TESIS

PENINGKATAN AKURASI PADA *SLOWFAST NETWORK* MENGGUNAKAN *MULTI-HEAD SELF ATTENTION LAYER*

disusun oleh:

Muhammad Reza

23.55.2507

Konsentrasi : Business Intelligence

FAKULTAS ILMU KOMPUTER

UNIVERSITAS AMIKOM YOGYAKARTA

YOGYAKARTA

2026

TESIS

PENINGKATAN AKURASI PADA *SLOWFAST NETWORK* MENGGUNAKAN *MULTI-HEAD SELF ATTENTION LAYER*

ACCURACY IMPROVEMENT IN SLOWFAST NETWORK USING MULTI-HEAD SELF ATTENTION LAYER

Diajukan untuk memenuhi salah satu syarat memperoleh derajat Pascasarjana

Program Studi Magister PJJ Informatika

disusun oleh:

Muhammad Reza

23.55.2507

Konsentrasi : Business Intelligence

FAKULTAS ILMU KOMPUTER

PROGRAM MAGISTER UNIVERSITAS AMIKOM YOGYAKARTA

YOGYAKARTA

2026

HALAMAN PERSETUJUAN

PENINGKATAN AKURASI PADA *SLOWFAST NETWORK* MENGGUNAKAN *MULTI-HEAD SELF ATTENTION LAYER*

ACCURACY IMPROVEMENT IN SLOWFAST NETWORK USING MULTI-HEAD SELF ATTENTION LAYER

yang disusun dan diajukan oleh

Muhammad Reza

23.55.2507

telah disetujui oleh Dosen Pembimbing Tesis
pada tanggal 02 Januari 2026

Dosen Pembimbing,

Prof. Arief Setyanto, S.Si., M.T., Ph.D.
NIK. 190302036

HALAMAN PENGESAHAN

PENINGKATAN AKURASI PADA *SLOWFAST NETWORK* MENGGUNAKAN *MULTI-HEAD SELF ATTENTION LAYER*

ACCURACY IMPROVEMENT IN SLOWFAST NETWORK USING MULTI-HEAD SELF ATTENTION LAYER

yang disusun dan diajukan oleh

Muhammad Reza

23.55.2507

Telah dipertahankan di depan Dewan Penguji
pada tanggal 02 Januari 2026

Susunan Dewan Penguji

Nama Penguji	Tanda Tangan
I Made Artha Agastya, S.T., M.Eng., Ph.D. **NIK.190302352**	
Prof. Arief Setyanto, S.Si., M.T., Ph.D. **NIK.190302036**	
Dr. Andi Sunyoto, S.Kom., M.Kom. **NIK.190302052**	

Tesis ini telah diterima sebagai salah satu persyaratan
untuk memperoleh gelar Magister Komputer
Tanggal 02 Januari 2026

DEKAN FAKULTAS ILMU KOMPUTER

Prof. Dr. Kusrini, M.Kom.
NIK. 190302106

HALAMAN PERNYATAAN KEASLIAN TESIS

Yang bertandatangan di bawah ini,

Nama mahasiswa : Muhammad Reza
NIM : 23.55.2507

Menyatakan bahwa Tesis dengan judul berikut:

Peningkatan Akurasi Pada *Slowfast Network* Menggunakan *Multi-Head Self Attention Layer*

Dosen Pembimbing : Prof. Arief Setyanto, S.Si., M.T., Ph.D.

1. Karya tulis ini adalah benar-benar ASLI dan BELUM PERNAH diajukan untuk mendapatkan gelar akademik, baik di Universitas AMIKOM Yogyakarta maupun di Perguruan Tinggi lainnya.
2. Karya tulis ini merupakan gagasan, rumusan dan penelitian SAYA sendiri, tanpa bantuan pihak lain kecuali arahan dari Tim Dosen Pembimbing.
3. Dalam karya tulis ini tidak terdapat karya atau pendapat orang lain, kecuali secara tertulis dengan jelas dicantumkan sebagai acuan dalam naskah dengan disebutkan nama pengarang dan disebutkan dalam Daftar Pustaka pada karya tulis ini.
4. Perangkat lunak yang digunakan dalam penelitian ini sepenuhnya menjadi tanggung jawab SAYA, bukan tanggung jawab Universitas AMIKOM Yogyakarta.
5. Pernyataan ini SAYA buat dengan sesungguhnya, apabila di kemudian hari terdapat penyimpangan dan ketidakbenaran dalam pernyataan ini, maka SAYA bersedia menerima SANKSI AKADEMIK dengan pencabutan gelar yang sudah diperoleh, serta sanksi lainnya sesuai dengan norma yang berlaku di Perguruan Tinggi.

Yogyakarta, 02 Januari 2026

Yang Menyatakan,

Muhammad Reza

HALAMAN PERSEMBAHAN

Karya ini saya persembahkan istimewa untuk keluarga tercinta, Ayah dan Ibu, yang doanya senantiasa menjadi jembatan bagi setiap kemudahan yang saya lalui, terima kasih atas kasih sayang dan pengorbanan yang tak terbatas. Kepada istri, pendamping hidup yang selalu sabar memberikan dukungan, pengertian, serta menjadi sumber ketenangan dikala lelah, terima kasih telah menjadi alasan terkuat bagi saya untuk terus melangkah maju. Tak lupa untuk sahabat seperjuangan, rekan-rekan Magister Informatika yang telah berbagi tawa, diskusi, dan semangat dalam menapaki jalan ilmu hingga dititik ini. Semoga karya ini menjadi bagian kecil yang bermanfaat dalam pengembangan ilmu pengetahuan dan praktik strategis di masa depan.

KATA PENGANTAR

Segala puji bagi Allah SWT, yang telah melimpahkan rahmat, taufik, dan hidayah-Nya sehingga penulis dapat menyelesaikan tesis ini yang berjudul **"Peningkatan Akurasi Pada *Slowfast Network* Menggunakan *Multi-Head Self Attention Layer*".** Sholawat, salam senantiasa tercurah kepada Nabi Muhammad SAW, beserta keluarga, sahabat, dan seluruh umatnya hingga akhir zaman.

Tesis ini disusun sebagai salah satu syarat untuk memperoleh gelar Magister Komputer pada Jurusan Informatika di Universitas Amikom Yogyakarta. Penyusunan penelitian ini tidak terlepas dari bantuan, bimbingan, serta dukungan dari berbagai pihak. Dengan penuh rasa syukur, penulis menyampaikan terima kasih kepada keluarga tercinta atas doa dan semangat yang tak pernah henti kepada dosen pembimbing dan seluruh sivitas akademika atas arahan dan ilmu yang diberikan, serta kepada sahabat-sahabat seperjuangan dan semua pihak yang telah membantu dalam proses penelitian ini.

Akhir kata, semoga tesis ini dapat memberikan manfaat bagi dunia akademik serta menjadi kontribusi bagi perkembangan ilmu pengetahuan, khususnya dalam bidang Informastika. Semoga Allah SWT senantiasa meridhoi setiap langkah kita. Wassalamu'alaikum warahmatullahi wabarakatuh.

Yogyakarta, 02 Januari 2026

Penulis

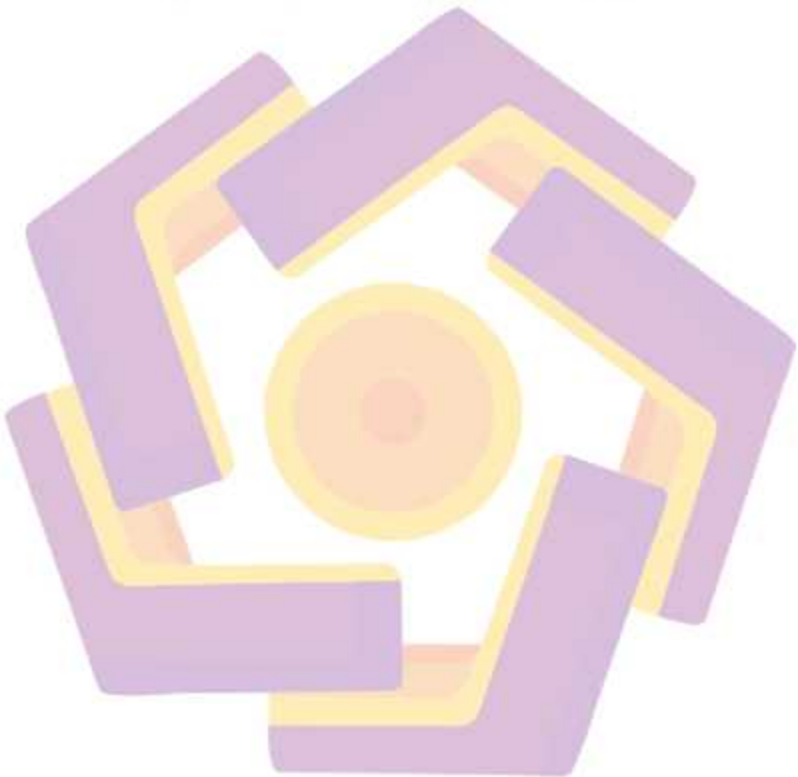
DAFTAR ISI

HALAMAN JUDUL i
HALAMAN PERSETUJUAN ii
HALAMAN PENGESAHAN iii
HALAMAN PERNYATAAN KEASLIAN TESIS iv
HALAMA PERSEMBAHAN v
KATA PENGANTAR vi
DAFTAR ISI vi
DAFTAR TABEL iix
DAFTAR GAMBAR x
DAFTAR LAMBANG DAN SINGKATAN vii
DAFTAR ISTILAH xii
INTISARI vii
ABSTRACT xiv
BAB I PENDAHULUAN 1
1.1. Latar Belakang 1
1.2. Rumusan Masalah 5
1.3. Batasan Masalah 5
1.4. Tujuan Penelitian 6
1.5. Manfaat Penelitian 6
1.6. Hipotesis 8
BAB II TINJAUAN PUSTAKA 9
2.1. Tinjauan Pustaka 9
2.2. Keaslian Penelitian 16
2.3. Landasan Teori 19
BAB III METODE PENELITIAN 30
3.1. Jenis, Sifat, dan Pendekatan Penelitian 30
3.2. Metode Pengumpulan Data 31
3.3. Metode Analisis Data 32
3.4. Alur Penelitian 34
BAB IV HASIL PENELITIAN DAN PEMBAHASAN 46
4.1. Deskripsi Umum Penelitian 46
4.2. Hasil dan Evaluasi Model 55
4.3. Komparasi dengan Penelitian Terdahulu 64
4.4. Analisis Visual dan Efektivitas MHSA 67
4.5. Analisis Dampak Preprocessing 70

BAB V PENUTUP 72
5.1. Kesimpulan 72
5.2. Saran 73
DAFTAR PUSTAKA 74

DAFTAR TABEL

Tabel 2.1. Matriks literatur review dan posisi penelitian 16
Tabel 4.1. Hasil Evaluasi Model SlowFast + Dataset SUST 56
Tabel 4.2. Hasil Evaluasi Model SlowFast + MHSA + Dataset SUST 59
Tabel 4.3. Hasil Evaluasi Model SlowFast + MHSA + SUST + NITYMED..... 62
Tabel 4.4. Komparasi Komprehensif dengan Penelitian Terdahulu 65



DAFTAR GAMBAR

Gambar 2.1. Arsitektur Slowfast Network 28
Gambar 3.1. Alur Penelitian SlowFast + MHSA 34
Gambar 3.2. Arsitektur SlowFast dan Multi-Head Self-Attention 42
Gambar 4.1. Dataset SUST-DDD dan NITYMEDD 49
Gambar 4.2. Dataset Setelah Preprocessing 50
Gambar 4.3. Alur Pemrosesan Paralel pada Arsitektur SlowFast 52
Gambar 4.4. Alur Pemrosesan Multi-Head Self Attention 53
Gambar 4.5. Arsitektur model SlowFast dengan penambahan MHSA 54
Gambar 4.6. Confusion Matrix Model SlowFast pada Dataset SUST 58
Gambar 4.7. Confusion Matrix Model SlowFast + MHSA pada Dataset SUST .. 61
Gambar 4.8. Confusion Matrix Model SF + MHSA + Dataset Gabungan 64
Gambar 4.9. Perbandingan Visualisasi Atensi (Grad-CAM) Dataset SUST 67
Gambar 4.10. Visualisasi Atensi Model Final pada Dataset NITYMED 69

DAFTAR LAMBANG DAN SINGKATAN

α	Learning rate atau bobot alpha pada focal loss
β	Rasio channel antara jalur slow dan fast
τ	Stride temporal (jarak antar frame)
S	Jalur Slow (Slow Pathway)
F	Jalur Fast (Fast Pathway)
Q	Query (pada mekanisme Attention)
K	Key (pada mekanisme Attention)
V	Value (pada mekanisme Attention)
W	Matriks Bobot (Weight)
b	Bias
σ	Fungsi Sigmoid

DAFTAR ISTILAH

MHSA	Multi-Head Self Attention
NITYMED	Night Time Yawning Microsleep Eyeblink Distraction dataset
3D CNN	3-Dimensional Convolutional Neural Network
AI	Artificial Intelligence
CNN	Convolutional Neural Network
FPS	Frames Per Second
Grad-CAM	Gradient-weighted Class Activation Mapping
ReLU	Rectified Linear Unit
RGB	Red Green Blue
SlowFast	Arsitektur deep learning yang menggunakan dua jalur (slow dan fast) untuk pemrosesan video
SUST-DDD	SUST Driver Drowsiness Dataset

INTISARI

Kecelakaan lalu lintas akibat kelelahan pengemudi merupakan masalah keselamatan global yang memerlukan solusi deteksi dini yang akurat. Penelitian ini bertujuan untuk meningkatkan kinerja sistem deteksi kantuk berbasis video menggunakan arsitektur SlowFast Network yang diintegrasikan dengan mekanisme Multi-Head Self Attention (MHSA). Penelitian eksperimental ini menggunakan pendekatan kuantitatif dengan menggabungkan dataset SUST-DDD dan NITYMED guna mencakup variasi kondisi pencahayaan siang dan malam hari. Metode pengembangan model melibatkan pra-pemrosesan data menggunakan teknik sliding window berdurasi 2 detik, normalisasi, dan augmentasi data, serta divalidasi menggunakan analisis visual Grad-CAM untuk mengukur efektivitas fokus atensi model.

Hasil penelitian menunjukkan bahwa integrasi MHSA secara signifikan meningkatkan kemampuan model dalam menangkap fitur spasial-temporal yang kritis. Model final berhasil mencapai akurasi sebesar 96,65% pada data uji gabungan, dengan nilai Recall untuk kelas drowsy mencapai 99%. Capaian ini mengungguli model baseline tanpa atensi yang hanya mencapai akurasi 84,48% pada pengujian sejenis. Analisis visual membuktikan bahwa mekanisme atensi berhasil memfokuskan pembelajaran model pada area wajah yang relevan, seperti mata dan mulut, serta secara efektif melakukan supresi terhadap gangguan visual dari latar belakang maupun kondisi minim cahaya.

Dapat disimpulkan bahwa kombinasi arsitektur SlowFast dengan MHSA serta penggunaan dataset yang bervariasi mampu menghasilkan model deteksi kantuk yang handal (robust) dan memiliki generalisasi tinggi. Sistem ini terbukti efektif dalam mengenali tanda-tanda kantuk baik yang bersifat halus maupun eksplisit dalam berbagai kondisi lingkungan.

Kata kunci: Deteksi Kantuk Pengemudi, SlowFast Network, Multi-Head Self Attention, Computer Vision, Video Classification

ABSTRACT

Traffic accidents caused by driver fatigue represent a significant global safety issue, necessitating accurate early detection solutions. This study aims to enhance the performance of video-based drowsiness detection systems by utilizing the SlowFast Network architecture integrated with a Multi-Head Self Attention (MHSA) mechanism. This experimental research employs a quantitative approach by combining the SUST-DDD and NITYMED datasets to encompass variations in both daytime and nighttime lighting conditions. The model development method involves data preprocessing using a 2-second sliding window technique, normalization, and data augmentation, further validated through Grad-CAM visual analysis to assess the effectiveness of the model's attentional focus.

The results indicate that the integration of MHSA significantly improves the model's capability to capture critical spatiotemporal features. The final model achieved an accuracy of 96.65% on the combined test set, with a Recall value for the drowsy class reaching 99%. This performance outperforms the baseline model without attention, which only achieved an accuracy of 84.48% in similar testing. Visual analysis demonstrates that the attention mechanism successfully focuses the model's learning on relevant facial areas, such as the eyes and mouth, while effectively suppressing visual noise from the background and low-light conditions.

It can be concluded that the combination of the SlowFast architecture with MHSA, along with the use of diverse datasets, is capable of producing a robust drowsiness detection model with high generalization capabilities. The system proves effective in recognizing both subtle and explicit signs of drowsiness across various environmental conditions.

Keyword: *Driver Drowsiness Detection, SlowFast Network, Multi-Head Self Attention, Computer Vision, Video Classification*

BAB I

PENDAHULUAN

1.1. Latar Belakang Masalah

Keselamatan berkendara menjadi isu penting di berbagai negara, terutama dengan meningkatnya jumlah kecelakaan lalu lintas yang sebagian besar disebabkan oleh perilaku pengemudi yang tidak aman, seperti mengantuk, menggunakan ponsel saat berkendara, atau kurangnya fokus. Menurut data dari Organisasi Kesehatan Dunia (WHO), kecelakaan lalu lintas adalah salah satu penyebab utama kematian di seluruh dunia, dan perilaku pengemudi yang lalai berkontribusi besar terhadap tingginya angka kecelakaan tersebut (WHO, *Global Status Report on Road Safety*, 2023). Seiring perkembangan teknologi, berbagai upaya telah dilakukan untuk mengurangi kecelakaan ini, salah satunya adalah penggunaan teknologi *computer vision* untuk mendeteksi perilaku pengemudi (Ajay Kumar, 2023).

