmark* HAR, diantaranya pada Kinetics-400 dan Kinetics-600. Dibandingkan dengan 3D *Convolutional Network*, TimeSformer dapat meraih tes efisiensi lebih baik dalam waktu latih lebih cepat.

Salah satu model video transformer lain yang cukup menonjol adalah *Video Masked Autoencoder* (VideoMAE), diperkenalkan oleh Tong dkk. (2022). VideoMAE menggunakan pendekatan *self-supervised learning* dengan cara menyembunyikan sebagian besar frame dalam video (*masking*) dan melatih model untuk merekonstruksi informasi yang hilang. Strategi ini memungkinkan model untuk memahami representasi spasial dan temporal secara efisien tanpa memerlukan data dalam jumlah besar. VideoMAE terbukti unggul dalam berbagai tugas pengenalan aksi pada *benchmark* dataset seperti Kinetics-400 dan Something-Something v2, dengan performa yang kompetitif dibandingkan metode lain yang lebih kompleks. Keunggulan utama VideoMAE terletak pada efisiensinya dalam pelatihan dan kemampuannya untuk menangkap dinamika gerakan dari video dengan lebih baik, menjadikannya kandidat yang kuat untuk penelitian ini. Model akan dilatih mengklasifikasikan aksi di lingkungan kelas pada dataset kecil.

Menurut Chathanadath (2023), seorang profesor dan pakar perilaku, menuliskan di platform *Quora* bahwa sikap positif merupakan sifat yang memicu pertumbuhan dan memungkinkan organisme untuk berkembang, sedangkan sikap negatif mengandung benih-benih kerusakan yang merugikan pertumbuhan dan mengurangi peluang untuk berkembang. Dalam penelitian ini, sikap positif menunjukkan ketertarikan atau partisipasi aktif di dalam kelas yang direpresentasikan dengan aksi mengangkat tangan dan mengangguk. Sebaliknya, sikap negatif menggambarkan ketidak tertarikan terhadap pembelajaran di dalam

kelas ketika pengajar menyampaikan materi, direpresentasikan dengan aksi menggunakan ponsel, menopang kepala di atas meja, dan menunduk. Dataset yang digunakan diambil oleh peneliti karena belum tersedia dataset standar yang sesuai untuk tugas pengenalan dan klasifikasi aksi peserta didik di dalam ruang kelas saat penelitian ini dilakukan.

Oleh karena itu, penelitian ini mengembangkan model untuk mengenali dan mengklasifikasikan aksi peserta didik di ruang kelas menggunakan VideoMAE. Model akan mengklasifikasikan lima aksi yang menggambarkan sikap positif dan negatif peserta didik selama proses pembelajaran di dalam ruang kelas. Penelitian juga menggunakan TimeSformer sebagai model pembanding (*baseline*), mengingat TimeSformer merupakan salah satu pionir dalam penerapan Vision Transformer pada data video dan telah menunjukkan performa kompetitif pada berbagai benchmark pengenalan aksi, seperti pada dataset Kinetics-400. Untuk mengatasi kebutuhan komputasi yang tinggi pada model video *transformer*, penelitian ini memanfaatkan model pralatih (*pretrained*) dan melakukan *fine-tuning*, sebagaimana disarankan oleh Gowda dkk. (2023). Dengan sistem ini, diharapkan proses pengamatan aksi dapat dilakukan secara efisien dan objektif, serta menghasilkan informasi visual yang dapat digunakan untuk mendukung evaluasi proses pembelajaran dan pengembangan sistem pendukung keputusan di bidang pendidikan.

1.2 Rumusan Masalah

Berdasarkan latar belakang diatas, maka rumusan masalah yang akan dibahas dalam penelitian ini di antaranya adalah sebagai berikut:

1. Bagaimana proses membangun dataset video pengenalan aksi peserta didik di dalam ruang kelas?
2. Bagaimana implementasi VideoMAE dalam mengenali dan mengklasifikasikan aksi peserta didik di ruang kelas?
3. Bagaimana performa VideoMAE dalam mengenali dan mengklasifikasikan aksi peserta didik di ruang kelas?

4. Bagaimana perbandingan kinerja VideoMAE dengan model *baseline* TimeSformer dalam mengenali dan mengklasifikasikan aksi peserta didik di ruang kelas?

1.3 Tujuan Penelitian

Berdasarkan rumusan masalah di atas, maka tujuan penelitian ini adalah sebagai berikut:

1. Membangun dataset video pengenalan dan klasifikasi aksi peserta didik di dalam ruang kelas.
2. Mengimplementasi VideoMAE dalam mengenali dan mengklasifikasikan aksi peserta didik di ruang kelas.
3. Mengevaluasi performa VideoMAE dalam mengenali dan mengklasifikasikan aksi peserta didik di ruang kelas.
4. Membandingkan hasil performa VideoMAE dengan model baseline TimeSformer dalam mengenali dan mengklasifikasikan aksi peserta didik di ruang kelas.

