BAB II
TINJAUAN PUSTAKA DAN LANDASAN TEORI

2.1 Tinjauan Pustaka

Penelitian mengenai deteksi kantuk telah banyak dilakukan terlebih penelitian penting karena dapat mewaspadi sering terjadinya kecelakaan pada lalu lintas dijalan raya. Salah satu yang sudah dilakukan Sistem deteksi kantuk yang dikembangkan menggunakan teknik *Eye Aspect Ratio* (EAR) mencapai akurasi 90% dalam kondisi eksperimental, berhasil mengenali apakah pengemudi mengantuk atau tidak. Sistem terus menganalisis keadaan penutupan mata secara real-time menggunakan teknik pemrosesan citra dan perhitungan rasio mata. Sistem ini menggabungkan penggunaan kamera Pi, Raspberry Pi 4, dan modul GPS untuk mendeteksi dan menganalisis keadaan pengemudi. Makalah ini menyarankan penggunaan LED inframerah untuk meningkatkan kinerja sistem deteksi kantuk berbasis citra. Nilai EAR 0,25 ditemukan bekerja paling baik untuk aplikasi ini, karena dengan cepat turun ke nol selama berkedip, menunjukkan kantuk. Para peneliti disini menyarankan untuk dikembangkan memberikan sensor tambahan untuk menganalisis fisiologi seperti detak jantung. (Saravanaraj Sathasivam dkk., 2020)

Penelitian dilakukan kembali dengan deteksi kantuk secara real time menggunakan analisis mata kantuk Sistem yang dikembangkan mencapai akurasi sekitar 92,5% pada dataset menguap (YawDD) yang digunakan untuk evaluasi. Sistem yang diusulkan ternyata dapat diterapkan pada mesin, kuat, dan dapat diandalkan untuk digunakan. Dalam makalah tersebut disebutkan penggunaan rasio aspek mata (EAR) dan durasi/hitungan kedipan sebagai parameter untuk mengetahui kondisi mengantuk pengemudi. Ketika nilai EAR turun di bawah ambang batas yang telah diinisialisasi sebelumnya dan penghitung kedipan melebihi batas tertentu, pesan peringatan akan dihasilkan untuk pengemudi (Pandey & Muppalaneni, 2021).Penelitian lainnya mengusulkan suatu metode yang memanfaatkan kamera dan modul untuk mendeteksi tingkat kelelahan pengemudi

dengan menganalisis frekuensi kepala miring dan kedipan mata. Keakuratan identifikasi wajah dan mata dievaluasi melalui penilaian terhadap relawan. Sistem terus memantau pola mata pengemudi dan membandingkannya dengan nilai default yang disimpan dalam *database*. Perbandingan ini dilakukan dengan menggunakan Eye Aspect Ratio (EAR) untuk mengidentifikasi tingkat kantuk. Metode yang diusulkan mencapai akurasi yang signifikan dalam mendeteksi kantuk pengemudi, dengan tingkat akurasi sekitar 85 hingga 90% berdasarkan nilai *Eye Aspect Ratio* (EAR). Sistem terus memantau pola mata pengemudi secara real-time dan memicu alarm ketika nilai EAR mencapai nol, menunjukkan rasa kantuk. Selain itu, SMS dikirim ke otoritas terkait untuk tindakan segera Data real-time yang dikumpulkan dari proses pemantauan membantu dalam menggambar diagram waktu untuk mengidentifikasi pola kantuk dan memahami perilaku pengemudi. Informasi ini dapat digunakan untuk memprediksi perilaku pengemudi dan mencegah tabrakan otomatis (Satyanarayana dkk., 2021).

Penerapan mesin learning banyak digunakan pada penelitian salah satunya dalam bidang lalu lintas seperti deteksi kantuk berbasis sinyal fisiologis menggunakan mesin pembelajaran: Pendekatan sinyal tunggal dan hibrida, data yang digunakan pada penelitian ini terdiri dari pengamatan peserta melalui webcam selama tes kewaspadaan psikomotor (PVT) Sebanyak 26 data peserta digunakan untuk analisis selanjutnya, setelah mengecualikan peserta yang menunjukkan tanda-tanda kantuk yang ekstrem atau sangat terganggu selama tugas. Hasil penelitian menunjukkan bahwa ukuran fisiologis tunggal menunjukkan pola kinerja tertentu, dengan sensitivitas yang lebih tinggi dan spesifisitas yang lebih rendah atau sebaliknya. Sebaliknya, model berbasis biosignal hibrida memberikan profil kinerja yang lebih baik, mengurangi perbedaan antara sensitivitas dan spesifisitas. Keakuratan pendekatan tunggal bervariasi, dengan EOG mengungguli EEG dan EKG, sedangkan pendekatan hibrida yang menggabungkan sinyal EEG, EOG, dan EKG mencapai akurasi terbaik. Temuan menunjukkan perbedaan yang lebih tinggi antara sensitivitas dan spesifisitas dalam pendekatan tunggal, yang berkurang dalam strategi hibrida. Pengklasifikasi ANN menunjukkan kinerja keseluruhan tertinggi di antara model pembelajaran mesin yang diawasi yang digunakan dalam