Kecelakaan yang disebabkan oleh perilaku berbahaya pengemudi, seperti mengantuk atau kurangnya fokus, dapat mengakibatkan kerugian yang signifikan, baik dari segi manusia maupun material. Oleh karena itu, diperlukan suatu sistem yang dapat mendeteksi dan mencegah perilaku berbahaya tersebut untuk meningkatkan keselamatan pengemudi (Fangming Qu, 2024). Namun, penelitian sebelumnya menghadapi beberapa keterbatasan. Sebagai contoh, algoritma berbasis *Convolutional Neural Networks* (CNN) sering kali terbatas dalam menangkap dinamika temporal dari perilaku pengemudi, seperti perubahan cepat

pada gerakan tangan atau wajah. Selain itu, metode seperti *Haar Cascade* dan pendekatan berbasis fitur (seperti *Eye Aspect Ratio* atau EAR) memiliki akurasi yang terbatas dalam kondisi pencahayaan yang buruk atau sudut pandang kamera yang tidak ideal. Model deteksi berbasis pengenalan fitur wajah sensitif terhadap perubahan kondisi lingkungan, seperti pencahayaan yang buruk dan jarak kamera yang terlalu dekat dengan pengemudi. Sensitivitas ini menjadi kelemahan signifikan dalam aplikasi dunia nyata. Penelitian tersebut merekomendasikan pengujian yang lebih luas dengan variasi demografi dan kondisi jalan yang lebih beragam untuk memastikan bahwa model dapat berfungsi secara andal dalam berbagai skenario (J. Robert Theivadas, 2024).

Penerapan teknik *computer vision* dalam deteksi perilaku pengemudi memberikan peluang untuk menganalisis perubahan pola visual secara otomatis melalui kamera yang dipasang di dalam kendaraan. Deteksi *real-time* dari perilaku berisiko, seperti mengemudi dengan kantuk atau kehilangan fokus, dapat memberikan peringatan dini kepada pengemudi dan meningkatkan keselamatan berkendara. Salah satu tantangan utama adalah menangkap pola perilaku secara konsisten meskipun terdapat variasi pencahayaan dan posisi kamera yang tidak seragam (Jing Liu, 2021).

Penelitian ini mengadopsi arsitektur SlowFast Network yang dirancang untuk menangkap informasi spasial dan temporal secara bersamaan melalui dua jalur pemrosesan (slow dan fast pathway). Arsitektur ini efektif dalam memahami detail gerakan halus sekaligus mendeteksi perubahan visual yang signifikan pada

rangkaian video. Namun, model ini masih memiliki kelemahan dalam memprioritaskan fitur-fitur yang benar-benar relevan untuk tugas deteksi kantuk. Dalam kondisi nyata, video pengemudi mengandung banyak informasi visual yang tidak terkait langsung dengan perilaku kantuk, seperti perubahan pencahayaan, gerakan latar belakang, atau noise akibat getaran kamera. SlowFast Network cenderung memproses seluruh informasi ini secara merata, sehingga performa dapat terpengaruh oleh fitur yang tidak penting. Keterbatasan ini sejalan dengan temuan bahwa arsitektur konvolusi 3D, termasuk model dua jalur, memiliki keterbatasan dalam menangkap hubungan global antar frame dan perubahan gerakan yang halus, sehingga rentan melewatkan pola temporal yang jarang terjadi (Qianyi Jiang, 2023). Untuk mengatasi kelemahan ini, penelitian ini menambahkan mekanisme Multi-Head Self Attention (MHSA) yang mampu memfokuskan perhatian model pada fitur-fitur spasio-temporal yang paling signifikan, serta mengabaikan informasi yang tidak relevan. Dengan demikian, diharapkan integrasi MHSA dapat meningkatkan kemampuan model dalam mengidentifikasi pola perilaku kantuk secara lebih akurat, bahkan pada kondisi pencahayaan yang buruk atau sudut kamera yang bervariasi.

Selain tantangan arsitektur model, kualitas dan karakteristik dataset juga berperan penting dalam keberhasilan sistem deteksi kantuk. Dataset deteksi kantuk yang tersedia saat ini memiliki perbedaan signifikan dalam hal ekspresi, pencahayaan, dan latar belakang, sehingga model yang dilatih pada satu dataset sering kali sulit melakukan generalisasi pada dataset lain. Misalnya, SUST Driver Drowsiness Dataset (SUST-DDD) berisi rekaman video dari 19 pengemudi yang

merekam dirinya sendiri menggunakan ponsel pribadi dalam kondisi mengemudi nyata tanpa instruksi khusus. Meskipun data ini merefleksikan ekspresi alami, sifat natural tersebut membuat perbedaan antara label drowsy dan not_drowsy menjadi sangat halus, sehingga rentan terjadi ambiguitas label (Esra Kavalci Yilmaz, 2022). Sebaliknya, NITYMED Nighttime Drowsiness Dataset merekam pengemudi di malam hari dengan dua kondisi kantuk yang eksplisit yawning dan microsleep, di bawah pencahayaan rendah yang realistis. Data ini memiliki ekspresi kantuk yang jelas, tetapi terbatas pada skenario malam hari (Nikos Petrellis, 2022).

Berdasarkan perbedaan karakteristik tersebut, penelitian ini melakukan penggabungan kedua dataset untuk memanfaatkan kelebihan masing-masing. SUST-DDD memberikan variasi kondisi siang/malam dan ekspresi natural, sementara NITYMED memberikan contoh kantuk yang eksplisit dan tantangan pencahayaan rendah. Penggabungan ini diharapkan dapat meningkatkan kemampuan generalisasi model dalam mengenali perilaku kantuk di berbagai kondisi nyata. Oleh karena itu, penelitian ini bertujuan untuk mengembangkan sistem deteksi perilaku pengemudi menggunakan teknik computer vision dengan SlowFast Network yang dipadukan dengan Multi-Head Self-Attention. Sistem ini diharapkan dapat mendeteksi perilaku berbahaya pengemudi secara real-time, menganalisis data video untuk mengidentifikasi pola risiko, dan memberikan peringatan dini. Penelitian ini mencakup beberapa langkah utama, mulai dari pengumpulan data video pengemudi, prapemrosesan data, pengembangan dan pelatihan model deteksi, hingga evaluasi kinerja model secara eksperimental.

1.2. Rumusan Masalah

Terdapat beberapa rumusan masalah pada penelitian yang nantinya akan diselesaikan seperti :

a. Bagaimana membangun model deteksi perilaku kantuk pengemudi berbasis video menggunakan arsitektur SlowFast Network dan Multi-Head Self Attention (MHSA)?
b. Apa pengaruh penambahan Multi-Head Self Attention terhadap performa model SlowFast Network dalam mendeteksi perilaku kantuk pengemudi?
c. Bagaimana pengaruh penggabungan dua dataset SUST-DDD dan NITYMED terhadap performa model dalam mendeteksi perilaku kantuk pengemudi?

1.3. Batasan Masalah

Penelitian ini memiliki beberapa batasan yang perlu diperhatikan agar ruang lingkup penelitian tetap fokus dan terarah, yaitu:

a. Penelitian ini hanya mencakup deteksi perilaku kantuk pengemudi, tanpa mencakup perilaku lain seperti penggunaan ponsel atau gangguan visual lainnya.
b. Fokus penelitian adalah pada pengembangan dan analisis performa model SlowFast Network dengan penambahan mekanisme Multi-Head Self Attention (MHSA).
c. Sistem yang dibangun tidak menggunakan teknik segmentasi wajah atau deteksi fitur wajah eksplisit (seperti mata atau mulut).

d. Sistem deteksi yang dikembangkan hanya akan diuji pada skenario simulasi dan rekaman video pengemudi, bukan pada kendaraan yang sedang bergerak di jalan raya
e. Dataset yang akan digunakan dalam penelitian ini yaitu dataset SUST-dataset

1.4. Tujuan Penelitian

Adapun tujuan dari penelitian ini yaitu:

a. Mengembangkan model deteksi perilaku kantuk pengemudi berbasis video dengan menggunakan arsitektur SlowFast Network dan mekanisme Multi-Head Self Attention (MHSA).
b. Menganalisis pengaruh penggabungan dataset SUST-DDD dan NITYMED terhadap akurasi model deteksi kantuk.
c. Menganalisis pengaruh penambahan Multi-Head Self Attention terhadap performa model SlowFast Network.

1.5. Manfaat Penelitian

Penelitian ini diharapkan dapat memberikan manfaat berikut:

a. Memberikan kontribusi terhadap pengembangan metode deteksi perilaku pengemudi berbasis video, khususnya dalam konteks analisis temporal menggunakan arsitektur SlowFast Network yang dikombinasikan dengan Multi-Head Self Attention (MHSA).
b. Sistem yang dikembangkan dapat digunakan untuk mendeteksi perilaku pengemudi yang berisiko, seperti kantuk, sehingga berpotensi mengurangi kecelakaan dan kerugian material.

c. Dapat berfungsi sebagai bagian dari program Keselamatan dan Kesehatan Kerja (K3) di perusahaan transportasi atau logistik, dengan membantu mengidentifikasi dan mengurangi perilaku berbahaya secara otomatis.

1.6. Hipotesis

Berdasarkan tinjauan pustaka dan landasan teori, penelitian ini mengajukan dua hipotesis utama. Pertama, terkait pengaruh penambahan Multi-Head Self Attention (MHSA) pada arsitektur SlowFast Network. H0 tidak terdapat perbedaan signifikan dalam performa deteksi perilaku kantuk pengemudi antara model SlowFast Network tanpa MHSA dan model SlowFast Network dengan MHSA. H1 terdapat perbedaan signifikan dalam performa deteksi perilaku kantuk pengemudi antara model SlowFast Network tanpa MHSA dan model SlowFast Network dengan MHSA, di mana penambahan MHSA meningkatkan akurasi, presisi, recall, dan F1-score, terutama pada kondisi pencahayaan rendah dan sudut kamera yang bervariasi. Kedua, terkait pengaruh penggabungan dataset SUST-DDD dan NITYMED. H0 tidak terdapat perbedaan signifikan dalam performa deteksi perilaku kantuk pengemudi antara model yang dilatih menggunakan satu dataset (SUST-DDD atau NITYMED) dengan model yang dilatih menggunakan gabungan kedua dataset tersebut. H1 terdapat perbedaan signifikan dalam performa deteksi perilaku kantuk pengemudi antara model yang dilatih menggunakan satu dataset dan model yang dilatih menggunakan gabungan kedua dataset, di mana penggabungan dataset meningkatkan kemampuan generalisasi model terhadap variasi ekspresi kantuk, kondisi pencahayaan, dan latar belakang.

BAB II

TINJAUAN PUSTAKA

2.1. Tinjauan Pustaka

Penelitian ini memerlukan tinjauan pustaka, oleh karena itu peneliti mengumpulkan data dari berbagai penelitian sebelumnya untuk mengevaluasi keuntungan dan hambatan yang ditemukan. Tujuan langkah ini adalah agar peneliti dapat memahami keterkaitan antara penelitian terdahulu dengan yang sedang dilakukan, serta untuk menghindari duplikasi penelitian. Penelitian sebelumnya menjadi bagian penting yang dianalisis dalam studi ini. Tinjauan literatur ini juga bertujuan untuk menunjukkan bahwa penelitian yang dilakukan oleh penulis memiliki nilai yang signifikan dan memberikan kontribusi besar terhadap ilmu pengetahuan. Berikut adalah beberapa tinjauan terkait penelitian sebelumnya yang berkaitan dengan data dan metode yang digunakan sebagai acuan.

Penelitian oleh (Tsiktsiris et al., 2024) mengusulkan sistem deteksi kejadian abnormal dalam transportasi umum berbasis arsitektur multimodal yang diperluas dari SlowFast Network. Penelitian ini mengintegrasikan data video RGB, kedalaman (depth), dan audio untuk mendeteksi peristiwa berisiko seperti perkelahian, jatuh, atau pencurian di dalam kendaraan otonom. Untuk menangkap dinamika gerakan cepat dan informasi spasial secara bersamaan, digunakan jalur lambat (slow path) untuk menangkap detail spasial dan jalur cepat (fast paths) untuk menangkap perubahan cepat pada tiap modality. Hasil dari penelitian ini menunjukkan bahwa model multimodal (RGB + depth + audio) mencapai akurasi

deteksi sebesar 85,1%, lebih tinggi dibandingkan arsitektur video lain seperti MoViNet dan X3D. Penelitian ini menunjukkan bahwa arsitektur berbasis SlowFast Network yang diperluas sangat efektif untuk klasifikasi video aksi yang kompleks dan cepat, serta menunjukkan potensi besar untuk digunakan dalam sistem keselamatan transportasi. Kelebihan dari penelitian ini adalah pendekatannya yang menyeluruh terhadap pemrosesan multimodal dan kemampuan deteksi waktu nyata. Namun, kelemahan yang dicatat adalah meningkatnya kompleksitas dan konsumsi komputasi akibat banyaknya jalur input yang digunakan. Relevansi penelitian ini dengan penelitian yang dilakukan penulis terletak pada penggunaan arsitektur SlowFast dalam mendeteksi perilaku dalam video berdurasi pendek secara efisien, yang dalam konteks ini difokuskan untuk deteksi kantuk pada pengemudi berbasis video.

Penelitian oleh (Zhao et al., 2022) membahas pengenalan aksi pengemudi dalam video naturalistik dengan pendekatan temporal action localization (TAL) dan klasifikasi aktivitas menggunakan kombinasi arsitektur modern seperti Swin Transformer, I3D, serta berbagai teknik post-processing untuk meningkatkan ketepatan deteksi. Studi ini dilakukan pada track 3 AI City Challenge, dengan data dari 15 pengemudi yang melakukan 18 aktivitas berbeda, termasuk distraksi seperti menggunakan ponsel dan makan di dalam kendaraan. Untuk menangani kompleksitas data multi-view dan durasi aksi yang bervariasi, peneliti menerapkan pendekatan berbasis multi-perspektif dan modul temporal proposal untuk menangkap batas waktu aksi dengan lebih akurat. Hasil akhir menunjukkan nilai F1-score sebesar 29.05%, dengan peningkatan presisi dari 6.67% menjadi 33.33%

setelah integrasi modul koreksi berbasis deteksi objek dan keypoint. Namun, meskipun pendekatan mereka cukup komprehensif, penelitian ini masih menghadapi tantangan dalam hal performa klasifikasi yang rendah akibat keterbatasan kualitas dan jumlah data video, serta kompleksitas perilaku yang terekam. Oleh karena itu, penelitian ini masih menyisakan ruang untuk pengembangan sistem yang lebih efisien dan akurat. Penelitian yang dilakukan oleh penulis merupakan pengembangan dari pendekatan ini, dengan mengadopsi arsitektur video berbasis dual-stream (SlowFast) yang dikombinasikan dengan Multi-Head Self Attention (MHSA) untuk meningkatkan akurasi deteksi kantuk tanpa memerlukan segmentasi perilaku secara manual.

Penelitian oleh (Liang et al., 2024) mengkaji pengenalan dan prediksi perubahan jalur kendaraan sebagai masalah pengenalan aksi berbasis video dengan menggunakan berbagai arsitektur 3D Convolutional Neural Networks (3D CNN), termasuk I3D, X3D, dan SlowFast. Penelitian ini mengembangkan dua pendekatan: RGB+3DN yang hanya menggunakan video kamera depan, serta RGB+BB+3DN yang menambahkan informasi bounding box kendaraan pada tiap frame. Model diuji pada dataset PREVENTION dan dievaluasi berdasarkan skenario klasifikasi serta prediksi perubahan jalur dengan variasi waktu (TTE: Time-To-Event). Hasil menunjukkan bahwa SlowFast-R50 mencapai akurasi klasifikasi 81.04% dengan data RGB murni, dan model X3D-S bahkan mencapai 84.79%. Penelitian ini juga menunjukkan bahwa penggunaan bounding box secara signifikan meningkatkan akurasi hingga 98.50%, namun membutuhkan informasi tambahan yang tidak selalu tersedia secara real-time. Kelebihan dari studi ini adalah desain end-to-end tanpa

segmentasi manual dan validasi terhadap berbagai konfigurasi temporal. Namun demikian, keterbatasannya terletak pada ketergantungan performa terhadap variasi data dan resolusi visual target, serta kebutuhan akan data anotasi tambahan untuk prediksi tingkat lanjut. Penelitian yang dilakukan oleh penulis merupakan bentuk pengembangan dari studi ini, dengan fokus pada penggunaan SlowFast Network yang dikombinasikan dengan Multi-Head Self Attention (MHSA) untuk meningkatkan deteksi perilaku kantuk pengemudi secara lebih efisien dan akurat, tanpa bergantung pada anotasi bounding box atau fitur wajah eksplisit.