1.4 Batasan Masalah

Berdasarkan latar belakang, rumusan masalah, dan tujuan penelitian yang dikemukakan di atas, diperlukan beberapa batasan masalah agar penyelesaian masalah lebih terarah. Batasan masalah dari penelitian ini adalah:

1. Penelitian ini akan fokus pada klasifikasi aksi peserta didik yang dapat diamati secara visual dalam ruang kelas.
2. Model mengenali dan mengklasifikasikan satu peserta didik sedang melakukan satu aksi dalam satu video masukan.
3. Model utama yang digunakan adalah VideoMAE, dan hanya dibandingkan dengan satu model baseline yaitu TimeSformer.
4. Dataset yang digunakan untuk mengembangkan model merupakan dataset buatan sendiri (*custom dataset*) dengan struktur yang telah disesuaikan untuk kebutuhan tugas klasifikasi, terdiri dari 403 video aksi. Aksi yang dapat diklasifikasikan pada penelitian ini meliputi mengangkat tangan, mengangguk, menggunakan telepon genggam, menundukkan kepala, dan bersandar di atas meja.

5. Evaluasi model dilakukan berdasarkan matriks *Accuracy*, presisi, *Recall*, dan *F1-score* pada dataset uji, dengan model terbaik ditentukan berdasarkan nilai akurasi tertinggi.

1.5 Manfaat Penelitian

Penelitian ini diharapkan dapat memberikan solusi yang lebih objektif dan efisien untuk pengamatan aksi peserta didik di dalam kelas, menggantikan metode manual oleh pengawas. Informasi yang dihasilkan dari sistem klasifikasi ini dapat membantu pendidik dalam memahami perilaku siswa secara mendalam, sehingga dapat merancang strategi pembelajaran yang lebih efektif, meningkatkan kualitas interaksi di dalam kelas, dan menciptakan lingkungan belajar yang nyaman. Penelitian ini juga diharapkan dapat memperluas cakupan implementasi teknologi VideoMAE atau pada lingkup video *transformer* dalam domain pendidikan, yang hingga saat ini masih jarang dijelajahi. Hasil penelitian dapat menjadi acuan untuk pengembangan lebih lanjut, baik dalam optimasi performa model pada dataset yang lebih kecil maupun dalam pengaplikasian metode VideoMAE untuk tugas-tugas lain di bidang *computer vision*, khususnya dalam pengenalan aksi manusia.

Penelitian ini memberikan contoh bagaimana teknologi *video transformer*, lebih spesifiknya model *self-supervised learning* seperti VideoMAE dapat dimanfaatkan untuk mengatasi keterbatasan data pelatihan, yang menjadi salah satu tantangan utama dalam penerapan teknologi kecerdasan buatan di berbagai bidang.

1.6 Sistematika Penulisan

Untuk mempermudah melihat dan mengetahui pembahasan pada penelitian ini secara menyeluruh, maka dikemukakan sistematika penulisan penelitian karya ilmiah. Adapun sistematika penulisan penelitian ini terdiri dari lima bagian. Bagian-bagian tersebut adalah sebagai berikut:

BAB I PENDAHULUAN

Pada bab ini diuraikan latar belakang pengusulan penelitian. Dari latar belakang tersebut terbentuk rumusan masalah, tujuan penelitian, dan manfaat penelitian. Di akhir bab terdapat sistematika penulisan yang menjelaskan rincian dari isi setiap bab.

BAB II TINJAUAN PUSTAKA

Bab ini memaparkan landasan teori dan hasil dari penelitian terdahulu terkait masalah yang dibahas pada penelitian ini. Teori yang dibahas pada bab ini termasuk pengenalan aksi manusia, *student engagement,* computer vision, deep learning, transformer, dan VideoMAE, TimeSformer.

BAB III METODE PENELITIAN

Pada bab ini dijelaskan metodologi yang digunakan dalam penelitian, termasuk desain penelitian, alat dan bahan yang digunakan dalam penelitian, serta metode pengumpulan data.

BAB IV HASIL DAN PEMBAHASAN

Bab ini membahas hasil serta analisis dari proses yang telah dilakukan selama penelitian, dimulai dari pengambilan data sampai dengan evaluasi kinerja model.

BAB V SIMPULAN DAN SARAN

Pada bab ini dipaparkan kesimpulan dari hasil yang telah didapatkan terkait klasifikasi aksi peserta didik di dalam kelas menggunakan *Video Masked Autoencoder*. Kemudian dipaparkan saran untuk pengembangan penelitian terkait selanjutnya.