penelitian ini. (Hasan dkk., 2022). Penelitian Drowsiness detection using portable wireless EEG, penelitian tersebut menggunakan perangkat Electroencephalogram (EEG) kelas konsumen nirkabel ringan yang dapat dikenakan yang disebut Muse-2 untuk merekam sinyal EEG dari subjek. Data EEG mampu mengklasifikasikan keadaan waspada dan mengantuk dengan akurasi 78,3% menggunakan pengklasifikasi Support Vector Machine (SVM) dengan validasi lintas subjek. Proses pemilihan fitur mengungkapkan bahwa fitur EEG yang diekstraksi dari elektroda temporal (TP9 dan TP10) lebih signifikan untuk deteksi kantuk daripada fitur dari elektroda frontal (AF7 dan AF8). Fitur EEG yang diekstraksi dari elektroda temporal menunjukkan koefisien korelasi yang lebih tinggi dengan detak jantung, yang konsisten dengan hasil pemilihan fitur. Awalnya, klasifikasi menggunakan semua 38 fitur menghasilkan akurasi rendah 43,65%. Namun, setelah menerapkan kriteria pemilihan fitur, set fitur yang dikurangi diperoleh, yang meningkatkan akurasi klasifikasi. Makalah ini bertujuan untuk menyelidiki beragam fitur EEG untuk deteksi kantuk dan mendapatkan serangkaian fitur yang berkurang yang secara akurat memprediksi kantuk.(Gangadharan K & Vinod, 2022). Real-Time EAR Based Drowsiness Detection Model for Driver Assistant System, Sistem yang diusulkan mendeteksi fitur wajah dan mengubah citra menjadi skala abu-abu sebelum menerapkan algoritma deteksi Fitur Wajah Dlib. Eye Aspect Ratio (EAR) dihitung menggunakan rumus jarak Euclidean antara berbagai titik di mata. Model deteksi kantuk yang diusulkan mencapai presisi 94,44% dalam mendeteksi subjek mengantuk menggunakan visi komputer dan teknik pembelajaran mesin. Model memproses dataset untuk pelatihan dan pengujian selama 300 zaman, dan grafik akurasi dan kerugian . Skenario waktu nyata dari model yang diusulkan tercermin pada Gambar. 3, di mana kedua mata dilacak dan kedip-kedip diidentifikasi untuk mendeteksi kantuk. Teknik ambang batas EAR digunakan untuk menentukan sensitivitas deteksi kantuk. EAR kurang dari 0,25 memberikan “Alarm Mengantuk,” sedangkan EAR yang sama dengan atau lebih besar dari 0,25 dianggap tidak mengantuk. Makalah penelitian tidak memberikan rincian spesifik tentang kinerja keseluruhan model atau analisis komparatif apa pun dengan model lain yang ada(Singh dkk., 2022)

Monitoring System of Drowsiness and Lost Focused Driver Using Raspberry Pi. Metode Haar Cascade Classifier digunakan untuk mengidentifikasi area wajah, mata, dan mulut untuk menentukan kondisi mengantuk Sampel video driver ditangkap menggunakan kamera yang terhubung ke Raspberry Pi, dan sampel ini diproses menggunakan metode Haar Cascade Classifier untuk mendeteksi kantuk dan kehilangan fokus. Sistem menganalisis keadaan mata pengemudi (terbuka, setengah terbuka, dan tertutup) menggunakan pemrosesan citra dan jaringan saraf buatan (ANN). Keakuratan sistem dalam mendeteksi fokus yang hilang adalah 88,00%, sedangkan akurasi dalam mendeteksi kantuk adalah 90,40% Sistem menganalisis keadaan mata pengemudi dan menentukan dua parameter kehilangan fokus dan kantuk.(Adi dkk., 2020). Drowsy Detection System by Facial Landmark and Light Gradient Boosting Machine Method, Studi ini mengembangkan sistem deteksi kantuk menggunakan landmark wajah dan metode Light Gradient Boosting Machine (Light GBM). Akurasi prototipe sistem meningkat sebesar 21,8% dibandingkan dengan penelitian sebelumnya. Prototipe sistem diintegrasikan ke dalam aplikasi PC yang menghasilkan suara peringatan selama lima detik ketika driver mengantuk terdeteksi Sistem deteksi kantuk berkinerja baik dalam pengujian, dengan daya ingat tinggi, presisi, Skor F1, dan nilai akurasi. Algoritma klasifikasi LightGBM menghasilkan hasil terbaik dalam hal akurasi, ingatan, presisi, dan Skor F1. Modifikasi dalam matriks kebingungan menunjukkan bahwa perubahan dalam metode klasifikasi meningkatkan prediksi yang benar sekaligus mengurangi kesalahan. (Asdyo dkk., 2023).