Penelitian oleh (Yılmaz & Akcayol, 2024) memperkenalkan dataset baru bernama SUST-DDD (Sivas University of Science and Technology Driver Drowsiness Dataset), yang dirancang untuk mendeteksi kantuk pengemudi secara lebih realistis melalui data video dari pengendara sungguhan di kondisi jalan sebenarnya. Berbeda dengan dataset simulasi seperti NTHU-DDD atau RLDD, data pada SUST-DDD dikumpulkan langsung oleh 19 partisipan dengan menggunakan kamera ponsel saat mereka benar-benar merasa kantuk atau normal selama berkendara. Dataset ini terdiri dari 2.074 video berdurasi 10 detik, dengan resolusi 224x224x3, yang diklasifikasikan oleh juri independen. Dalam evaluasi performa, penulis menggunakan beberapa arsitektur CNN yang dikombinasikan dengan LSTM, seperti VGG19+LSTM, VGG16+LSTM, dan AlexNet+LSTM. Hasil terbaik diperoleh dari model VGG19+LSTM dengan akurasi 90,53%, precision 91,74%, recall 91,28%, dan F1-score 91,46%. Kelebihan dari dataset ini adalah sifatnya yang realistis dan tidak menginstruksikan perilaku tertentu selama perekaman, sehingga ekspresi kantuk bersifat alami. Namun, salah satu

keterbatasannya adalah resolusi dan pencahayaan video yang bervariasi tergantung perangkat peserta, yang dapat memengaruhi stabilitas ekstraksi fitur. Penggunaan dataset ini dalam penelitian penulis menjadi sangat relevan karena mendukung pendekatan berbasis video utuh dan temporal, tanpa ketergantungan pada deteksi fitur wajah eksplisit. Selain itu, penelitian ini mengembangkan model yang berbeda dari pendekatan CNN+LSTM sebelumnya, dengan menerapkan SlowFast Network yang diperkuat Multi-Head Self Attention (MHSA) untuk meningkatkan efisiensi klasifikasi perilaku kantuk secara langsung dari data video.

Penelitian oleh (Khattak, 2024) berfokus pada hubungan antara dinamika pose kepala pengemudi dan kejadian keselamatan kritis (Safety Critical Events, SCEs) dalam konteks mengemudi. Melalui data studi naturalistik dan algoritma seperti RetinaFace untuk deteksi wajah serta Hopenet untuk estimasi pose kepala (yaw, pitch, roll), penelitian ini mengevaluasi bagaimana perubahan arah kepala berhubungan dengan peningkatan risiko kecelakaan. CNN digunakan untuk mengekstraksi fitur wajah dan memprediksi tingkat kantuk berdasarkan sudut Euler, dengan hasil menunjukkan nilai Mean Absolute Error (MAE) sebesar 6.1 derajat menggunakan Multi-Loss ResNet50. Kelebihan penelitian ini adalah penggunaan data mengemudi nyata dan pendekatan berbasis fitur geometrik yang eksplisit. Namun, pendekatan ini sangat bergantung pada kualitas deteksi wajah, yang menjadi kelemahan signifikan dalam kondisi pencahayaan rendah atau sudut pandang kamera yang tidak ideal. Oleh karena itu, penelitian ini masih menyisakan ruang untuk pendekatan yang tidak bergantung pada deteksi bagian wajah tertentu. Penelitian penulis merupakan pengembangan dari hal tersebut dengan mengadopsi

SlowFast Network dan MHSA untuk mendeteksi perilaku kantuk secara langsung dari video penuh tanpa perlu segmentasi wajah.

Dalam penelitian (Yang, 2024), sistem bernama DANet (Driver Attention Network) dikembangkan sebagai kerangka kerja pembelajaran multitugas berbasis CNN untuk memantau perhatian pengemudi. Model ini mampu secara simultan mendeteksi pose kepala, arah pandangan, deteksi landmark wajah, serta kondisi seperti menguap dan mata tertutup. Arsitektur yang digunakan menggabungkan multi-task CNN dan Dual-loss Block (DLB) untuk mengoptimalkan prediksi berbasis klasifikasi dan regresi. Hasilnya menunjukkan performa deteksi yang efisien dengan nilai MAE sebesar 5.76 derajat untuk pose kepala dan 2.61 piksel untuk landmark wajah. Keunggulan utama dari pendekatan ini adalah efisiensi pemrosesan berbagai kondisi visual wajah dalam satu arsitektur. Namun, ketergantungan pada segmentasi wajah menjadikan pendekatan ini kurang adaptif dalam kondisi nyata yang bervariasi. Penelitian yang dilakukan oleh penulis menyempurnakan arah ini dengan berfokus pada pemrosesan video tanpa segmentasi wajah, melalui kombinasi SlowFast Network dan MHSA untuk meningkatkan akurasi dan generalisasi dalam deteksi perilaku kantuk pengemudi.

Berdasarkan tinjauan terhadap enam penelitian terdahulu tersebut, dapat disimpulkan bahwa metode deteksi kantuk pengemudi berbasis video masih memiliki tantangan dalam hal akurasi, efisiensi komputasi, dan adaptasi terhadap kondisi nyata. Beberapa pendekatan terdahulu masih bergantung pada segmentasi wajah atau anotasi bounding box, yang rentan terhadap gangguan visual dan bising. Oleh karena itu, penelitian ini dilakukan sebagai pengembanga dari studi

sebelumnya, dengan mengusulkan arsitektur SlowFast Network yang dikombinasikan dengan Multi-Head Self Attention (MHSA) untuk meningkatkan akurasi klasifikasi temporal, tanpa bergantung pada segmentasi wajah eksplisit maupun sensor tambahan.

2.2. Keaslian Penelitian

Tabel 2.1. Matriks literatur review dan posisi penelitian
Tuliskan Judul Tesis di Baris Ini

No	Judul	Peneliti, Media Publikasi, dan Tahun	Tujuan Penelitian	Kesimpulan	Saran atau Kelemahan	Perbandingan
1	Multimodal Abnormal Event Detection in Public Transportation	Dimitris Tsiktsiris1, Antonios Lalas2, Minas Dasygenis3, dan Konstantinos Votis4, IEEE, 2024	Mengembangkan sistem deteksi kejadian abnormal di transportasi umum dengan pendekatan SlowFast Network berbasis data multimodal (RGB, depth, audio)	Model SlowFast multimodal mampu mendeteksi kejadian kompleks secara akurat dengan akurasi 85,1%, serta meningkatkan performa klasifikasi pada sistem pemantauan kendaraan umum	Kompleksitas tinggi dan konsumsi komputasi besar karena banyaknya jalur input yang harus diolah secara simultan.	Penelitian ini menggunakan multimodalitas untuk kejadian publik, sedangkan penelitian penulis menggunakan monomodal video untuk deteksi kantuk pengemudi, serta menambahkan MHSA untuk efisiensi dan akurasi deteksi temporal
2	PAND: Precise Action Recognition on Naturalistic Driving	Hangyue Zhao1, Yuchao Xiao2, Yanyun Zhao3. IEEE, 2022	Mengembangkan metode klasifikasi perilaku pengemudi berdasarkan video naturalistik menggunakan Swin Transformer, I3D, dan modul temporal untuk action localization	Sistem mampu meningkatkan presisi klasifikasi aksi pengemudi dengan integrasi post-processing object detection, namun akurasi F1 masih terbatas pada 29,05%.	Performa masih rendah karena keterbatasan dataset dan kompleksitas durasi aksi. Penelitian selanjutnya disarankan menggunakan arsitektur temporal yang lebih ringan dan efisien	Penelitian ini menggunakan pendekatan Transformer dan arsitektur kompleks untuk klasifikasi perilaku, sedangkan penelitian penulis mengusulkan SlowFast Network yang dikombinasikan dengan MHSA untuk mendeteksi kantuk tanpa segmentasi perilaku manual

No	Judul	Peneliti, Media Publikasi, dan Tahun	Tujuan Penelitian	Kesimpulan	Saran atau Kelemahan	Perbandingan
3	Lane Change Classification and Prediction with Action Recognition Networks	Kai Liangg, Jun Wang, dan Abhir Bhalerao. arXiv, 2024	Memprediksi dan mengklasifikasi aksi berpindah jalur kendaraan menggunakan arsitektur video seperti I3D, X3D, dan SlowFast berbasis data kamera depan	Model X3D dan SlowFast berhasil meningkatkan akurasi klasifikasi lane-change hingga 84,79% dan mencapai 98,5% jika menggunakan bounding box	Ketergantungan terhadap bounding box atau data anotasi tambahan membatasi generalisasi sistem; saran penelitian diarahkan untuk mengurangi dependensi tersebut	Penelitian ini menekankan prediksi aksi berbasis bounding box dan kamera eksternal, sedangkan penelitian penulis mengembangkan SlowFast untuk deteksi kantuk dari video wajah utuh tanpa anotasi tambahan, serta meningkatkan efisiensi melalui MHSA
4	SUST-DDD: A Real-Drive Dataset for Driver Drowsiness Detection	Esra Kavalci Yilmaz dan M. Ali Akcayol, Springer, 2024	Menyediakan dataset video berkendara nyata untuk mendeteksi kantuk pengemudi dengan ekspresi alami tanpa instruksi eksplisit	Dataset ini menghasilkan hasil tinggi dengan VGG19+LSTM (akurasi 90,53%), dan mencerminkan kondisi nyata dari kantuk pengemudi di jalan	Kualitas video bervariasi, pencahayaan tidak seragam, perangkat perekaman berbeda, dan orang yang berada dalam video tidak memperlihatkan tanda kantuk, menjadi tantangan dalam generalisasi model	Penelitian ini berfokus pada penyediaan data dan menggunakan CNN+LSTM, sedangkan penelitian penulis menggunakan dataset yang sama tetapi menerapkan arsitektur SlowFast + MHSA untuk eksplorasi pendekatan klasifikasi video yang lebih efisien dan modern
5	Relationship of Driver Head Pose Dynamics and Safety-Critical Events	Zulkarnain H. Khattak dan Asad J. Khattak. IEEE, 2024	Menganalisis keterkaitan antara perubahan arah kepala pengemudi dan kejadian keselamatan kritis menggunakan	CNN dan Hopenet berhasil mendeteksi dinamika pose kepala sebagai indikator kantuk, namun bergantung pada deteksi wajah yang presisi	Sistem sangat bergantung pada kualitas deteksi wajah; saran diarahkan untuk pendekatan yang lebih robust dalam kondisi pencahayaan buruk	Penelitian ini menggunakan analisis fitur wajah spesifik (pose kepala), sedangkan penelitian penulis menggunakan video penuh dan menghilangkan ketergantungan terhadap

No	Judul	Peneliti, Media Publikasi, dan Tahun	Tujuan Penelitian	Kesimpulan	Saran atau Kelemahan	Perbandingan
			deteksi wajah dan estimasi pose			segmentasi wajah dengan mengintegrasikan SlowFast dan MHSA untuk klasifikasi temporal
6	Driver Attention Network (DANet): A Multi-task Learning CNN	Dawein Yang, Yan Wang, Ran Wei, Jaipeng Guan, Xiaohua Huang, Wei Cai, dan Zhe Jiang, ScienceDirect, 2024	Mengembangkan sistem multitugas berbasis CNN (DANet) untuk mendeteksi berbagai kondisi wajah pengemudi seperti arah pandangan, menguap, dan mata tertutup secara simultan	Model DANet menunjukkan efisiensi tinggi dalam menggabungkan tugas-tugas visual menggunakan Dual-loss Block, dengan performa presisi tinggi untuk deteksi landmark wajah	Ketergantungan pada segmentasi wajah membuatnya kurang andal dalam kondisi nyata yang tidak terstruktur; saran untuk sistem yang lebih fleksibel tanpa segmentasi eksplisit	Penelitian ini fokus pada multitugas berbasis wajah dengan deteksi fitur eksplisit, sementara penelitian penulis memproses video utuh tanpa fitur wajah, dan mengusulkan kombinasi SlowFast + MHSA untuk deteksi kantuk yang efisien dan lebih general
7	Research on Insdurstrial Human Action Recognition based on Improved SlowFast	Zhen Lian, Ying Hu, Ranshcng Yang. IEEE, ICoIAS,	Meningkatkan akurasi deteksi aksi manusia di lingkungan industri dengan memperbaiki arsitektur SlowFast	Model SlowFast berhasil ditingkatkan dengan modul C3DS Attention dan MH-C3DS Multi-Head Attention, yang meningkatkan mAP dari 67.6% menjadi 77.4% dalam klasifikasi video industri.	Belum diuji pada sistem real-time dan hanya terbatas pada data industri, tidak mencakup konteks pengemudi atau kondisi real-life yang variatif.	Penelitian ini menambahkan attention untuk video aksi umum, sedangkan penelitian penulis berfokus pada deteksi kantuk pengemudi menggunakan MHSA untuk optimasi temporal-spasial berbasis video singkat.

2.3. Landasan Teori

Penelitian ini memerlukan landasan teori yang akan digunakan untuk mendukung penelitian ataupun membantu dalam mengembangkan system deteksi perilaku pengemudi menggunakan computer vision dengan SlowFast network dan Self-Attention.

A. Computer Vision

Computer vision adalah bidang dalam ilmu komputer yang berfokus pada bagaimana komputer dapat "melihat" dan memahami informasi visual dari gambar atau video digital. Menurut Yasunari Matsuzaka dan Ryu Yashiro dalam artikel mereka di jurnal "AI-Based Computer Vision Techniques and Expert Systems," computer vision bertujuan untuk memberikan fungsi penglihatan pada komputer, mirip dengan cara kerja mata manusia. Bidang ini telah berkembang pesat dengan munculnya pembelajaran mendalam (deep learning), yang secara signifikan meningkatkan akurasi pengenalan gambar dan deteksi objek.

Computer vision adalah bidang ilmu komputer yang bertujuan untuk memungkinkan komputer untuk memahami dan menafsirkan dunia visual, mirip dengan cara manusia melihat dan memahami lingkungan mereka. Computer vision melibatkan akuisisi, pemrosesan, dan analisis data gambar dan video untuk mengekstrak informasi yang berarti

Sejarah computer vision dimulai pada tahun 1960-an ketika ilmuwan mulai mengeksplorasi cara agar komputer bisa melihat dan memahami gambar. Proyek ini dianggap sebagai kelahiran resmi computer vision sebagai bidang

ilmiah. Pada tahun 1982, David Marr, seorang ahli neurobiologi, memperkenalkan kerangka kerja representasional untuk visi, yang mencakup konsep-konsep seperti "primal sketch" dan model 3D. Karya Marr ini penting dalam pemahaman hierarkis tentang bagaimana visi bekerja. Atallah, S. (2020).

Pada tahun 1980-an, Kunihiko Fukushima mengembangkan Neocognitron, jaringan saraf buatan yang bisa mengenali pola. Ini merupakan cikal bakal dari jaringan saraf convolutional (CNN) modern. Kemudian, Yann LeCun memperkenalkan LeNet-5 pada tahun 1989, yang merupakan salah satu arsitektur CNN pertama yang digunakan secara luas, terutama untuk pengenalan karakter.

Pada tahun 2010-an, kemajuan signifikan terjadi dengan penerapan deep learning dalam computer vision. Teknik ini mengandalkan arsitektur CNN yang lebih dalam dan dataset yang lebih besar, seperti ImageNet, yang memungkinkan komputer untuk mencapai tingkat akurasi yang sangat tinggi dalam berbagai tugas pengenalan gambar dan video

B. Metode Computer Vision

Terdapat beberapa metode atau tahapan yang digunakan dalam computer vision seperti, image processing yang didalamnya terdapat image filtering dan edge detection, lalu Feature Extraction yang terdiri dari SIFT, SURF, FOG dan beberapa tahapan lainnya yang akan dijelaskan lebih lengkap berikut ini :

1. Image Processing
    a. Thresholding yaitu Teknik untuk memisahkan objek dari latar belakang gambar berdasarkan nilai piksel.

b. Filtering melibatkan penggunaan filter untuk menghilangkan noise dan meningkatkan kualitas gambar.

c. Edge Detection yaitu untuk Mengidentifikasi tepi objek dalam gambar menggunakan algoritma seperti Canny atau Sobel.

2. Feature Extraction

a. SIFT (Scale-Invariant Feature Transform) yaitu untuk Mengidentifikasi dan menggambarkan titik-titik kunci dalam gambar. Metode ini sangat stabil terhadap perubahan skala dan rotasi, serta tahan terhadap variasi pencahayaan. SIFT melibatkan deteksi titik kunci, deskripsi fitur, dan pencocokan fitur antara gambar.

b. SURF (Speeded Up Robust Features) adalah alternatif yang lebih cepat dari SIFT, yang menggunakan pendekatan integral image dan deteksi fitur berbasis Haar wavelet untuk meningkatkan kecepatan dan efisiensi. SURF tetap mempertahankan robust terhadap perubahan skala dan rotasi.

c. HOG (Histogram of Oriented Gradients) yaitu untuk Menggambarkan penampilan dan bentuk objek dalam gambar. Algoritma ini menghitung histogram gradien arah dalam blok-blok kecil dari gambar, menghasilkan representasi fitur yang kuat terhadap variasi lokal dalam pencahayaan dan pose objek.

3. Object Recognition

    a. Template matching adalah metode pencocokan pola yang mencari bagian dari gambar yang mirip dengan template yang telah ditentukan sebelumnya

    b. Bag of Words (BoW) adalah representasi gambar yang menggunakan fitur lokal seperti SIFT untuk membuat representasi visual dari gambar. Metode ini mengelompokkan fitur-fitur lokal ke dalam "kata-kata" visual, yang kemudian digunakan untuk tugas-tugas klasifikasi dan pengenalan objek.

4. Deep Learning Techniques

    a. Convolutional Neural Networks (CNN) menggunakan lapisan konvolusi untuk mendeteksi fitur dalam gambar, diikuti oleh lapisan pooling untuk mengurangi dimensi fitur, dan lapisan fully connected untuk klasifikasi. CNN sangat efektif untuk tugas-tugas seperti pengenalan objek, klasifikasi gambar, dan segmentasi gambar.

    b. Recurrent Neural Networks (RNNs) dan Long Short-Term Memory (LSTM) digunakan untuk analisis video dan pengenalan pola temporal dalam data sekuensial. RNN dan LSTM dapat menangani urutan data yang panjang dan kompleks, seperti gerakan dalam video atau urutan aksi.

    c. Generative Adversarial Networks (GANs) digunakan untuk menghasilkan gambar baru yang tampak realistis dengan belajar dari dataset gambar. GAN terdiri dari dua jaringan saraf, generator dan

discriminator, yang berinteraksi dalam proses pelatihan adversarial untuk meningkatkan kualitas gambar yang dihasilkan.

Pada penelitian ini peneliti akan mmenggunakan SlowFast dan Self-attention pada computer vision untuk membantu membuat system deteksi perilaku pengemudi. Berikut penjelasan SlowFast dan juga Self-attention

C. SlowFast Network

SlowFast Network adalah model dua alur yang canggih yang dikembangkan untuk meningkatkan pengenalan aksi dalam video dengan menangani gerakan lambat dan cepat secara efektif. Jaringan ini dirancang untuk menangkap berbagai informasi temporal dengan memanfaatkan dua cabang yang berjalan paralel. Slow Branch memproses video dengan kecepatan frame rendah untuk fokus pada informasi yang berkembang secara perlahan, seperti perubahan konteks dalam sebuah adegan dan gerakan bertahap. Sementara itu, Fast Branch menangani video dengan kecepatan frame tinggi untuk menangkap gerakan cepat, memungkinkan model mengenali dinamika temporal yang cepat dan detail. Arsitektur ini bekerja dengan menjalankan cabang lambat dan cepat secara bersamaan, masing-masing memproses frame video pada tingkat yang berbeda. Slow branch memproses lebih sedikit input frame, memberikan pemahaman temporal jangka panjang, sedangkan Fast branch memproses lebih banyak frame dengan kecepatan lebih tinggi, menghasilkan pola gerakan yang lebih detail. Christoph Feichtenhofer (2019).

Salah satu aspek penting dari SlowFast Network adalah fusion mechanismnya, di mana fitur gerakan dari cabang fast diintegrasikan ke dalam

cabang slow. Kombinasi ini memungkinkan model memanfaatkan fitur gerakan yang mendetail sekaligus konteks adegan secara keseluruhan, sehingga meningkatkan kinerja action recognition

1. Slow Pathway

Slow Pathway mengambil input frame video dengan sampling rate rendah sehingga informasi temporal yang lebih lama dapat tertangkap. Setiap frame yang diterima diproses dengan konvolusi 3D (3D convolution), yang digunakan untuk mengekstraksi fitur dari dimensi spasial dan temporal secara bersamaan. Arsitektur Slow Pathway biasanya menggunakan ResNet-3D (seperti ResNet-50 atau ResNet-101) sebagai backbone, dengan kernel 3D untuk menangkap pola spasial dan temporal secara serentak. Rumus yang digunakan untuk menentukan jumlah frame yang diproses dalam jalur lambat adalah :

$$T_s = \frac{T}{\tau}$$

Di sini, T adalah total frame dalam video asli, sedangkan τ adalah faktor sampling, yang biasanya bernilai 4 atau 8. Ini berarti bahwa hanya setiap 4 atau 8 frame yang dipilih untuk diproses dalam jalur lambat, sehingga menghasilkan Ts, yaitu jumlah frame yang akhirnya diproses dalam jalur ini. Untuk memproses frame dalam jalur lambat, digunakan convolution 3D yang bertujuan untuk menangkap detail spasial secara mendalam. Biasanya, model yang digunakan adalah ResNet-3D atau varian lain dari model convolutional 3D yang memiliki kemampuan menangani input tiga dimensi

(spasial dan temporal). Proses konvolusi pada jalur lambat ini mengikuti rumus:

$$y_t = \sum_{i=0}^{N} W_i * X_{t-i}$$

Di sini, $Xt-i$ adalah input frame pada waktu $t-i$, dan Wi adalah filter konvolusi pada waktu i. Hasil keluaran dari proses ini adalah yt , yaitu feature map yang terbentuk pada waktu t. Operasi ini memungkinkan model untuk menangkap informasi spasial yang lebih akurat dari frame yang diproses, mengingat jalur lambat bertujuan untuk mengolah video secara mendalam pada aspek spasial, meskipun dengan frame rate yang lebih rendah. Operasi ini bertujuan untuk menangkap fitur-fitur yang relevan secara spasial dan temporal dari gerakan lambat, memungkinkan model mengenali perubahan jangka panjang dengan lebih efektif

2. Fast Pathway

Fast Pathway menerima input frame dengan sampling rate tinggi, memungkinkan deteksi pola temporal yang lebih cepat dan rinci. Konvolusi 3D juga digunakan dalam jalur ini, tetapi dengan ukuran kernel yang lebih kecil dan kedalaman jaringan yang lebih rendah, sehingga dapat memproses frame dengan lebih cepat. dalam jalur ini, semua frame video diproses, sehingga jumlah frame yang diproses (Tf) sama dengan jumlah frame asli video (T), yang berarti frame rate-nya tetap sama dengan frame rate asli. Hal ini memungkinkan jalur cepat untuk menangkap perubahan temporal yang lebih mendetail dalam video, seperti gerakan tangan yang cepat. Proses

konvolusi 3D pada jalur cepat bekerja dengan cara yang sama seperti pada jalur lambat, di mana operasi ini menggunakan filter konvolusi untuk mengekstraksi fitur dari video dalam ruang tiga dimensi (spasial dan temporal). Namun, karena resolusi yang lebih rendah, jumlah parameter yang dibutuhkan lebih sedikit, sehingga konvolusi di jalur cepat menjadi lebih ringan dan lebih efisien dalam hal komputasi. Hal ini menjadikan jalur cepat cocok untuk menangani aspek temporal yang cepat dalam video, tanpa memerlukan sumber daya komputasi yang besar

3. Fusion Mechanism

Setelah kedua pathway memproses input video secara terpisah, terdapat proses fusi untuk menggabungkan informasi dari kedua pathway tersebut. Fusi dilakukan dengan cara mengintegrasikan fitur dari Fast Pathway ke dalam Slow Pathway, sehingga Slow Pathway dapat memperkaya pemahaman konteksnya dengan informasi yang lebih cepat dari Fast Pathway. Penggabungan antara jalur lambat dan cepat dalam SlowFast Network dilakukan melalui koneksi lateral (lateral connection), yang memungkinkan kedua jalur berbagi informasi sehingga dapat saling melengkapi. Penggabungan ini adalah inti dari arsitektur SlowFast, memastikan bahwa informasi spasial yang ditangkap oleh jalur lambat dapat diperkaya dengan informasi temporal yang lebih dinamis dari jalur cepat. Proses penggabungan ini dilakukan dalam beberapa langkah utama, yang masing-masing dirancang untuk menyelaraskan dan mengintegrasikan fitur dari kedua jalur. Langkah pertama adalah menyamakan dimensi fitur dari

kedua jalur dengan menggunakan konvolusi 1x1. Pada tahap ini, fitur dari jalur cepat (*Ffast*) diproyeksikan ke resolusi yang sama dengan fitur dari jalur lambat, sehingga dimensi spasial dan jumlah channel menjadi serasi. Proses ini dapat dirumuskan sebagai:

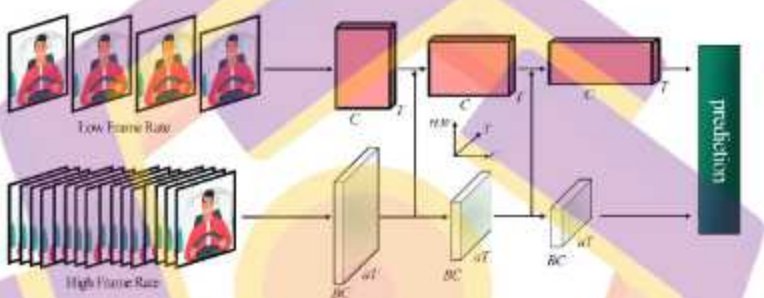
$$F_{proj} = W_{1x1} * F_{fast}$$

Dalam rumus ini, *Ffast* adalah fitur dari jalur cepat yang diproyeksikan menggunakan filter konvolusi 1x1 ($W_{1\times1}$), menghasilkan *Fproj*, yaitu fitur yang telah diselaraskan dengan jalur lambat. Proyeksi ini penting karena fitur dari jalur cepat biasanya memiliki resolusi yang lebih rendah dan jumlah channel yang berbeda, sehingga perlu disesuaikan agar dapat digabungkan dengan jalur lambat. Setelah fitur dari jalur cepat diproyeksikan, langkah berikutnya adalah penjumlahan lateral. Pada tahap ini, fitur yang diproyeksikan dari jalur cepat (*Fproj*) ditambahkan secara langsung ke fitur dari jalur lambat (*Fslow*). Proses ini dapat dirumuskan sebagai:

$$F_{slow}^{(t)} = F_{slow}^{(t)} + F_{proj}$$

Di sini, $F_{slow}^{(t)}$ adalah fitur dari jalur lambat pada waktu *t*, setelah digabungkan dengan fitur yang diproyeksikan dari jalur cepat. Penjumlahan lateral ini memungkinkan jalur lambat menerima informasi temporal yang lebih mendetail dari jalur cepat, sehingga memperkaya representasi fitur keseluruhan dengan baik. Setelah penggabungan fitur antara kedua jalur selesai, model melanjutkan dengan operasi konvolusi dan pooling tambahan untuk mengurangi dimensi dan memperkaya representasi fitur lebih lanjut.

Langkah ini penting agar model dapat menghasilkan fitur yang lebih ringkas dan bermakna, yang kemudian digunakan untuk tugas akhir seperti klasifikasi video atau deteksi objek. Proses ini memastikan bahwa fitur gabungan tidak hanya kuat secara spasial tetapi juga tajam dalam menangkap dinamika temporal, menjadikan arsitektur SlowFast sangat efektif untuk berbagai tugas pengenalan video.



Gambar 2.1. Arsitektur Slowfast Network

D. Self-attention

Self-attention adalah metode yang memungkinkan model untuk menghitung korelasi antara elemen-elemen dalam sebuah urutan data, memberikan bobot yang lebih besar pada elemen-elemen yang lebih relevan terhadap elemen target. Proses self-attention melibatkan tiga komponen utama yang disebut *query* (Q), *key* (K), dan *value* (V), yang merupakan representasi vektor dari elemen-elemen dalam urutan input. Setiap elemen dalam urutan bertindak sebagai pusat perhatian, di mana korelasinya dengan elemen-elemen lain dalam urutan dihitung berdasarkan kesamaan antara query dan key. Tingkat kesamaan ini kemudian digunakan untuk memberi bobot pada nilai (value) masing-

masing elemen, menghasilkan output yang memperkuat informasi kontekstual yang lebih relevan dalam urutan tersebut. Sangwoo Cho (2020). Rumus utama dari self-attention adalah:

$$Attention(Q, K, V) = softmax\left(\frac{QK^T}{\sqrt{d_k}}\right)V$$

Dalam persamaan ini, QK^T adalah hasil perkalian dot product antara query dan key, yang kemudian diskalakan dengan $\frac{1}{\sqrt{dk}}$ untuk mengurangi risiko nilai yang terlalu besar sehingga bisa menghambat stabilitas perhitungan. Fungsi softmax digunakan untuk mengubah hasil perkalian ini menjadi distribusi probabilitas yang menunjukkan bobot relatif antar-elemen dalam urutan. Setelah bobot ini diterapkan pada elemen value, model dapat mengidentifikasi informasi mana yang paling relevan untuk setiap elemen target dalam urutan.

Dalam implementasinya, self-attention sering menggunakan *multi-head attention*, di mana perhitungan attention dilakukan beberapa kali secara paralel dengan berbagai parameter yang berbeda. Teknik ini memungkinkan model untuk menangkap berbagai pola korelasi di antara elemen-elemen dalam urutan, meningkatkan kemampuan model untuk memahami konteks dengan lebih kaya. Output dari berbagai head ini digabungkan dan diproyeksikan kembali ke dimensi yang sama dengan input aslinya, memastikan bahwa informasi konteks tetap sesuai dengan skala input.

BAB III

METODE PENELITIAN

3.1. Jenis, Sifat, dan Pendekatan Penelitian

Penelitian ini termasuk dalam kategori penelitian eksperimental yang menggunakan pendekatan kuantitatif. Penelitian ini dilakukan dengan mengembangkan dan menguji kinerja model deep learning untuk mendeteksi perilaku kantuk pengemudi berdasarkan video berdurasi pendek yang diperoleh dari kamera di dalam kendaraan. Dari segi sifatnya, penelitian ini bersifat deskriptif dan analitik, di mana peneliti berusaha mendeskripsikan kondisi perilaku pengemudi dalam keadaan kantuk dan tidak kantuk, serta menganalisis hasil klasifikasi dari sistem yang dibangun. Tujuan utamanya adalah untuk mengetahui apakah penambahan mekanisme Multi-Head Self Attention (MHSA) pada arsitektur SlowFast Network dapat meningkatkan akurasi dalam mengenali perilaku kantuk secara efisien dan tepat. Secara garis besar, pendekatan penelitian ini meliputi empat tahap yaitu pengumpulan data, praproses data, pengembangan model, serta pelatihan dan evaluasi model.

1. Pengumpulan Data

   Dataset yang digunakan dalam penelitian ini meliputi SUST Driver Drowsiness Dataset sebagai data utama, serta NITYMED Dataset sebagai data tambahan untuk menguji kemampuan generalisasi model pada kondisi pencahayaan malam hari. Keduanya berisi video pengemudi nyata dalam kondisi kantuk dan tidak kantuk.

2. Pre-processing

   Tahap ini meliputi resizing frame video ke resolusi 112x112 piksel, pembentukan klip temporal menggunakan metode sliding window, serta normalisasi nilai piksel dan augmentasi data untuk meningkatkan variasi input bagi arsitektur jaringan.

3. Pengembangan Model

   Model dikembangkan menggunakan SlowFast Network dengan konfigurasi dua jalur (slow dan fast) untuk menangkap gerakan lambat dan cepat. Pada bagian penggabungan informasi, ditambahkan Multi-Head Self Attention (MHSA) untuk meningkatkan representasi fitur dan fokus perhatian model.

4. Pelatihan dan Evaluasi Mode

   Dataset dibagi menjadi tiga bagian proporsional untuk pelatihan (training), validasi, dan pengujian (testing). Evaluasi kinerja dilakukan secara komprehensif menggunakan metrik standar klasifikasi, yaitu Akurasi, Presisi, Recall, dan F1-Score.

3.2. Metode Pengumpulan Data

Pengumpulan data dalam penelitian ini dilakukan dengan memanfaatkan data sekunder dari dua sumber dataset publik yang memiliki karakteristik saling melengkapi, yaitu SUST Driver Drowsiness Dataset (SUST-DDD) dan NITYMED Dataset. Penggunaan dua sumber data ini bertujuan untuk memastikan model memiliki kemampuan generalisasi yang baik terhadap berbagai kondisi pencahayaan dan karakteristik subjek. Dataset utama yang digunakan adalah SUST-DDD yang dikembangkan oleh Yılmaz dan Akcayol (2024). Dataset ini

memuat 2.074 video rekaman perilaku pengemudi yang terbagi dalam dua kondisi, yaitu kantuk (drowsy) dan tidak kantuk (not drowsy). Seluruh video direkam secara langsung oleh 19 partisipan (terdiri dari 3 perempuan dan 16 laki-laki) menggunakan kamera ponsel di kendaraan masing-masing tanpa adanya skenario yang diatur sebelumnya. Kondisi perekaman ini menghasilkan ekspresi yang bersifat alami (in-the-wild) dan merepresentasikan kondisi pengemudi pada pencahayaan siang hari atau standar kabin. Selanjutnya, sebagai data uji tambahan untuk mengevaluasi ketahanan (robustness) model pada kondisi lingkungan yang lebih menantang, penelitian ini menggunakan NITYMED Dataset. Dataset ini berisi 130 video yang direkam dalam kondisi kendaraan bergerak di malam hari (nighttime conditions), di mana deteksi kantuk menjadi lebih krusial namun sulit akibat minimnya pencahayaan. Integrasi kedua dataset ini memungkinkan pengembangan sistem deteksi yang tidak hanya akurat pada kondisi ideal sebagaimana direpresentasikan oleh SUST, tetapi juga teruji handal dalam menangani tantangan visual seperti pencahayaan rendah dan ekspresi kantuk yang eksplisit pada dataset NITYMED. Seluruh data video dari kedua sumber tersebut kemudian direpresentasikan dalam bentuk sekuens frame untuk diproses lebih lanjut dalam pemodelan Deep Learning.