Beberapa penelitian yang berkaitan dengan deteksi kantuk LSTM-CNN model of drowsiness detection from multiple consciousness states acquired by EEG, Model LSTM mencapai akurasi tertinggi 86% untuk panjang jendela 1 detik, sedangkan model LSTM-CNN menghasilkan indeks kappa terbaik 0,77 untuk panjang jendela 4 detik. Model LSTM-CNN yang diusulkan menghasilkan akurasi rata-rata 85,6% dan indeks kappa 0,77 untuk masalah klasifikasi tiga kelas. Untuk klasifikasi biner keadaan kesadaran normal (terjaga) dan kesadaran rendah (kantuk dan tidur), model LSTM-CNN mencapai skor F1 0,95 dalam 4000-ms dan skor F1 0,94 saat menggunakan data input pendek 500 msec. Kinerja pengklasifikasi BP-

LDA dan MVAR-LDA hampir sama, seperti halnya kinerja pengklasifikasi BP-SVM dan MVAR-SVM. Perbedaan akurasi antara pengklasifikasi yang sama kurang dari 1, dan perbedaan indeks kappa di bawah 0,05, menunjukkan bahwa fitur BP dan MVAR berbagi banyak informasi (Lee & An, 2023). Prediction of drowsiness using EEG signals in young Indonesian drivers, Uji signifikansi statistik, regresi logistik, dan mesin vektor pendukung digunakan untuk analisis data. Anova digunakan sebagai metode pemilihan fitur untuk menghilangkan fitur yang tidak perlu dan berlebihan. Analisis komponen utama digunakan sebagai teknik transformasi fitur. Kurang tidur secara signifikan meningkatkan gelombang alfa, beta, dan theta dalam sinyal EEG. Durasi mengemudi juga memiliki efek yang signifikan, meningkatkan daya theta dan semua rasio EEG, sambil menurunkan daya beta pada kelompok waspada. Rasio $(\theta + \alpha) / \beta$ dan θ/β menunjukkan peningkatan yang cukup besar pada kelompok SD pada akhir mengemudi. Status tidur dan durasi mengemudi keduanya memengaruhi kantuk subjektif. Sinyal EEG dikombinasikan dengan status tidur dan faktor durasi mengemudi menghasilkan akurasi model yang dapat diterima masing-masing 77,1% dan 90,2% dalam pelatihan dan pengujian, dengan sensitivitas 90,5% dan spesifisitas 90% dalam uji data. Mesin vektor pendukung menunjukkan klasifikasi yang lebih baik daripada regresi logistik, dengan kernel linier sebagai pengklasifikasi terbaik. Daya theta memiliki efek tertinggi dalam model dibandingkan dengan sinyal EEG lainnya (Puspasari dkk., 2023). Analysis of the effect of thermal comfort on driver drowsiness progress with Predicted Mean Vote: An experiment using real highway driving conditions, penelitian ini menemukan bahwa pengemudi paling mengantuk dalam kondisi yang sedikit lebih hangat (PMV + 0,2) dan hampir tidak mengantuk dalam kondisi yang lebih dingin dan hangat setelah 15 menit mengemudi. Karakteristik berbentuk U terbalik ini dimodelkan secara numerik sebagai efek lingkungan termal pada kantuk pengemudi. Pemilihan model berdasarkan Akaike Information Criterion (AIC) menunjukkan bahwa model regresi kuartik memberikan kecocokan terbaik untuk data. Perkembangan kantuk memuncak pada PMV 0,207. Lingkungan termal pribadi, seperti yang ditunjukkan oleh indeks Predicted Mean Vote (PMV), memiliki pengaruh yang lebih besar pada

perkembangan kantuk daripada waktu mengemudi untuk periode mengemudi yang singkat. Penelitian ini memberikan wawasan tentang hubungan antara kenyamanan termal dan kantuk pengemudi di bawah kondisi mengemudi di jalan raya dunia nyata. Namun, hasilnya dibatasi oleh ukuran sampel yang kecil (Sunagawa et al., 2023).

Deteksi kantuk dalam artikel (Qianyi Jiang dkk., 2023) memberikan kontribusi penting dalam menjelaskan pendekatan berbasis video yang menggabungkan dimensi spasial dan temporal secara simultan dalam mendeteksi kantuk pengemudi. Artikel ini memperkenalkan model 2s-STTN (Two-Stream Spatial–Temporal Transformer Network), yang memanfaatkan mekanisme self-attention dalam dua jalur terpisah: *spatial self-attention* untuk mempelajari hubungan antara titik-titik wajah dalam satu frame, dan *temporal self-attention* untuk menangkap dinamika ekspresi wajah antar frame. Representasi spasial dan temporal dari titik-titik landmark wajah dibangun dalam bentuk *spatial-temporal graph*, dan ditangani menggunakan teknologi *graph convolution* serta *class activation mapping (CAM)* guna mengatasi kondisi nyata seperti occlusion (misalnya penggunaan kacamata hitam), pencahayaan redup, dan perubahan pose kepala.