3.3. Metode Analisis Data

Metode analisis data dalam penelitian ini bertujuan untuk mengevaluasi kinerja sistem deteksi perilaku kantuk pengemudi berbasis video, yang dikembangkan menggunakan arsitektur SlowFast Network dan diperkuat dengan Multi-Head Self Attention (MHSA). Analisis dilakukan terhadap hasil klasifikasi

video ke dalam dua kategori, yaitu drowsy (kantuk) dan not drowsy (normal). Tahap awal analisis dimulai setelah model selesai dilatih. Dataset pengujian yang telah melalui proses pre-processing akan diprediksi oleh model, dan hasil prediksi akan dibandingkan dengan label ground-truth. Selanjutnya, evaluasi dilakukan dengan menggunakan metode confusion matrix, yang mencakup beberapa metrik utama, yaitu:

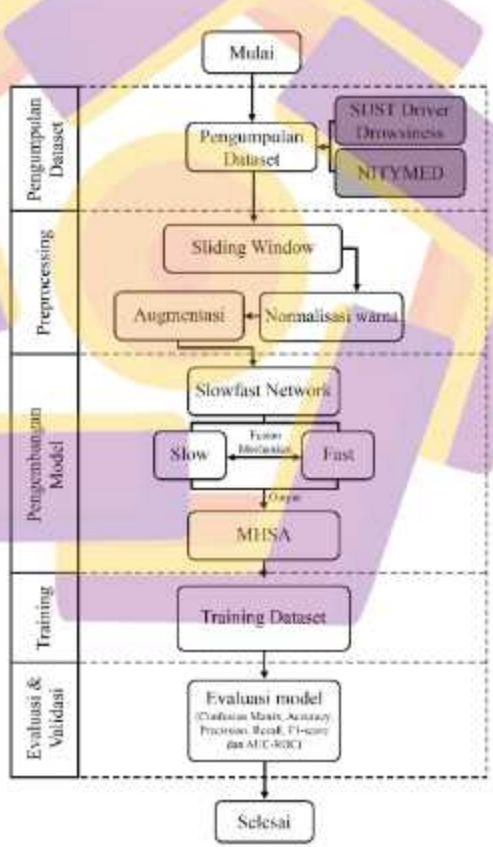
- Akurasi (Accuracy) untuk Mengukur proporsi prediksi yang benar terhadap seluruh data.
- Presisi (Precision) untuk Mengukur proporsi prediksi positif yang benar-benar relevan (true positive dibandingkan dengan semua hasil positif).
- Recall (Sensitivity) untuk Mengukur proporsi data positif yang berhasil dikenali oleh model.
- F1-Score yaitu Merupakan harmonic mean dari precision dan recall, berguna dalam kondisi data tidak seimbang.

Semua perhitungan dilakukan menggunakan library Python seperti Scikit-learn untuk metrik evaluasi, serta PyTorch untuk proses pelatihan dan prediksi model. Model dijalankan pada lingkungan komputasi dengan spesifikasi GPU untuk mempercepat proses pelatihan dan evaluasi. Analisis ini bertujuan untuk mengetahui sejauh mana integrasi MHSA dalam arsitektur SlowFast dapat meningkatkan performa model dalam mendeteksi perilaku kantuk berbasis data video utuh tanpa memerlukan segmentasi wajah atau input multimodal tambahan.

3.4. Alur Penelitian

Penelitian ini terdiri dari tiga konfigurasi eksperimen utama berdasarkan kombinasi arsitektur model (SlowFast Network, dengan/ tanpa Multi-Head Self Attention) dan jenis dataset yang digunakan (SUST-DDD saja atau SUST-DDD digabung dengan NITYMED). Setiap konfigurasi memiliki alur proses yang mirip secara umum, namun memiliki perbedaan pada tahap input data dan struktur model.

3.4.1. Alur Penelitian SlowFast + MHSA



Gambar 3.1. Alur Penelitian SlowFast + MHSA

Alur penelitian ini terdiri dari lima tahap utama, yang disusun secara sistematis dari perumusan permasalahan hingga evaluasi model. Langkah-langkah ini divisualisasikan dalam bentuk diagram alir penelitian yang menggambarkan urutan proses dan komponen utama yang digunakan dalam sistem deteksi kantuk pengemudi berbasis video. Berikut penjelasan tiap tahap:

A. Pengumpulan Data

Dataset yang digunakan dalam penelitian ini yaitu penggabungan dua dataset, yaitu SUST Driver Drowsiness Dataset (SUST-DDD) dan NITYMED Nighttime Drowsiness Dataset. Kombinasi dua dataset ini dirancang untuk memperkaya keberagaman data dari sisi pencahayaan, ekspresi kantuk, serta demografi peserta. Tujuannya adalah untuk meningkatkan kemampuan generalisasi model terhadap berbagai kondisi dunia nyata.

1. SUST-DDD (Sivas University of Science and Technology – Driver Drowsiness Dataset): Dataset ini terdiri dari 2.074 video berdurasi 10 detik, dikumpulkan dari 19 partisipan (3 perempuan dan 16 laki-laki), yang merekam dirinya sendiri menggunakan kamera ponsel saat berkendara. Tidak ada instruksi atau skenario tertentu selama proses perekaman, sehingga ekspresi yang terekam bersifat alami. Setiap partisipan memiliki video dengan penamaan d_X.mp4 untuk kondisi drowsy, dan n_X.mp4 untuk kondisi not drowsy. Namun, dalam praktiknya, berdasarkan observasi dan uji coba model, diketahui bahwa ekspresi drowsy pada dataset ini terlalu halus dan tidak dapat dipastikan kebenarannya secara visual maupun fisiologis. Oleh karena itu, dalam konteks penelitian ini,

seluruh video dari SUST-DDD diklasifikasikan ulang sebagai kategori not_drowsy. Hal ini dilakukan untuk menjaga validitas klasifikasi biner yang lebih tegas, serta mencegah bias label dari data yang ambigu.

2. NITYMED (Nighttime Driver Drowsiness Dataset – Patras, Greece): Dataset ini terdiri dari 130 video, direkam di dalam mobil sungguhan saat malam hari, dengan pencahayaan alami ditambah sedikit cahaya interior. Partisipannya berjumlah 21 orang (11 laki-laki dan 10 perempuan) dengan fitur visual yang beragam. Terdapat dua jenis video dalam dataset ini yaitu Yawning 107 video dengan durasi 15–25 detik dan pengemudi yang menguap sebanyak tiga kali, lalu Microsleep 21 video dengan durasi 2 menit dan pengemudi mengalami microsleep saat berbicara atau melihat sekitar. Semua video dalam format .mp4, mute, 25 fps, dan tersedia dalam resolusi HDTV720 dan Full HD. Dalam penelitian ini, resolusi diubah menjadi 112×112 piksel dan dipotong menjadi klip 2 detik untuk konsistensi input.

Karena video pada dataset ini merekam ekspresi kantuk secara eksplisit, maka seluruh video dari NITYMED digunakan sebagai kategori drowsy dalam sistem klasifikasi yang dikembangkan. Penelitian ini diawali dengan penggabungan dan relabeling dataset, di mana seluruh video dari SUST-DDD dikategorikan sebagai not_drowsy karena sifat ekspresinya yang natural dan tidak terstruktur, sedangkan semua video dari NITYMED diklasifikasikan sebagai drowsy karena menampilkan tanda kantuk yang eksplisit. Gabungan kedua dataset ini membentuk sistem klasifikasi biner yang lebih seimbang dan

jelas. Tahap preprocessing mencakup segmentasi temporal dengan memotong video menjadi klip berdurasi 2 detik menggunakan metode sliding window, dilanjutkan dengan augmentasi serta normalisasi warna untuk memperkaya variasi data dan meningkatkan konsistensi input. Selanjutnya, pengembangan model dilakukan menggunakan arsitektur SlowFast, di mana output dari jalur Slow dan Fast digabungkan dan diproses oleh lapisan Multi-Head Self Attention (MHSA) untuk menyoroti fitur spasio-temporal yang relevan sebelum diteruskan ke layer klasifikasi sigmoid. Data kemudian dibagi menjadi 70% untuk pelatihan, 15% untuk validasi, dan 15% untuk pengujian, dan model dievaluasi menggunakan metrik akurasi, presisi, recall, F1-score, AUC-ROC, serta confusion matrix. Tujuan dari eksperimen ini adalah untuk menguji efektivitas kombinasi antara arsitektur SlowFast + MHSA dengan data gabungan yang mencakup ekspresi kantuk eksplisit dari NITYMED dan ekspresi netral dari SUST-DDD, sehingga model diharapkan mampu mengenali pola kantuk secara lebih akurat dan general terhadap berbagai kondisi dunia nyata.

B. Preprocessing

Untuk meningkatkan kualitas data sebelum pelatihan model, dilakukan beberapa tahap preprocessing, yaitu:

1. Sliding Window

   Setiap video berdurasi 10 detik dipotong menjadi klip pendek berdurasi sekitar 2 detik menggunakan metode sliding window dengan panjang 60

frame dan overlap 50%. Segmentasi ini bertujuan untuk mempertahankan kontinuitas pola temporal dan meningkatkan jumlah sampel pelatihan.

2. Normalisasi Citra

Setiap frame yang diambil dari video terlebih dahulu dikonversi dari format warna BGR (standar default pada OpenCV) menjadi format RGB yang umum digunakan dalam pemrosesan citra. Kemudian, ukuran gambar diubah menjadi resolusi tetap 112×112 piksel agar sesuai dengan kebutuhan input model. Pengubahan ukuran ini dilakukan melalui proses resampling piksel, sehingga seluruh gambar tetap dipertahankan namun dengan kemungkinan perubahan rasio aspek (distorsi kecil) apabila rasio panjang dan lebar aslinya tidak 1:1. Setelah itu, nilai warna pada setiap piksel, yang semula berada dalam rentang 0 hingga 255, disesuaikan dengan cara dibagi 255. Hasilnya adalah nilai yang berada pada skala antara 0 dan 1, yang dikenal sebagai normalisasi. Nilai yang telah dinormalisasi ini kemudian dikonversi ke tipe data float (desimal) agar dapat diproses lebih efisien oleh model deep learning. Proses ini bertujuan untuk menyamakan skala input, mengurangi variabilitas data, dan mempercepat proses pelatihan model.

C. Pengembangan Model

Model yang dikembangkan dalam penelitian ini menggunakan arsitektur SlowFast Network yang dimodifikasi dan dikombinasikan dengan mekanisme Self-Attention. Arsitektur ini dirancang untuk menangkap informasi temporal

dan spasial dari video secara lebih efektif dalam konteks deteksi perilaku kantuk pengemudi.

1. SlowFast Network

   Model yang dikembangkan dalam penelitian ini menggunakan arsitektur SlowFast Network yang dimodifikasi dan dikombinasikan dengan mekanisme Self-Attention. Arsitektur ini dirancang untuk menangkap informasi temporal dan spasial dari video secara lebih efektif dalam konteks deteksi perilaku kantuk pengemudi.

   - Fast Pathway menerima 30 frame berurutan dari video berdurasi pendek. Jalur ini dirancang untuk mendeteksi perubahan cepat seperti kedipan mata, gerakan kepala mendadak, atau aktivitas mikro lainnya dalam waktu singkat
   - Slow Pathway menerima 8 frame yang dipilih secara merata dari 30 frame tersebut. Jalur ini fokus pada perubahan lambat dan konteks global, seperti arah pandangan, penurunan aktivitas, atau ekspresi umum wajah.

   Setiap jalur kemudian memproses inputnya menggunakan layer Conv3D, yang bertugas mengekstraksi pola spasial (bentuk) dan temporal (gerakan) secara bersamaan. Tensor input pada setiap jalur direpresentasikan sebagai:

   $$X \in R^{TxHxWxC}$$

   Dengan:

   - T: jumlah frame,
   - H ×W: resolusi frame,

- C: jumlah channel warna (RGB = 3).
- R: himpunan bilangan real, menunjukkan bahwa nilai piksel pada tensor X berada dalam domain bilangan riil (floating point), bukan bilangan bulat.

Operasi Conv3D dilakukan menggunakan kernel ukuran 3×3×3, yang dijalankan untuk seluruh channel:

$$\text{Output}(t,h,w,c)=\sum_{i,j,k}\sum_{c'} X(t+i,h+j,w+k,c') \cdot W(i,j,k,c',c)$$

- t: indeks waktu (frame ke-t dalam klip video).
- h: indeks tinggi (baris pixel pada frame).
- w: indeks lebar (kolom pixel pada frame).
- c: channel warna input (misalnya, R, G, atau B).
- i, j, k: offset/indeks lokal dari kernel 3D pada dimensi waktu (i), tinggi (j), dan lebar (k).
- c': channel warna dari input X saat dilakukan konvolusi untuk menghasilkan channel output c

Hasil dari Conv3D kemudian diringkas menggunakan Global Average Pooling (GAP) 3D untuk menghasilkan representasi vector:

$$f_c \frac{1}{T \cdot H \cdot W} \sum_{t=1}^{T}\sum_{h=1}^{H}\sum_{w=1}^{W} Output(t,h,w,c)$$

Masing-masing jalur akan menghasilkan vektor berdimensi tetap, misalnya 1×128. Kedua vektor dari Slow dan Fast Pathway kemudian digabungkan (concatenated) menjadi:

$$\text{FusedFeature} = [\text{Ffast} \parallel \text{Fslow}] \in R^{1x256}$$

Vektor ini menjadi representasi akhir dari video, yang memuat informasi komprehensif dari kedua perspektif: cepat dan lambat. Representasi ini diteruskan ke tahap selanjutnya, yaitu Self-Attention dan klasifikasi

2. Multi-Head Self-Attention

Setelah dua jalur fitur digabungkan, model menggunakan mekanisme Self-Attention untuk meningkatkan relevansi fitur sebelum klasifikasi akhir. Penelitian ini menggunakan layer Multi-Head Self-Attention (MHSA), yang memungkinkan model menangkap hubungan antar fitur dari berbagai perspektif secara paralel. Pertama, vektor fitur gabungan $X \in R^{1\times256}$ diproyeksikan menjadi Query (Q), Key (K), dan Value (V) menggunakan bobot transformasi:

$$Q = XW^Q, K = XW^K, V = XW^V$$

Keterangan:

- X: Vektor input (hasil penggabungan fitur Slow dan Fast) berukuran 1×2561
- W^Q, W^K, W^V:Matriks bobot transformasi (trainable parameters) untuk membentuk Query, Key, dan Value
- Q, K, V: Vektor hasil proyeksi input ke ruang Query, Key, dan Value

Kemudian, perhatian dihitung menggunakan fungsi Softmax dari hasil dot-product antara Q dan K, dibagi dengan akar dari dimensi key (dk):

$$\text{Attention(Q,K,V)} = \text{softmax}\left(\frac{QK^T}{\sqrt{dk}}\right)V$$

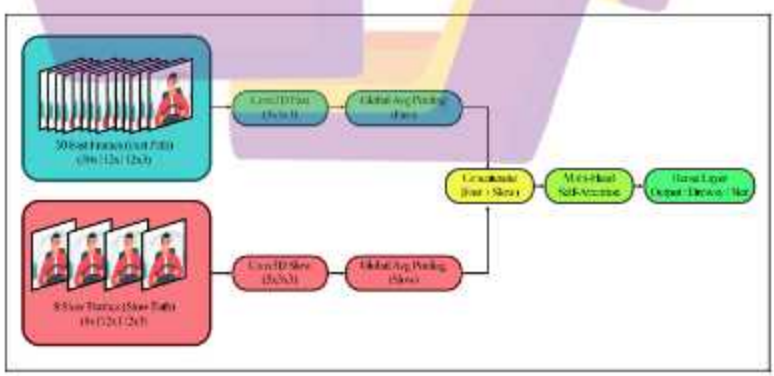
Mekanisme ini menghasilkan vektor fitur baru yang memprioritaskan bagian-bagian penting dari input gabungan, seperti ekspresi wajah atau gerakan kepala yang relevan terhadap deteksi kantuk. Setelah perhatian dihitung, vektor hasil Self-Attention diteruskan ke dense layer atau layer klasifikasi. Fungsi aktivasi sigmoid digunakan untuk menghasilkan nilai prediksi biner:

$$\hat{y}=\sigma(W_o \cdot AttentionOutput + b$$

Jika $\hat{y}$ >0.5, maka model memprediksi drowsy; jika tidak, maka not drowsy. Gambar 2 menyajikan arsitektur model yang diusulkan dalam penelitian ini. Struktur dua jalur pemrosesan, yaitu Slow Pathway dan Fast Pathway, diadaptasi dari penelitian yang dilakukan oleh Feichtenhofer et al. (2019)

Gambar 3.2. Arsitektur SlowFast dan Multi-Head Self-Attention

terkait pengenalan video menggunakan SlowFast Networks. Penelitian ini melakukan modifikasi dengan mengintegrasikan mekanisme Self-Attention ringan setelah proses penggabungan fitur dari kedua jalur, dengan tujuan



untuk meningkatkan relevansi representasi fitur sebelum memasuki tahap klasifikasi akhir.

D. Pelatihan Model

Model SlowFast yang telah dikembangkan dilatih menggunakan dataset video yang sebelumnya telah melalui tahap segmentasi dan normalisasi. Dataset dibagi menjadi tiga bagian, yaitu 70% untuk pelatihan (training), 15% untuk validasi (validation), dan 15% untuk pengujian (testing). Dalam arsitektur SlowFast, setiap video klip dipecah menjadi dua jalur input, yakni Fast Pathway yang terdiri dari 30 frame berurutan, dan Slow Pathway yang terdiri dari 8 frame yang diambil secara merata dari klip yang sama. Kedua jalur ini diproses secara paralel dalam jaringan untuk menangkap dinamika temporal dari video.

Selama proses pelatihan, digunakan pendekatan berbasis custom data generator dengan memanfaatkan tf.keras.utils.Sequence, yang memungkinkan pemuatan data secara bertahap dalam bentuk batch. Hal ini bertujuan untuk menghindari beban memori berlebih selama training. Model dilatih menggunakan algoritma optimasi Adam dan fungsi loss binary cross-entropy, yang sesuai untuk tugas klasifikasi biner antara kondisi kantuk dan tidak kantuk. Setiap batch terdiri dari 4 klip video dengan resolusi 112×112 piksel, dan proses pelatihan dilakukan selama 50 epoch. Tujuan utama dari pelatihan ini adalah agar model dapat mempelajari pola perilaku pengemudi berdasarkan informasi temporal dan spasial dalam video, serta mampu melakukan generalisasi dengan baik terhadap data baru.

E. Evaluasi Model

Evaluasi model dilakukan untuk mengukur performa sistem dalam mengklasifikasikan kondisi pengemudi berdasarkan data video menjadi dua kelas, yaitu kantuk (drowsy) dan tidak kantuk (not drowsy). Proses evaluasi dilakukan menggunakan data pengujian (test set) yang sebelumnya telah dipisahkan secara eksplisit dan tidak digunakan selama proses pelatihan maupun validasi, guna menjamin keaslian hasil evaluasi dan mencegah kebocoran data.

Evaluasi dilakukan dengan menghitung sejumlah metrik performa klasifikasi biner, yaitu akurasi, presisi, recall, dan F1-score. Seluruh metrik ini dihitung berdasarkan hasil prediksi model terhadap data uji, yang kemudian dibandingkan langsung dengan label ground-truth. Di samping itu, hasil klasifikasi divisualisasikan dalam bentuk confusion matrix, yang menggambarkan jumlah prediksi benar dan salah untuk masing-masing kelas. Melalui confusion matrix ini, dapat dianalisis sejauh mana model mengalami kesalahan klasifikasi, terutama terhadap kelas yang lebih minoritas seperti kondisi kantuk. Penelitian ini juga melakukan perbandingan antara dua model, yaitu Model baseline SlowFast Network tanpa mekanisme Self-Attention, dan Model yang diusulkan yaitu SlowFast Network yang ditambahkan dengan Multi-Head Self Attention (MHSA). Tujuan dari perbandingan ini adalah untuk mengetahui sejauh mana kontribusi MHSA dalam meningkatkan akurasi dan ketepatan klasifikasi temporal pada data video pengemudi. Hasil evaluasi dari kedua model digunakan untuk menarik kesimpulan tentang efektivitas

arsitektur yang diusulkan dalam mendeteksi perilaku kantuk secara lebih akurat dan adaptif terhadap kondisi nyata.

BAB IV

HASIL PENELITIAN DAN PEMBAHASAN

4.1. Deskripsi Umum Penelitian

Deteksi kantuk pada pengemudi merupakan salah satu tantangan krusial dalam sistem keselamatan berkendara, khususnya di sektor transportasi dan industri logistik. Kondisi kantuk sering kali tidak terdeteksi secara eksplisit, namun dapat dikenali melalui pola gerakan mikro seperti kedipan mata lambat, penurunan posisi kepala, serta perubahan ekspresi wajah yang halus. Oleh karena itu, diperlukan pendekatan berbasis computer vision dan deep learning yang mampu mengekstraksi informasi spasial (ruang) dan temporal (waktu) dari video pendek untuk mendeteksi kondisi tersebut secara akurat. Dalam penelitian ini, diterapkan arsitektur SlowFast Network yang dirancang khusus untuk analisis video dengan dua jalur pemrosesan paralel: jalur cepat (Fast Pathway) yang memproses 32 frame untuk menangkap detail gerakan cepat, dan jalur lambat (Slow Pathway) yang memproses 8 frame untuk menangkap konteks spasial global. Kebaruan utama dalam penelitian ini terletak pada integrasi lapisan Multi-Head Self-Attention (MHSA) setelah proses penggabungan fitur (feature fusion). Mekanisme atensi ini bertujuan untuk memperkuat kemampuan model dalam memahami keterkaitan temporal antar-frame secara kontekstual, sehingga sistem dapat membedakan sinyal kantuk yang halus maupun eksplisit dengan lebih presisi dibandingkan arsitektur konvensional.

4.1.1. Dataset

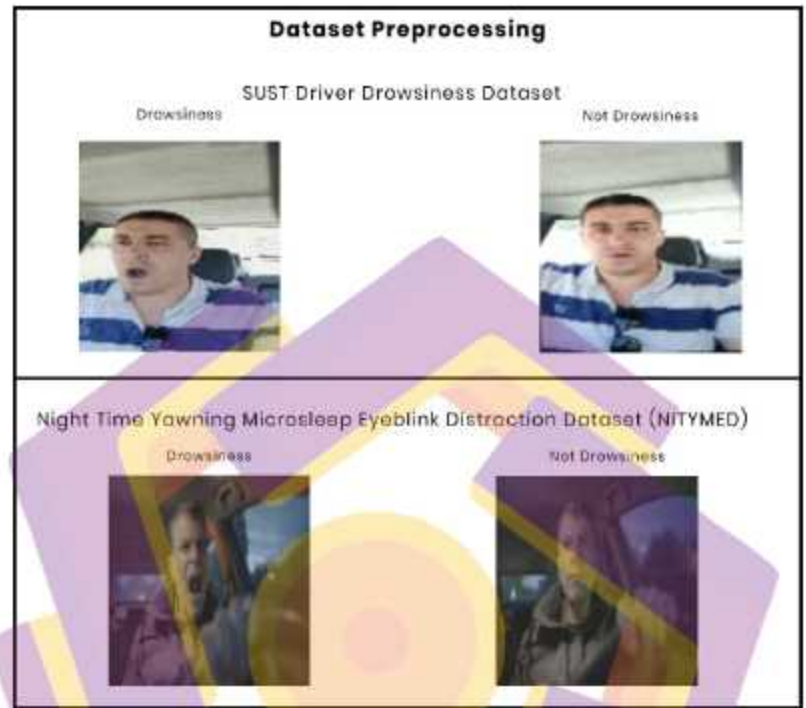
Penelitian ini memanfaatkan dua sumber data utama, yaitu SUST *Driver Drowsiness Dataset* (SUST-DDD) dan NITYMED *Night time yawning microsleep eyeblink distraction dataset*, untuk menguji kehandalan model dalam berbagai skenario. Penggunaan dataset ini dibagi ke dalam dua skenario eksperimen; skenario pertama hanya menggunakan SUST-DDD untuk keperluan validasi baseline dan perbandingan adil (fair comparison) dengan penelitian terdahulu, sedangkan skenario kedua menggabungkan SUST-DDD dan NITYMED untuk menguji kemampuan generalisasi model terhadap variasi ekspresi kantuk yang lebih luas. Karakteristik kedua dataset ini saling melengkapi. SUST-DDD berisi rekaman video dari 19 partisipan yang merekam diri sendiri menggunakan kamera ponsel saat berkendara, dengan keunggulan sifatnya yang natural (naturalistic) di mana ekspresi kantuk muncul spontan. Namun, tantangan utamanya adalah perbedaan visual antara kelas drowsy dan not drowsy yang sering kali sangat halus (subtle). Untuk melengkapi kekurangan tersebut, dataset NITYMED ditambahkan pada eksperimen tahap kedua. Dataset ini menyediakan rekaman pengemudi di malam hari dengan ekspresi kantuk yang sangat eksplisit, seperti menguap lebar atau microsleep nyata, yang membantu model mempelajari fitur kantuk yang lebih tegas. Sebelum digunakan untuk pelatihan, seluruh video dari kedua dataset melewati tahapan pra-pemrosesan yang ketat guna menstandarisasi input model. Tahap pertama adalah segmentasi temporal, di mana video asli dipotong menjadi klip-klip pendek berdurasi 2 detik menggunakan metode sliding window dengan overlap sebesar 50%. Penerapan overlap ini tidak hanya berfungsi untuk

memperbanyak sampel data, tetapi juga menjaga kontinuitas informasi temporal antar-klip agar tidak ada transisi gerakan mikro yang terputus. Selanjutnya, setiap frame dalam klip diubah resolusinya (resizing) menjadi 112×112 piksel dan dikonversi dari format BGR ke RGB, kemudian nilai piksel dinormalisasi ke rentang [0, 1] untuk mempercepat konvergensi gradien saat pelatihan. Selain standardisasi input, strategi augmentasi data juga diterapkan untuk mencegah overfitting dan meningkatkan ketahanan model terhadap variasi kondisi berkendara. Augmentasi mencakup Horizontal Flip untuk membalik gambar, Random Rotation (-10° hingga 10°) untuk mensimulasikan guncangan atau posisi kepala miring. Brightness Adjustment (faktor 0.7 - 1.3) untuk simulasi perubahan cahaya, serta Gamma Correction (0.8 - 1.2) untuk mengatur kontras gambar. Setelah seluruh proses preprocessing dan segmentasi selesai, dataset SUST-DDD yang menjadi fokus utama perbandingan terbagi menjadi ribuan klip video pendek dengan distribusi sebagai berikut:

- Training Set: Terdiri dari 14.040 klip video drowsy dan 15.822 klip video not drowsy. Data ini telah melalui proses augmentasi untuk memperkaya variasi pembelajaran.
- Validation Set: Terdiri dari 873 klip video drowsy dan 981 klip video not drowsy, digunakan untuk memantau performa model selama pelatihan.
- Test Set: Terdiri dari 876 klip video drowsy dan 1.005 klip video not drowsy, digunakan sebagai data uji final yang tidak pernah dilihat oleh model sebelumnya

Gambar 4.1. Dataset SUST-DDD dan NITYMEDD

Visualisasi pada Gambar 4.1 di atas memperlihatkan perbandingan karakteristik visual data mentah (raw data) antara SUST-DDD yang cenderung natural dengan NITYMED yang ekspresif. Sementara itu, untuk memastikan model menerima input yang konsisten dan kaya variasi, data tersebut ditransformasi melalui tahapan preprocessing dan augmentasi. Hasil dapat dilihat pada Gambar 4.2 berikut.



Gambar 4.2. Dataset Setelah Preprocessing

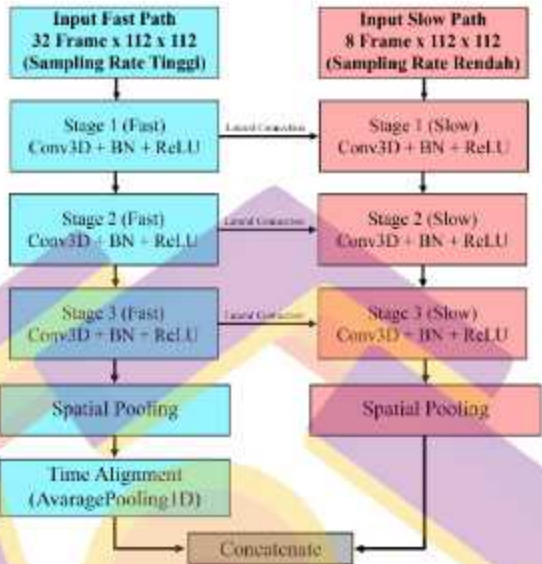
4.1.2. Arsitektur dan Metode

Dalam penelitian ini, peneliti mengembangkan model berbasis video dengan arsitektur SlowFast Network yang telah dimodifikasi dengan penambahan Multi-Head Self Attention Layer untuk meningkatkan kemampuan model dalam mendeteksi perilaku kantuk pada pengemudi.

a. Arsitektur Slowfast

Model SlowFast memproses video melalui dua jalur paralel, yaitu Fast Pathway dan Slow Pathway, yang masing-masing menerima input frame dengan frekuensi berbeda. Fast Pathway menerima 32 frame berukuran

112×112 piksel dengan sampling rate tinggi dan bertugas menangkap perubahan cepat secara temporal seperti kedipan mata atau gerakan kepala. Sementara itu, Slow Pathway menerima 8 frame dengan sampling rate rendah dan fokus pada konteks spasial yang lebih stabil, seperti postur wajah atau arah pandangan. Berbeda dengan arsitektur standar, model ini menerapkan tiga tahapan (stages) blok konvolusi 3D (Conv3D Blocks) pada setiap jalurnya. Setiap blok terdiri dari lapisan Conv3D, Batch Normalization, dan aktivasi ReLU. Selain itu, terdapat mekanisme Lateral Connection yang menghubungkan jalur Fast ke jalur Slow pada setiap tahap, yang bertujuan untuk memfusikan informasi gerakan detail ke dalam jalur spasial. Pada tahap akhir ekstraksi fitur, model tidak melakukan Global Average Pooling (GAP) secara penuh, melainkan melakukan Spatial Pooling (merata-ratakan dimensi tinggi dan lebar, namun mempertahankan dimensi waktu). Jalur Fast kemudian diselaraskan waktunya (time alignment) menggunakan Average Pooling 1D agar memiliki panjang sekuens yang sama dengan jalur Slow (yaitu 8 time steps). Hasil dari kedua jalur kemudian digabungkan (Concatenate) menghasilkan tensor fitur gabungan yang mempertahankan urutan temporal untuk diproses lebih lanjut oleh lapisan Attention.



Gambar 4.3. Alur Pemrosesan Paralel pada Arsitektur SlowFast

b. Multi-Head Self Attention

Setelah keluaran dari jalur Fast dan Slow disatukan melalui operasi concatenation, tensor fitur gabungan tersebut diproses oleh lapisan Multi-Head Self Attention (MHSA). Tujuannya adalah untuk menangkap dependensi temporal jangka panjang antar-frame yang mungkin terlewatkan oleh operasi konvolusi standar. Dalam implementasinya, lapisan ini menggunakan 4 heads yang bekerja secara paralel untuk mempelajari berbagai aspek dinamika kantuk. Struktur MHSA ini menerapkan mekanisme Residual Connection dan Layer Normalization untuk menjaga stabilitas aliran gradien selama pelatihan. Selanjutnya, dimensi temporal dirangkum menjadi satu vektor fitur

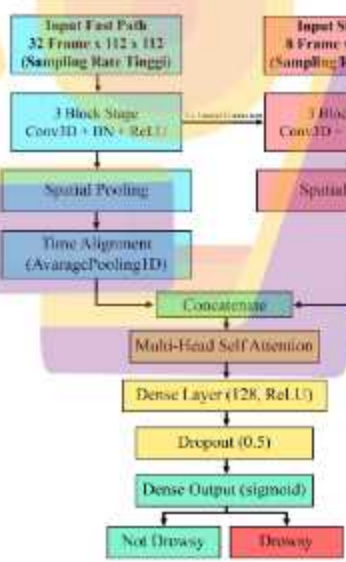
representatif menggunakan Global Average Pooling 1D. Vektor fitur tersebut kemudian diteruskan ke lapisan Dense dengan 128 unit (aktivasi ReLU) dan regulerisasi Dropout sebesar 0.5 untuk mencegah overfitting, sebelum akhirnya masuk ke lapisan output dengan aktivasi Sigmoid untuk klasifikasi biner (Drowsy / Not Drowsy).

Gambar 4.4. Alur Pemrosesan Multi-Head Self Attention

c. Struktur Model Akhir

Model yang diusulkan dalam penelitian ini dibangun dengan struktur integrasi dua jalur input video yang diproses secara paralel. Jalur Fast menerima 32 frame dan jalur Slow menerima 8 frame, masing-masing dengan resolusi 112×112 piksel. Setiap jalur melewati tiga tahapan blok Conv3D untuk mengekstraksi fitur spasial dan temporal secara mendalam, dengan mekanisme lateral connection di setiap tahapnya. Hasil ekstraksi fitur dari kedua jalur tidak

langsung diringkas habis, melainkan melalui tahap Spatial Pooling (mempertahankan dimensi waktu) dan penyelarasan waktu (Time Alignment) pada jalur cepat. Fitur kemudian digabung melalui operasi concatenation. Vektor gabungan ini selanjutnya diproses oleh lapisan Multi-Head Self Attention (MHSA) untuk memperkuat pemahaman terhadap keterkaitan antar fitur dalam domain spasio-temporal. Output dari lapisan attention ini dirangkum menggunakan Global Average Pooling 1D dan diteruskan ke lapisan Dense dengan 128 unit aktivasi ReLU, diikuti oleh Dropout sebesar 0.5 sebagai langkah regularisasi. Akhirnya, prediksi dilakukan melalui satu neuron output dengan fungsi aktivasi Sigmoid yang mengklasifikasikan kondisi pengemudi sebagai drowsy atau not drowsy.