Hasil penelitian lain menunjukkan bahwa metode deteksi kantuk yang diusulkan, yaitu model Vis-Net yang menggabungkan Vision Transformer (ViT) dengan jaringan neural lainnya seperti MobileNet-V2, ResNet152-V2, Inception-V3, DenseNet, dan NASNet, berhasil mendeteksi berbagai tingkat kantuk pengemudi dengan akurasi tinggi. Model ViT mencatatkan akurasi tertinggi yaitu 98,94% dalam mendeteksi tingkat kantuk pengemudi, sementara model lain seperti Vis-MobileNet, Vis-ResNet, Vis-IncepNet, dan Vis-DenNet menunjukkan kinerja yang sangat baik dengan akurasi lebih dari 95%. Namun, model Vis-NASNet memiliki performa lebih rendah dengan akurasi antara 58% hingga 86%. Model ini berhasil mengklasifikasikan kantuk dalam lima level, dengan akurasi hampir 100% untuk level 1 (terjaga), namun akurasi menurun pada level 3 (kantuk ringan). Prediksi untuk level 5 (tertidur) sangat baik dengan akurasi 99% pada ViT. Integrasi kerangka deteksi emosi dalam fase pengujian juga membantu mengurangi waktu

pemrosesan hingga 8%, mempercepat inferensi. Pengujian di kondisi dunia nyata, seperti pencahayaan rendah dan pengemudi yang mengenakan masker atau kacamata, menunjukkan ketahanan model, namun akurasi tetap menurun saat pengemudi menggunakan masker atau kacamata, terutama pada model selain ViT (Phan et al., 2025). Berikut merupakan tabel rujukan peneliti dalam melakukan pengembangan penelitian yang sudah dilakukan sebelumnya, secara detail ditampilkan pada Tabel 2.1.

Tabel 2.1 Tabel Literatur Review

Penulis dan Tahun	**Metode**	**Hasil**
(Adi dkk., 2020)	Teknik pemrosesan citra dan metode Haar Cascade Classifier untuk mendeteksi kantuk dan kehilangan fokus pada driver	Sistem yang dikembangkan menggunakan metode pemrosesan citra mampu memantau kantuk dan kehilangan fokus pada pengemudi dengan akurasi tinggi. Nilai akurasi tertinggi untuk mendeteksi fokus yang hilang pada pengemudi adalah 88,00%, sedangkan nilai akurasi tertinggi untuk mendeteksi kantuk pada pengemudi adalah 90,40%
(Pandey & Muppalaneni, 2021)	Eye aspect ratio (EAR) ,Histogram of Oriented Gradients (HOG) method	Sistem yang dikembangkan mencapai akurasi sekitar 92,5% pada dataset menguap (YawDD) yang digunakan untuk evaluasi. Sistem yang diusulkan ternyata dapat diterapkan pada mesin, kuat, dan dapat diandalkan untuk

Tabel 2.1 Tabel Literatur Review

Penulis dan Tahun	Metode	Hasil
		digunakan. Metode ini dinilai sangat cocok dibandingkan dengan metode fisiologis seperti EEG dan EOG
(Hasan dkk., 2022)	Metodologi melibatkan perekaman sinyal dengan tes kewaspadaan psikomotor (PVT), pra-pemrosesan, ekstraksi fitur, dan penentuan fitur penting dari sinyal fisiologis untuk deteksi kantuk	Hasil penelitian menunjukkan bahwa ukuran fisiologis tunggal menunjukkan pola kinerja tertentu, dengan sensitivitas yang lebih tinggi dan spesifisitas yang lebih rendah atau sebaliknya. Sebaliknya, model berbasis biosignal hibrida memberikan profil kinerja yang lebih baik, mengurangi perbedaan antara sensitivitas dan spesifisitas.
(Qianyi Jiang dkk., 2023)	Two-Stream Spatial–Temporal Transformer Network (2s-STTN	Penelitian ini diuji pada dua dataset, yaitu YAW-DD dan NTHU-DDD. Pada dataset YAW-DD, model ini mencapai akurasi sebesar 94,3% dan F1-score sebesar 0,948. Sedangkan pada dataset NTHU-DDD, model mencatat akurasi rata-rata sebesar 93%, dengan nilai presisi dan true positive rate yang stabil pada berbagai skenario,

Tabel 2.1 Tabel Literatur Review

Penulis dan Tahun	**Metode**	**Hasil**
		termasuk penggunaan kacamata dan pencahayaan malam.
(T.-C. Phan dkk., 2025a)	Model hibrida Vis-Net. Vision transformer	Model Vision Transformer (ViT) mencatatkan akurasi tertinggi, yaitu 98,94% pada data pengujian. Ini menunjukkan bahwa model ViT sangat efektif dalam mendeteksi berbagai tingkat kantuk pengemudi
(Bai dkk., 2022)	Berbasis jaringan konvolusional graf spasial-temporal dua alur, yaitu *Two-stream Spatial–Temporal Graph Convolutional Network (2s-STGCN)*,	Deteksi kantuk pengemudi dataset utama yang digunakan adalah YawDD dan NTHU-DDD, akurasi 93,4% pada YawDD dan 92,7% pada NTHU-DDD, dan hasil uji ablation membuktikan bahwa kombinasi dua alur lebih efektif dibanding penggunaan alur tunggal
(J. Zhao dkk., 2024)	CNN untuk mengekstrak fitur spasial dari setiap frame, kemudian membaginya menjadi patch yang dikodekan secara linear dan diproses melalui *Spatial Transformer Encoder* untuk interaksi spasial serta *Temporal Transformer Encoder* untuk menangkap perubahan fitur	Metode ini terbukti lebih kuat terhadap perubahan cuaca, pencahayaan, serta objek dinamis seperti kendaraan dan pejalan kaki. Walau memiliki kompleksitas yang lebih tinggi dan waktu ekstraksi sedikit lebih lama, model ini tetap efisien dan mampu menghasilkan performa yang lebih stabil dan akurat dalam berbagai kondisi pengujian.