Gambar 4.5. Arsitektur model SlowFast dengan penambahan Multi-Head Self Attention

4.2. Hasil dan Evaluasi Model

Paparan hasil eksperimen dalam penelitian ini mencakup analisis kinerja dari tiga skenario pengembangan model, yaitu model SlowFast Network sebagai baseline, model SlowFast dengan integrasi Multi-Head Self Attention (MHSA), serta model final yang dilatih menggunakan dataset gabungan. Seluruh rangkaian eksperimen, mulai dari pelatihan hingga pengujian, dilaksanakan menggunakan infrastruktur komputasi berbasis GPU (Graphics Processing Unit) RTX 3060 12GB untuk mengoptimalkan efisiensi komputasi paralel pada operasi konvolusi 3D. Konsistensi dan validitas komparasi antar-model dijaga melalui penerapan konfigurasi hyperparameter yang seragam. Proses pelatihan dilaksanakan selama 50 epoch dengan ukuran batch sebesar 8, serta menggunakan algoritma optimasi Adam dengan laju pembelajaran (learning rate) awal sebesar 1e-4. Fungsi kerugian Binary Cross-Entropy diterapkan sebagai fungsi objektif, sesuai dengan karakteristik klasifikasi biner pada penelitian ini. Selain itu, mekanisme adaptif ReduceLROnPlateau diimplementasikan untuk menurunkan laju pembelajaran secara otomatis apabila akurasi validasi mengalami stagnasi, sedangkan fitur Early Stopping diaktifkan guna menghentikan proses pelatihan lebih awal sebagai langkah preventif terhadap overfitting. Strategi augmentasi data diimplementasikan secara acak pada setiap batch pelatihan untuk meningkatkan variabilitas visual dan generalisasi model. Teknik augmentasi yang diterapkan meliputi Horizontal Flip, rotasi acak, serta penyesuaian intensitas cahaya (Brightness) dan koreksi gamma. Selanjutnya, pengukuran kinerja model dilakukan berdasarkan metrik evaluasi standar yang mencakup Akurasi, Precision, Recall, dan F1-Score. Perhitungan

metrik tersebut didasarkan pada hasil prediksi model terhadap data uji (test set) yang terisolasi dan tidak pernah dilibatkan dalam proses pelatihan maupun validasi.

4.2.1. Evaluasi Model SlowFast dan Dataset SUST

Sebagai langkah awal untuk menetapkan standar kinerja (baseline), arsitektur SlowFast diuji dalam bentuk murninya tanpa penambahan lapisan Multi-Head Self Attention (MHSA). Eksperimen ini menggunakan dataset SUST dengan konfigurasi preprocessing yang sama persis dengan eksperimen utama, bertujuan untuk mengukur kemampuan dasar arsitektur SlowFast dalam menangkap fitur kantuk tanpa bantuan mekanisme atensi. Hasil pengujian pada data uji (test set) menunjukkan bahwa model baseline ini mampu mencapai Akurasi sebesar 84.48%. Meskipun hasil ini cukup baik, analisis mendalam pada setiap kelas menunjukkan adanya keseimbangan kinerja yang belum optimal antara deteksi kondisi terjaga dan mengantuk. Rincian evaluasi performa untuk setiap kelas disajikan secara lengkap pada Tabel 4.1.

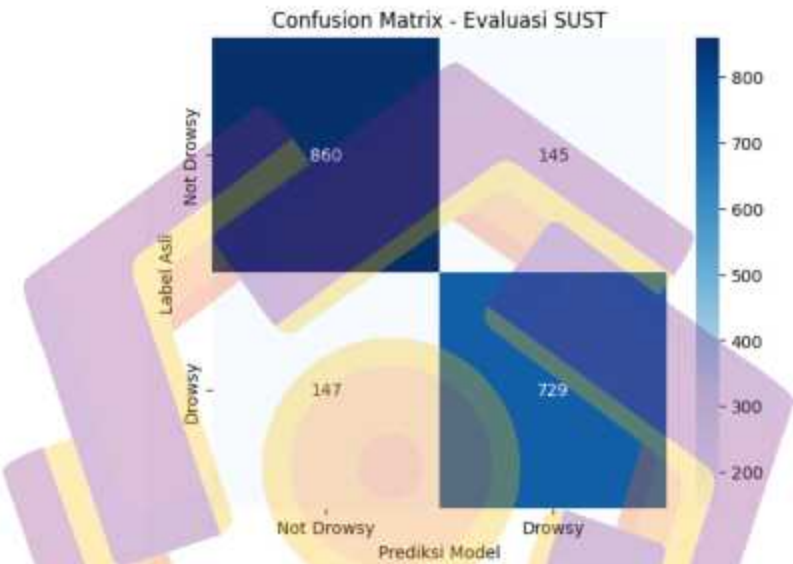
Tabel 4.1. Hasil Evaluasi Model SlowFast + Dataset SUST

Evaluasi	**Nilai**	**Keterangan**
Accuracy	84.48%	Proporsi prediksi yang benar dari seluruh data uji.
Precision Not Drowsy	0.85	Kemampuan model memprediksi *not drowsy* secara tepat tanpa kesalahan.

Recall Not Drowsy	0.86	Kemampuan model menangkap semua kasus *not drowsy* yang sebenarnya.
Precision Drowsy	0.83	Tingkat ketepatan prediksi drowsy yang benar dari seluruh prediksi drowsy.
Recall Drowsy	0.83	Tingkat keberhasilan model mengenali semua video drowsy yang sebenarnya.
F1-Score Not Drowsy	0.85	Harmoni antara precision dan recall pada kelas not drowsy, mencerminkan keseimbangan klasifikasi.
F1-Score Drowsy	0.83	Harmoni antara precision dan recall pada kelas drowsy, mencerminkan keseimbangan klasifikasi.

Berdasarkan Tabel 4.1, terlihat bahwa performa model pada kelas Drowsy (Precision 0.83, Recall 0.83) sedikit lebih rendah dibandingkan kelas Not Drowsy (Precision 0.85, Recall 0.86). Hal ini mengindikasikan bahwa tanpa mekanisme attention, model SlowFast standar masih memiliki sedikit kesulitan dalam membedakan ciri-ciri kantuk yang halus (subtle) dibandingkan dengan ekspresi terjaga yang lebih jelas. Analisis kesalahan klasifikasi diperjelas melalui Confusion Matrix pada Gambar 4.6. Dari total sampel uji, model melakukan kesalahan prediksi yang cukup merata di

kedua sisi: terdapat 147 video Drowsy yang gagal dideteksi (terbaca sebagai Not Drowsy) dan 145 video Not Drowsy yang salah diprediksi sebagai kantuk



Gambar 4.6. Confusion Matrix Hasil Predikisi Model SlowFast (Baseline) pada Dataset SUST

4.2.2. Evaluasi Model SlowFast + MHSA dan Dataset SUST

Model integrasi SlowFast Network dan Multi-Head Self-Attention (MHSA) yang diusulkan dalam penelitian ini dievaluasi menggunakan dataset SUST, di mana dataset tersebut telah melewati serangkaian tahapan pra-pemrosesan standar. Tahapan ini meliputi segmentasi temporal dengan memotong video menjadi klip berdurasi 2 detik dan overlap 0.5 detik untuk menjaga kontinuitas informasi, serta augmentasi data yang menerapkan

teknik Horizontal Flip, Rotation (-10° hingga 10°), Brightness Adjustment, dan Gamma Correction secara acak khusus pada data latih (training set) guna meningkatkan ketahanan model terhadap variasi posisi kepala dan pencahayaan. Selain itu, dilakukan normalisasi melalui konversi ruang warna dari BGR ke RGB serta penskalaan nilai piksel (rescaling) menjadi rentang 0-1. Hasil pengujian menunjukkan bahwa penambahan mekanisme MHSA memberikan dampak krusial terhadap performa model, di mana mekanisme atensi ini terbukti berhasil meningkatkan kemampuan model dalam menangkap dependensi spasial-temporal jangka panjang sekaligus memfilter noise latar belakang yang tidak relevan. Berdasarkan pengujian pada dataset SUST, model berhasil mencapai akurasi sebesar 93.30%, dengan kinerja model yang dinilai melalui confusion matrix sebagaimana dirangkum dalam Tabel berikut:

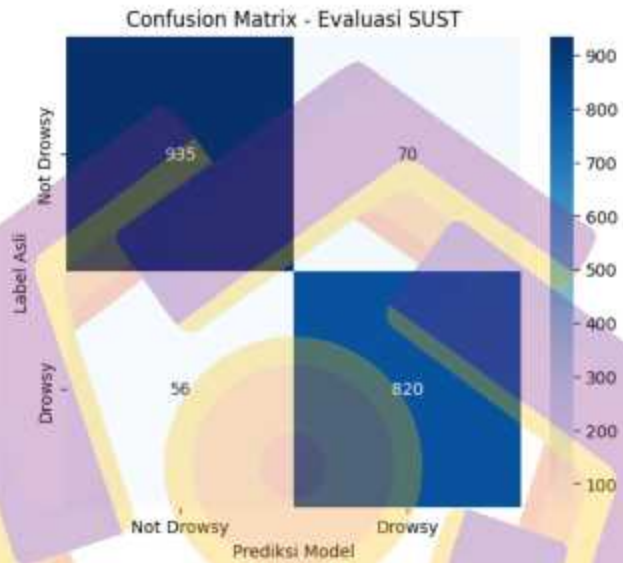
Tabel 4.2. Hasil Evaluasi Model SlowFast + MHSA + Dataset SUST

Evaluasi	**Nilai**	**Keterangan**
Accuracy	0.93	Proporsi prediksi yang benar dari seluruh data uji.
Precision Not Drowsy	0.94	Kemampuan model memprediksi *not drowsy* secara tepat tanpa kesalahan.
Recall Not Drowsy	0.93	Kemampuan model menangkap semua kasus *not drowsy* yang sebenarnya.

Precision Drowsy	0.92	Tingkat ketepatan prediksi drowsy yang benar dari seluruh prediksi drowsy.
Recall Drowsy	0.94	Tingkat keberhasilan model mengenali semua video drowsy yang sebenarnya.
F1-Score Not Drowsy	0.94	Harmoni antara precision dan recall pada kelas not drowsy, mencerminkan keseimbangan klasifikasi.
F1-Score Drowsy	0.93	Harmoni antara precision dan recall pada kelas drowsy, mencerminkan keseimbangan klasifikasi.

Berdasarkan Tabel 4.2, terlihat lonjakan performa yang signifikan dibandingkan model baseline (yang hanya mencapai 84.48%). Peningkatan paling krusial terlihat pada metrik Recall untuk kelas Drowsy yang mencapai angka 0.94 (94%). Kenaikan ini mengindikasikan bahwa integrasi MHSA berhasil mengatasi kelemahan utama arsitektur konvolusi standar, yaitu sensitivitas yang rendah terhadap sinyal kantuk yang halus. Dengan bantuan attention, model menjadi jauh lebih peka dalam mendeteksi tanda-tanda kantuk yang sebenarnya. Validasi lebih lanjut terhadap kinerja model dilakukan melalui analisis Confusion Matrix yang ditampilkan pada Gambar 4.7. Berdasarkan pengujian pada total 1.881 sampel data yang sama, integrasi mekanisme MHSA terbukti berhasil menekan tingkat kesalahan klasifikasi secara drastis dibandingkan dengan model baseline.

Peningkatan yang paling krusial terlihat pada penurunan angka False Negative, di mana kegagalan deteksi kantuk berkurang tajam dari 147 kasus menjadi hanya 56 kasus



Gambar 4.7. Confusion Matrix Hasil Predikisi Model SlowFast + MHSA pada Dataset SUST

4.2.3. Evaluasi Model Slowfast + MHSA dan Dataset SUST + NITYMED

Skenario pengujian ketiga merupakan evaluasi terhadap model final yang dilatih menggunakan dataset gabungan (SUST-DDD dan NITYMED). Pengujian ini bertujuan untuk mengukur kemampuan generalisasi model ketika dihadapkan pada variasi data yang ekstrim, mulai dari ekspresi kantuk natural hingga ekspresi eksplisit di malam hari. Evaluasi dilakukan pada data uji (test set) dengan total 2.537 klip video, yang terdiri dari 1.401

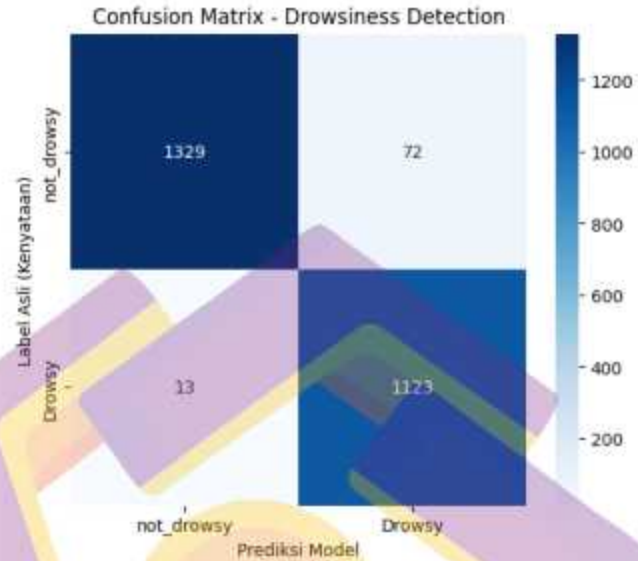
sampel kelas Not Drowsy dan 1.136 sampel kelas Drowsy. Hasil evaluasi menunjukkan performa puncak, di mana model berhasil mencapai Akurasi sebesar 96.65%. Capaian ini mengonfirmasi hipotesis bahwa penggabungan dataset yang memiliki karakteristik saling melengkapi (natural dan eksplisit), dikombinasikan dengan kemampuan atensi temporal MHSA, mampu menghasilkan model yang sangat robust. Rincian metrik evaluasi disajikan pada Tabel 4.3.

Tabel 4.3. Hasil Evaluasi Model SlowFast + MHSA + Dataset SUST + NITYMED

Evaluasi	**Nilai**	**Keterangan**
Accuracy	96.65%	Proporsi prediksi yang benar dari seluruh data uji.
Precision Not Drowsy	0.99	Kemampuan model memprediksi *not drowsy* secara tepat tanpa kesalahan.
Recall Not Drowsy	0.95	Kemampuan model menangkap semua kasus *not drowsy* yang sebenarnya.
Precision Drowsy	0.94	Tingkat ketepatan prediksi drowsy yang benar dari seluruh prediksi drowsy.
Recall Drowsy	0.99	Tingkat keberhasilan model mengenali semua video drowsy yang sebenarnya.

F1-Score (rata-rata)	0.97	Rata-rata harmonis precision dan recall, menggambarkan keseimbangan prediksi.

Analisis pada Tabel 4.3 menunjukkan capaian signifikan pada metrik Recall untuk kelas Drowsy yang mencapai angka 0.99 (99%). Tingginya nilai ini mengindikasikan bahwa model memiliki sensitivitas yang sangat baik dalam mengidentifikasi tanda-tanda kantuk. Capaian ini membuktikan efektivitas strategi penggabungan dataset (SUST dan NITYMED) yang memperkaya variasi data latih, mulai dari ekspresi kantuk yang halus hingga yang eksplisit. Selain itu, peran Multi-Head Self Attention (MHSA) terbukti krusial dalam menjaga konsistensi deteksi dengan cara memfokuskan perhatian model pada fitur temporal yang paling relevan di tengah keberagaman data tersebut. Dalam konteks sistem keselamatan berkendara, tingginya nilai Recall ini menjadi prioritas utama karena secara langsung berkorelasi dengan kemampuan sistem untuk meminimalkan False Negative, sehingga risiko kegagalan deteksi pada momen kritis dapat ditekan seminimal mungkin.



Gambar 4.8. Confusion Matrix Hasil Prediksi model SlowFast + MHSA + Dataset SUST + NITYMED

4.3. Komparasi dengan Penelitian Terdahulu

Perbandingan komprehensif dilakukan antara seluruh konfigurasi model yang dikembangkan dalam penelitian ini dengan penelitian terdahulu yang menjadi rujukan utama, yaitu karya Yılmaz dan Akcayol (2022). Langkah validasi ini melibatkan tiga skenario pengujian internal serta satu pembanding eksternal guna memberikan gambaran holistik terkait efektivitas arsitektur SlowFast Network yang dikombinasikan dengan lapisan Multi-Head Self Attention (MHSA). Validasi ini bertujuan untuk memverifikasi posisi kontribusi penelitian terhadap perkembangan metode deteksi kantuk saat ini.

Tabel 4.4. Komparasi Komprehensif dengan Penelitian Terdahulu

Metode	Dataset	Akurasi	Precision	Recall	F1-Score	Keterangan
VGG19+LSTM	SUST-DDD	90.53%	91.74%	91.28%	91.46%	Yılmaz & Akcayol (2022)
SlowFast (Baseline)	SUST-DDD	84.48%	83.41%	83.22%	83.31%	Model Dasar (Tanpa MHSA)
SlowFast + MHSA	SUST-DDD	93.30%	92.13%	93.61%	92.87%	Model Perbandingan
SlowFast + MHSA	SUST-DDD + NITYMED	96.65%	97%	97%	97%	Model Terbaik (Data Fusion)

Berdasarkan Tabel 4.4, analisis hasil menunjukkan bahwa model SlowFast murni tanpa MHSA menghasilkan akurasi sebesar 84.48%. Capaian ini masih berada di bawah performa penelitian terdahulu yang mencapai 90.53%. Rendahnya akurasi pada model dasar ini mengindikasikan bahwa arsitektur SlowFast standar

menghadapi kendala dalam menangkap fitur temporal kritis pada dataset SUST, terutama karena ekspresi kantuk yang terekam pada dataset tersebut cenderung halus (subtle) dan sulit dideteksi tanpa mekanisme atensi tambahan. Penambahan lapisan Multi-Head Self Attention (MHSA) pada arsitektur SlowFast terbukti memberikan kenaikan performa yang signifikan. Pada pengujian dengan dataset yang sama (SUST-DDD), akurasi meningkat dari 84.48% menjadi 93.30%. Peningkatan ini tidak hanya melampaui model *baseline internal*, tetapi juga berhasil mengungguli metode VGG19+LSTM dari penelitian rujukan dengan selisih sebesar +2.77%. Secara spesifik, nilai Recall mengalami kenaikan drastis menjadi 93.61%, yang membuktikan bahwa mekanisme attention berhasil membantu model untuk memfokuskan pembelajarannya pada detail gerakan mikro, seperti kedipan mata lambat, yang sebelumnya terlewatkan. Potensi maksimal model terlihat ketika dilatih menggunakan dataset gabungan (SUST + NITYMED), di mana performa mencapai titik puncak dengan akurasi 96.65%. Hasil ini mengonfirmasi bahwa arsitektur yang diusulkan memiliki kapasitas belajar (*learning capacity*) yang besar dan mampu memanfaatkan variasi data tambahan, khususnya ekspresi kantuk eksplisit dari NITYMED untuk meningkatkan generalisasi model secara keseluruhan. Dengan demikian, dapat disimpulkan bahwa meskipun model dasar SlowFast belum cukup untuk mengungguli metode terdahulu, integrasi MHSA terbukti menjadi komponen yang menaikan performa model, terutama dalam aspek sensitivitas (Recall) deteksi kantuk.