Tabel 2.1 Tabel Literatur Review

Penulis dan Tahun	**Metode**	**Hasil**
	antar-frame dalam jendela geser (*sliding window*)	
(H. Chen dkk., 2021)	Image Processing Transformer (IPT), yang dirancang untuk menyelesaikan berbagai tugas pemrosesan citra tingkat rendah seperti super-resolution, denoising, dan deraining, hanya dengan satu model pre-trained berbasis Transformer	Hasil eksperimen menunjukkan bahwa model IPT secara konsisten mengungguli metode state-of-the-art sebelumnya berbasis CNN seperti RCAN, RDN, dan EDSR. Sebagai contoh, untuk super-resolution skala ×2 pada dataset Urban100, IPT mencetak PSNR 33.76 dB, mengungguli metode terbaik sebelumnya dengan margin >0.4 dB. Untuk tugas denoising dengan noise σ=50, IPT menghasilkan PSNR 29.88 dB pada BSD68,
(Dosovitskiy dkk., 2020)	Transformer tanpa konvolusi	Hasil menunjukkan bahwa ketika ViT dilatih pada dataset besar, ia mampu menyaingi bahkan melampaui performa CNN canggih seperti ResNet dan EfficientNet, dengan akurasi 88.55% pada ImageNet dan 94.55% pada CIFAR-100
(Heo dkk., 2021)	Pooling-based Vision Transformer (PiT) sebagai bentuk pengembangan dari Vision Transformer (ViT)	Hasil eksperimen menunjukkan bahwa PiT mengungguli ViT pada hampir semua metrik: akurasi lebih tinggi dengan FLOPs dan parameter lebih rendah, kecepatan inferensi GPU lebih cepat, dan performa

Tabel 2.1 Tabel Literatur Review

Penulis dan Tahun	Metode	Hasil
		lebih tangguh terhadap gangguan visual dan latar belakang. Misalnya, PiT-S mencapai akurasi 81.9% pada ImageNet dengan distilasi, melampaui ViT-S (81.2%) dan ResNet50 (80.2%)
(Park dkk., 2022; W. Xu dkk., 2021)	CoaT (Co-Scale Conv-Attentional Transformer), yang menggabungkan keunggulan CNN dan Transformer dalam satu kerangka kerja untuk tugas pengenalan citra	Hasil eksperimen menunjukkan bahwa CoaT mencapai performa unggul dengan efisiensi tinggi. Misalnya, CoaT-Lite Small memperoleh akurasi top-1 sebesar 81.9% pada ImageNet-1K dengan hanya 20.0M parameter dan 4.0G FLOPs, mengungguli beberapa model kompetitor yang lebih besar. CoaT juga menunjukkan efisiensi komputasi tinggi dan skalabilitas yang baik untuk berbagai ukuran model.
(Park dkk., 2022)	Fusion ViViT, yaitu sistem end-to-end berbasis Vision Transformer (ViViT) yang dirancang untuk estimasi remote photoplethysmography (rPPG) dari video wajah	Hasil eksperimen menunjukkan bahwa model Fusion ViViT dengan SSL menghasilkan performa kompetitif, dengan RMSE sebesar pada VIPL-HR dan 16.94 pada MR-NIRP-Car. Dalam pengujian transfer learning, model berbasis SSL menunjukkan generalisasi lebih baik dibanding supervised learning, dan transformer-based Fusion ViViT mengungguli model

Tabel 2.1 Tabel Literatur Review

Penulis dan Tahun	Metode	Hasil
		CNN-based seperti PhysNet-late fusion dalam skenario nyata
(J. Wang dkk., 2021)	Spatial-Temporal Token Selection (STTS) sebagai kerangka kerja efisien untuk video transformer	Hasil penelitian menunjukkan bahwa STTS secara konsisten mengurangi beban komputasi lebih dari 30–50% sambil mempertahankan akurasi dengan penurunan kurang dari 1%, bahkan dalam beberapa konfigurasi mampu meningkatkan akurasi
(Krishna dkk., 2022a)	YoloV5 untuk pendeteksian wajah otomatis dan Vision Transformer (ViT) untuk klasifikasi status kantuk	Model ViT dilatih selama 150 epoch dan menghasilkan akurasi pelatihan sebesar 96.2%, validasi 97.4%, dan pengujian pada dataset custom mencapai 95.5%. Evaluasi lebih lanjut menunjukkan bahwa model tetap akurat bahkan dalam kondisi malam hari, meskipun akurasinya sedikit lebih rendah dibanding siang atau sore hari. Dibandingkan dengan pendekatan terdahulu berbasis Haar cascades, CNN LeNet, atau MobileNetV2, model ini terbukti memiliki keunggulan akurasi yang lebih tinggi dan ketahanan lebih baik dalam skenario nyata
(Arnab dkk., 2021)	ViViT (Video Vision Transformer), sebuah arsitektur pure-	ViViT-H/14x2 yang dilatih dengan pretraining dari dataset JFT mencapai Top-1 accuracy sebesar