4.4. Analisis Visual dan Efektivitas MHSA

Selain evaluasi kuantitatif menggunakan metrik akurasi dan recall, validasi kualitatif dilakukan untuk memahami bagaimana model mengambil keputusan. Teknik visualisasi Gradient-weighted Class Activation Mapping (Grad-CAM) diterapkan untuk menghasilkan peta panas (heatmap) yang menunjukkan area mana pada frame video yang dianggap paling penting oleh model saat memprediksi kondisi kantuk. Analisis ini bertujuan untuk membuktikan secara visual klaim bahwa mekanisme Multi-Head Self Attention (MHSA) mampu memfokuskan perhatian model pada fitur-fitur wajah yang relevan dan mengabaikan latar belakang yang tidak perlu.

a. Analisis Dampak MHSA pada Fokus Atensi Model

Perbandingan visual antara model baseline (tanpa MHSA) dan model usulan (dengan MHSA) pada dataset SUST disajikan dalam Gambar 4.9. Sampel yang diambil adalah kondisi pengemudi yang sedang memejamkan mata (drowsy).

Gambar 4.9. Perbandingan Visualisasi Atensi (Grad-CAM) pada Dataset SUST

Berdasarkan Gambar 4.9, dapat diamati perbedaan pola atensi yang sangat signifikan antara kedua model. Pada visualisasi model tanpa MHSA (bagian tengah), distribusi peta panas terlihat menyebar secara tidak beraturan, di mana model baseline tampak mengalami "kebingungan" fitur dengan memberikan bobot atensi yang tinggi pada area yang tidak relevan, seperti lipatan jilbab, leher, serta sebagian latar belakang mobil, sementara area mata justru tidak mendapatkan fokus utama. Distraksi oleh noise visual ini menjelaskan secara logis mengapa model baseline memiliki nilai Recall yang lebih rendah (83.22%), karena kegagalannya dalam mengisolasi fitur kunci kantuk. Sebaliknya, setelah mekanisme MHSA diintegrasikan (bagian kanan), peta panas menunjukkan lokalisasi fitur yang sangat presisi. Area dengan aktivasi tertinggi—yang ditandai dengan warna biru hingga kuning cerah—terkonsentrasi secara akurat pada wilayah periocular (sekitar mata) dan pangkal hidung, sembari melakukan supresi (suppression) yang efektif terhadap fitur tidak relevan seperti pakaian dan latar belakang. Fenomena ini membuktikan bahwa lapisan MHSA berhasil mempelajari dependensi temporal jangka panjang; mengingat gerakan menutup mata adalah fitur dinamis yang terikat waktu, MHSA memampukan model untuk memfokuskan "perhatian" pada area yang mengalami perubahan gerak mikro tersebut, sehingga menghasilkan keputusan klasifikasi yang jauh lebih akurat.

b. Analisis Robustness pada Dataset Gabungan

Analisis visual selanjutnya dilakukan pada model final yang dilatih dengan dataset gabungan (SUST + NITYMED). Gambar 4.10 memperlihatkan respons

model terhadap video dari dataset NITYMED yang memiliki kondisi pencahayaan malam hari dan ekspresi menguap.

Gambar 4.10. Visualisasi Atensi Model Final pada Dataset NITYMED

Analisis visual pada Gambar 4.10 memperlihatkan respons model terhadap subjek yang sedang menunjukkan ekspresi menguap disertai kondisi mata yang menyipit. Hasil visualisasi Grad-CAM mengindikasikan bahwa model mengalokasikan atensi ganda secara simultan, yang terpusat pada area mata dan mulut. Distribusi atensi ini menunjukkan bahwa model memiliki kemampuan untuk mengintegrasikan berbagai indikator visual kantuk secara komprehensif, dan tidak hanya terbatas pada satu fitur wajah tunggal. Selain itu, meskipun terdapat objek distraktif berupa penumpang bermasker di latar belakang, peta aktivasi pada area tersebut menunjukkan intensitas yang rendah atau tidak aktif. Hal ini mengonfirmasi tingkat ketahanan (robustness) model yang tinggi dalam mendiskriminasi subjek utama (pengemudi) terhadap objek latar belakang, bahkan dalam kondisi pencahayaan minim. Konsistensi visual ini selaras dengan capaian akurasi dan recall tinggi (99%) pada evaluasi kuantitatif, sekaligus menegaskan bahwa performa model didasari oleh pemahaman fitur visual, bukan sekadar korelasi statistik semata

4.5. Analisis Dampak Preprocessing

Kinerja tinggi yang dicapai oleh model tidak hanya bergantung pada arsitektur jaringan saraf semata, melainkan juga sangat dipengaruhi oleh kualitas dan kesiapan data input. Sesuai dengan hasil evaluasi yang telah dipaparkan, tahapan preprocessing dan augmentasi data terbukti memiliki dampak signifikan terhadap stabilitas pelatihan dan kemampuan generalisasi model. Berikut adalah analisis dampak dari setiap tahapan pemrosesan data berdasarkan bukti empiris hasil eksperimen:

a. Efektivitas Mekanisme Sliding Window dengan Overlap

   Penerapan teknik sliding window dengan durasi 2 detik dan overlap 50% terbukti menjadi strategi yang krusial. Bukti efektivitas ini terlihat pada nilai Recall model baseline (83.22%) dan model usulan (93.61%) yang cukup tinggi. Hal ini mengindikasikan bahwa jendela waktu 2 detik adalah durasi yang optimal untuk menangkap fitur dinamis kantuk seperti microsleep atau kedipan lambat, yang umumnya berlangsung dalam hitungan detik. Tanpa segmentasi yang tepat, informasi temporal ini berisiko hilang atau tertutup oleh durasi video yang terlalu panjang (sparse information), yang akan menyebabkan model kesulitan mengisolasi momen kritis kantuk.

b. Dampak Normalisasi terhadap Konvergensi

   Proses resizing ke 112×112 piksel dan normalisasi nilai piksel ke rentang [0, 1] memberikan kontribusi langsung terhadap stabilitas proses pelatihan. Hal ini dibuktikan melalui grafik Training Loss (sebagaimana diamati pada proses pelatihan) yang menunjukkan penurunan gradien yang stabil dan konvergensi

model yang cepat dalam 50 epoch. Tanpa normalisasi, variasi intensitas piksel yang ekstrim antar-video seringkali menyebabkan vanishing gradient atau osilasi loss yang membuat model sulit belajar.

c. Peran Augmentasi terhadap Robustness (Ketahanan)

Strategi augmentasi data yang meliputi Random Rotation, Brightness Adjustment, dan Gamma Correction terbukti sukses meningkatkan ketahanan (robustness) model terhadap variasi kondisi lingkungan yang dinamis. Bukti dari efektivitas ini terlihat jelas pada hasil pengujian dataset gabungan (Tabel 4.3), di mana model mampu mengenali data dari dataset NITYMED yang didominasi oleh kondisi malam hari dengan pencahayaan minim dengan presisi yang tinggi. Secara spesifik, penerapan teknik Gamma dan Brightness Adjustment memfasilitasi model untuk tetap mengenali fitur wajah esensial meskipun dalam kondisi gelap atau kontras rendah, sedangkan teknik Rotation memastikan akurasi tetap terjaga meskipun terdapat variasi posisi kepala akibat kemiringan atau guncangan kendaraan. Tanpa intervensi augmentasi ini, model memiliki risiko tinggi mengalami overfitting pada dataset latih SUST, yang akan berujung pada kegagalan generalisasi saat dihadapkan pada dataset NITYMED yang memiliki karakteristik visual sangat berbeda.

BAB V

PENUTUP

5.1. Kesimpulan

Penelitian ini bertujuan untuk membangun dan menguji efektivitas metode deteksi kantuk berbasis video menggunakan arsitektur SlowFast Network yang dimodifikasi dengan penambahan lapisan Multi-Head Self Attention (MHSA). Berdasarkan hasil perancangan, pelatihan, dan evaluasi mendalam yang telah dilakukan, dapat ditarik beberapa kesimpulan sebagai berikut:

1. Penelitian ini berhasil mengembangkan model deteksi kantuk yang menggabungkan dua jalur informasi, yaitu jalur lambat (Slow Pathway) untuk menangkap detail visual dan jalur cepat (Fast Pathway) untuk menangkap gerakan cepat. Integrasi lapisan Multi-Head Self Attention (MHSA) pada arsitektur ini terbukti mampu berjalan dengan baik dalam memproses input video pendek (2 detik) untuk membedakan kondisi pengemudi yang mengantuk dan tidak mengantuk.
2. Penambahan lapisan MHSA terbukti memberikan dampak positif yang signifikan terhadap kinerja model. Hal ini ditunjukkan oleh kenaikan akurasi dari 84.48% (pada model dasar tanpa MHSA) menjadi 93.30% (pada model dengan MHSA) saat diuji pada dataset yang sama. Secara visual, mekanisme ini terbukti mampu membantu model untuk lebih "fokus" pada area wajah yang penting, seperti mata dan mulut, serta mengabaikan bagian latar belakang yang tidak perlu. Tanpa MHSA, model cenderung kurang sensitif dalam mendeteksi tanda-tanda kantuk yang halus.

3. Model mencapai performa terbaiknya ketika dilatih menggunakan gabungan dataset (SUST dan NITYMED), dengan capaian akurasi tertinggi sebesar 96.65%. Yang paling penting, model ini memiliki kemampuan deteksi kantuk (Recall) sebesar 99%, yang berarti sistem hampir tidak pernah melewatkan kejadian kantuk yang sebenarnya. Hasil ini membuktikan bahwa variasi data yang lengkap (mulai dari ekspresi natural hingga ekspresi jelas di malam hari) sangat membantu model menjadi lebih pintar dan tangguh dalam berbagai situasi.

5.2. Saran

Berdasarkan hasil penelitian yang telah dilakukan, terdapat beberapa hal yang dapat disarankan untuk pengembangan penelitian selanjutnya. Sistem deteksi kantuk berbasis video yang dikembangkan memiliki potensi untuk diimplementasikan dalam bentuk sistem real-time, baik pada perangkat tertanam (embedded system) maupun aplikasi mobile yang ringan dan efisien agar dapat langsung digunakan oleh pengguna di kendaraan. Penelitian selanjutnya juga dapat memperluas cakupan klasifikasi dari dua kelas (kantuk dan tidak kantuk) menjadi multi-kelas berdasarkan tingkat kantuk atau jenis gangguan pengemudi lainnya. Selain itu, model sebaiknya diuji pada populasi pengguna yang lebih beragam secara demografis, sehingga performa dan keadilan sistem dapat dijaga untuk berbagai kelompok usia, jenis kelamin, dan latar belakang pengemudi.

DAFTAR PUSTAKA

Abdulrahman Abououf, Ibrahim Sobh, Mohammad Nasser, Omar Alsaqa, Omar Elezaby, & John F. W. Zaki. (2022). Multimodel System for Driver Distraction Detection and Elimination. IEEE Access (https://doi.org/10.1109/ACCESS.2022.3188715)

Andi Asvin Mahersatillah Suradi, Samsu Alam, Mushaf, Muhammad Furqan Rasyid, Imran Djafar. (2023). Sistem Deteksi Kantuk Pengemudi Mobil Berdasarkan Analisis Rasio Mata Menggunakan Computer Vision. JUKI : Jurnal Komputer dan Informatika (https://ioinformatic.org/index.php/JUKI/article/view/269)

Balasubramani S, John Aravindhar D, P.N. Renjith, & K. Ramesh. (2024). DDSS: Driver decision support system based on the driver behaviour prediction to avoid accidents in intelligent transport system. International Journal of Cognitive Computing in Engineering (https://doi.org/10.1016/j.ijcce.2023.12.001)

Dimitris Tsiktsiris, Antonios Lalas, Minas Dasygenis, Dan Konstatinos Votis. (2024). Multimodal Abnormal Event Detection in Public Transportation. IEEE Access (https://doi.org/10.1109/ACCESS.2024.3425308)

Esra Kavalci Yilmaz dan M. Ali Akcayol. (2022). SUST-DDD: A Real-Drive Dataset for Driver Drowsiness Detection. The 31st Conference of Fruct Association (https://zenodo.org/records/6519933)

Dimitris Tsiktsiris, Antonios Lalas, Minas Dasygenis, Dan Konstatinos Votis. (2024). Multimodal Abnormal Event Detection in Public Transportation. IEEE Access (https://doi.org/10.1109/ACCESS.2024.3425308)

Dawei Yang, Yan Wang, Ran Wei, Jiapeng Guan, Xiaohua Huang, Wei Cai, & Zhe Jiang. (2024). An efficient multi-task learning CNN for driver attention monitoring. Journal of Systems Architecture. ELSEVIER (https://doi.org/10.1016/j.sysarc.2024.103085)

Fangming Qu, Nolan Dang, Borko Furht & Mehrdad Nojoumian. (2024). Comprehensive study of driver behavior monitoring systems using computer vision and machine learning techniques. Journal of Big Data (https://doi.org/10.1186/s40537-024-00890-0)

Hangyue Zhao, Yuchao Xiao, dan Yanyun Zhao. (2022). PAND: Precise Action Recognition on Naturalistic Driving. IEEE Access (https://doi.org/10.1109/CVPRW56347.2022.00372)

J Robert Theivadas & Suresh Ponnan. (2024). VigilEye: Machine learning-powered driver fatigue recognition for safer roads. Science direct, Measurement: Sensors (https://doi.org/10.1016/j.measen.2024.101186)

Kai Liang, Jun Wang, dan Abhir Bhalerao. (2024). Lane Change Classification and Prediction with Action Recognition Networks. arXiv:2208.11650 (https://doi.org/10.48550/arXiv.2208.11650)

Maykol Santosa, Paulo Jorge Coelho, Ivan Miguel Pires, Pedro Goncalves, & Goncalo Paiva Dias. (2024). An Overview of Machine Learning Algorithms to Reduce

Driver Fatigue and Distraction-Related Traffic Accidents. Science direct, Procedia computer science (https://doi.org/10.1016/j.procs.2024.06.003)

Md. Uzzol Hossain, Md. Ataur Rahman , Md. Manowarul Islam, Arnisha Akhter, Md. Ashraf Uddin, Bikash Kumar Paul. (2022). Automatic driver distraction detection using deep convolutional neuralnetworks. Science direct, Intelligent Systems with Applications (https://doi.org/10.1016/j.iswa.2022.200075)

Shahzeb Ansari, Fazel Naghdy, Haiping Du, & Yasmeen Naz Pahnwar. (2021). Driver Mental Fatigue Detection Based on Head Posture Using New Modified reLU-BiLSTM. IEEE Transaction on Intelligent Transportation System (https://doi.org/10.1109/TITS.2021.3098309)

V. Uma Maheswari, Rajanikanth Aluvalu, Mvv Prasad Kantipudi, Krishna Keerthi Chennam, Ketan Kotecha, & Jatinderkumar R. Saini. (2022). Driver Drowsiness Prediction Based on Multiple Aspects Using Image Processing Techniques. IEEE Access (https://doi.org/10.1109/ACCESS.2022.3176451)

Ward Ahmed Al-Hussein, Lip Yee Por, Miss Laiha Mat Kiah, & Bilal Bahaa Zaidan (2022, Januari). Driver Behavior Profiling and Recognition Using Deep-Learning Methods: In Accordance with Traffic Regulations and Experts Guidelines. International Journal of Environmental Reasearch and Public Health. MDPI. (https://www.mdpi.com/1660-4601/19/3/1470)

Zhen Liang, Ying Hu, dan Ransheng Yang. (2023). Research on Industrial Human Action Recognition based on Improved Slowfast. International Conference on Intelligent Autonomous System (ICoIAS). IEEE (https://doi.org/10.1016/j.aap.2024.107545)

Zulqarnain H. Khattak, Wan Li, Thomas Karnowski, & Asad J. Khattak. (2024). *The role of driver head pose dynamics and instantaneous driving in safety critical events: Application of computer vision in naturalistic driving*. Accident Analysis and Prevention. ELSEVIER (DOI: 10.1109/ICOIAS61634.2023.00023)