Tabel 2.1 Tabel Literatur Review

Penulis dan Tahun	**Metode**	**Hasil**
	transformer pertama yang dirancang khusus untuk video classification tanpa mengandalkan konvolusi	84.9% pada Kinetics-400, melampaui semua metode sebelumnya termasuk TimeSformer dan X3D. Pada Epic Kitchens-100, ViViT mengungguli metode lain dalam prediksi aksi, verba, dan nomina secara signifikan. Hasil ablation menunjukkan bahwa factorised encoder (Model 2) menawarkan trade-off terbaik antara akurasi dan efisiensi komputasi
(Z. Wang & Liu, 2024)	explainable AI (XAI) baru bernama STAA (Spatio-Temporal Attention Attribution), yang dirancang khusus untuk menjelaskan keputusan model video berbasis Transformer, seperti TimeSformer dan ViViT	Hasil penelitian menunjukkan bahwa STAA (Enhanced) secara signifikan mengungguli SHAP dan LIME baik dari segi kualitas penjelasan maupun efisiensi. STAA Enhanced mencapai skor faithfulness 0.844 dan monotonicity 0.850, jauh lebih tinggi dibanding SHAP (0.769 / -0.535) dan LIME (0.525 / -0.411)
(Y. Tang dkk., 2025)	PredFormer, yaitu model video prediction berbasis pure transformer tanpa menggunakan *recurrent neural networks (RNN)* maupun *convolutional neural networks (CNN)*	Hasil penelitian menunjukkan bahwa PredFormer mengungguli semua metode state-of-the-art sebelumnya. Moving MNIST, varian Quadruplet-TSST mencatat MSE 12.4, jauh lebih rendah dari SimVP (23.8) dan SwinLSTM (17.7), dengan kecepatan inferensi

Tabel 2.1 Tabel Literatur Review

Penulis dan Tahun	Metode	Hasil
		302 FPS. Pada dataset Human3.6M, PredFormer mencapai MSE 110.9, mengungguli PredRNN++, SimVP, dan OpenSTL-ViT dengan FLOPs jauh lebih kecil. Untuk TaxiBJ, PredFormer menurunkan MSE menjadi 0.277 dan meningkatkan kecepatan hingga 4249 FPS, dibandingkan SimVP yang hanya 533 FPS. Bahkan di WeatherBench, PredFormer mempertahankan akurasi tinggi (MSE 1.100) dan efisiensi, membuktikan efektivitasnya dalam prediksi jangka panjang.
(Yang dkk., 2021)	Klasifikasi aksi wajah halus menggunakan model 3D-LTS (Long-Term Sampling) yang menggabungkan 3D CNN dan bidirectional LSTM.	Akurasi deteksi mencapai hasil terbaik dibanding metode state-of-the-art lainnya. Penggunaan keyframe selection secara signifikan meningkatkan efisiensi dan akurasi deteksi.
(Cabot dkk., 2024)	Penelitian ini mengusulkan deteksi menguap dan mata tertutup berdasarkan *ratio* tinggi/lebar area mata dan mulut, lalu diklasifikasi menggunakan CNN pretrained (Inception-V3,	Model yang telah ditransfer learning menunjukkan performa yang baik untuk deteksi tindakan mata dan mulut. Akurasi mendekati 100% pada training dan 95% pada testing (untuk yawn detection) menggunakan algoritma BOSS pada video yang dinormalisasi.

Tabel 2.1 Tabel Literatur Review

Penulis dan Tahun	Metode	Hasil
	VGG-16, MobileNet, Xception)	
(K. Chen dkk., 2021)	Logistic Regression (LR) dengan dua fitur utama: Mouth Aspect Ratio (MAR) dan Mouth Corner Angle (MCA). BOSS (Bag of SFA Symbols) untuk mengklasifikasikan urutan waktu dari status yawning (dalam bentuk time series).	Hasil penelitian menunjukkan bahwa algoritma yang diusulkan, yaitu LR+BOSS, memberikan performa yang unggul dibandingkan metode lainnya. algoritma LR menunjukkan akurasi sebesar 95% pada data image. Sementara itu, algoritma BOSS mencapai akurasi sempurna 100% pada data pelatihan dan 95% pada data pengujian untuk klasifikasi proses menguap dalam bentuk urutan waktu (time series).
(Rimal dkk., 2023)	Deteksi wajah dilakukan menggunakan MediaPipe FaceMesh yang mengidentifikasi 468 titik landmark wajah. Posisi kepala dihitung melalui metode Perspective-n-Point (PnP) dan Euler angles, sementara status mata dan mulut dievaluasi menggunakan rasio ADR dan MAR. Sistem juga menerapkan ambang	Hasil pengujian sistem di komputer menunjukkan rata-rata F1-score sebesar 0,91 dengan kecepatan eksekusi mencapai 32 hingga 60 fps, tergantung resolusi. Kesalahan utama terjadi ketika pengemudi menutupi mulut saat menguap atau ketika ekspresi tertawa terdeteksi sebagai menguap. Pada Raspberry Pi 3, sistem tetap menunjukkan kinerja tinggi dengan F1-score hingga 0,92 dan kecepatan eksekusi 24–26 fps, meskipun dilakukan penghematan proses dengan melewati beberapa frame.

Tabel 2.1 Tabel Literatur Review

Penulis dan Tahun	Metode	Hasil
	batas adaptif selama 10 detik awal untuk menyesuaikan dengan struktur wajah individu.	
(Kielty dkk., 2023)	MobileNetV2 untuk ekstraksi fitur, Self-attention module untuk fokus kontekstual, 2-layer bidirectional LSTM sebagai kepala model untuk menangkap pola temporal.	Sistem deteksi menguap berbasis event camera menunjukkan performa tinggi dengan skor F1 sebesar 95,3% pada data internal dan 90,4% pada dataset publik YawDD, meski tanpa pelatihan langsung. Waktu inferensi rata-rata 0,44 detik per 100 frame memungkinkan pengoperasian real-time. Visualisasi peta perhatian menunjukkan fokus model tetap pada area mulut, bahkan saat tertutup tangan, menandakan kemampuan adaptif terhadap variasi ekspresi wajah.

Kebaruan penelitian ini terletak pada Pengembangan arsitektur Factorized Video Vision Transformer (ViViT) sebagai arsitektur utama untuk deteksi kantuk pengemudi berbasis video, sehingga model tidak hanya menangkap fitur spasial pada tiap frame, tetapi juga dinamika temporal antar frame. Berbeda dengan pendekatan konvensional yang umumnya memproses fitur wajah secara terpisah atau hanya mengandalkan data citra statis, penelitian ini melakukan inovasi metodologi dengan mengadaptasi kemampuan *self-attention* pada ViViT untuk menangkap dinamika perubahan fitur visual antar-frame secara simultan.

Orisinalitas riset ini semakin dipertegas melalui representasi fitur terintegrasi yang mensinergikan tiga indikator perilaku utama, yakni kondisi mata, aktivitas mulut (menguap), dan kemiringan kepala. Integrasi ketiga elemen ini dalam sebuah *pipeline* deteksi berbasis video mulai dari ekstraksi menggunakan metode HRNet hingga proses tokenisasi *patch* menciptakan sistem yang lebih kokoh (*robust*) terhadap variasi visual dan perilaku pengemudi. Hal ini membuktikan bahwa model yang diusulkan tidak hanya unggul secara teoretis pada dataset NTHU-DDD, tetapi juga memiliki relevansi praktis yang kuat untuk diimplementasikan sebagai solusi keselamatan berkendara yang bersifat non-intrusif dan *real-time*.

2.2 Keaslian Penelitian

Dari penelitian yang telah dilakukan sebelumnya, maka penelitian ini memiliki keaslian yaitu :

1. Penggunaan metode ViViT (Video Vision Transformer) untuk deteksi kantuk, menggantikan metode konvensional CNN yang terbatas dalam menangkap dinamika temporal video.
2. Integrasi simultan dua fitur *visual non-intrusif*, yaitu:
   a. Kondisi mata (frekuensi kedipan, durasi penutupan).
   b. Kemiringan kepala (*head pose*), ke dalam satu sistem deteksi kantuk berbasis ViViT yang belom banyak diteliti secara gabungan dalam studinya.
   c. Mulut menguap sebagai fitur tambahan indikator kantuk.
3. Implementasi sistem berbasis video real-time, memungkinkan deteksi progresif kantuk dari urutan video yang terus berjalan, bukan hanya dari frame statis.
4. Kontribusi pada sistem deteksi perilaku, sebagai alternatif yang lebih aman dan praktis dibandingkan pendekatan berbasis fisiologis atau sensor kendaraan, dengan tetap mempertahankan akurasi tinggi.

2.3 Landasan Teori

2.3.1 Definisi Kantuk

Kantuk dalam perspektif neurofisiologi, merupakan keadaan pasif di mana aktivitas listrik otak beralih dari gelombang beta/alpha dominan saat terjaga menjadi gelombang theta (4–7 Hz), mencerminkan penurunan kewaspadaan dan kesiapsiagaan kognitif. Arif dkk (2023) menunjukkan bahwa spectral EEG khususnya peningkatan kekuatan gelombang theta dan delta serta penurunan rasio alpha beta merupakan indikator paling andal untuk mengidentifikasi kantuk pada pengemudi, dengan akurasi tinggi dalam skenario nyata . Analisis ini didukung oleh metode non-invasif seperti EEG dan fNIRS, yang menawarkan resolusi temporal dan spasial yang cukup untuk merekam perubahan neurofisiologis yang cepat saat kantuk(Arif dkk., 2023) .

Kelelahan dapat diartikan sebagai dorongan biologis untuk melakukan istirahat dalam rangka pemulihan kondisi (Prabaswara, 2013). Saat seseorang sedang mengalami kelelahan akan terjadi dorongan untuk tidur atau beristirahat atau yang sering disebut kantuk. Kelelahan atau fatigue memiliki dampak yang negatif. Keadaan fatigue yang menimbulkan rasa kantukdapat mengurangi kekuatan, kecepatan, kecepatan reaksi, kemampuan kordinasi, keseimbangan, dan juga kemampuan seseorang dalam mengambil keputusan. Kelelahan dapat disebabkan oleh berbagai hal.Secara umum penyebab kelelahan adalah penggunaan tenaga fisik atau mental yang berkepanjangan tanpa cukup waktu untuk beristirahat dan memulihkan diri (Dawson dkk., 2011).

Secara psikofisiologis kantuk juga berkaitan dengan penurunan perhatian dan respons terhadap rangsangan eksternal, yang dapat diukur lewat preferensi frekuensi gelombang otak dan reaktivitas terhadap stimulus visual atau auditori . Studi Stancin dkk. (2021) meninjau penggunaan berbagai fitur EEG seperti power spectral density, entropi, dan variabilitas temporal – untuk mendeteksi kantuk pada pengemudi, dan menemukan bahwa kombinasi fitur linear dan non-linear meningkatkan akurasi klasifikasi sampai mendekati 95% . Hal ini mendukung pendekatan multimodal dalam memahami kantuk, dengan

EEG sebagai *ground truth* fisiologis yang memverifikasi indikator fisik seperti kelopak mata dan kemiringan kepala (Stancin dkk., 2021).

2.3.2 Ekspresi Kantuk

Kantuk (*drowsiness*) ialah keadaan dimana seseorang ingin tidur. Namun kondisi kantuk yang tidak tepat dapat mengakibatkan hal yang fatal. Mengantuk dapat disebakan oleh beberapa faktor antara lain: kelelahan bekerja, kurangnya tidur yang cukup. Kondisi mengantuk terdapat beberapa macam sehingga dapat dikategorikan seseorang sedang mengantuk atau tidak. Kondisi mengantuk seseorang antara lain dapat dilihat dari kondisi kelopak mata mulai berat, pandangan kabur serta kepala mulai tidak seimbang menahan beban sehingga mengharuskan untuk berbaring dan istirahat. Hal tersebut akan berakibat fatal jika dialami oleh pengemudi. Pada tabel 2.2 menunjukkan *Karonlinska Sleepiness Scale* (KSS) yang merupakan skala tingkatan kantuk seseorang (Pauly & Sankar, 2016).

Tabel 2.2 Karonlinska Sleepiness Scale (Pauly & Sankar, 2016)

No	*Karonlinska Sleepiness Scale (KSS)*
1	*Extremely alert*
2	*Very Alert*
3	*Alert*
4	*Rather Alert*
5	*Neither Alert nor sleepy*
6	*Some sight of sleepiness*
7	*Sleepy, but not difficulty remaining awake*
8	*Sleepy, some effort to keep alert*
9	*Extremely sleepy, fighting sleep*

Salah satu pendekatan untuk pendeteksian kantuk adalah dengan behavioral apporach . Pendekatan ini mendeteksian kantuk seseorang melalui keadaan mata, posisi kepala dan menguap . Manusia dikatakan mengantuk jika

keadaan matanya tertutup selama kurang lebih 1 detik (Khan, 2014). Berdasarkan KSS, skala tingkat 7 merupakan kondisi dimana manusia mulai merasakan kantuk tetapi tidak membutuhkan usaha yang berat untuk tetap terjaga. Kondisi terebut dapat dikatakan sebagai kondisi lelah. Manusia dalam kodisi lelah ditandai dengan menguap. Menguap adalah suatu keadaan diamana mata dalam keadaan tertutup atau setengah terbuka disertai dengan mulut yang terbuka lebar selama kurang lebih 1 detik (Azim dkk., 2009)

Penelitian ini diawali dengan upaya menilai tingkat kantuk pengemudi melalui pengamatan ekspresi wajah yang terekam, yang kemudian dinilai menggunakan skala kantuk standar. Peneliti mengeksplorasi indeks prediktor efektif seperti frekuensi kedipan, durasi penutupan mata, serta perubahan ekspresi wajah lain yang berkaitan dengan munculnya rasa kantuk. Data diperoleh melalui eksperimen simulasi mengemudi dalam kondisi terkontrol, memungkinkan pemantauan perilaku visual secara detail. Hasil penelitian menunjukkan bahwa indikator visual seperti peningkatan frekuensi kedipan dan durasi mata tertutup memiliki korelasi yang signifikan dengan tingkat kantuk pengemudi yang diukur secara subyektif, menjadikannya prediktor potensial untuk sistem deteksi kantuk otomatis. Kelebihan penelitian ini adalah fokusnya pada ekspresi wajah yang dapat dipantau secara non-invasif, namun keterbatasan utamanya adalah pengujian hanya dilakukan dalam skenario simulasi, sehingga diperlukan validasi lebih lanjut di lingkungan mengemudi nyata . Skala kantuk yang digunakan dibagi menjadi 5 tingkat (*point scale*) yang dinilai berdasarkan rekaman video wajah pengemudi (Kitajima dkk., 1997):

1. Tingkat 1 – Waspada (*alert)*, tidak ada tanda kantuk: mata terbuka normal, tidak ada penurunan frekuensi kedipan.
2. Tingkat 2 – Sedikit mengantuk: mulai menunjukkan peningkatan frekuensi kedipan, namun tetap responsif.
3. Tingkat 3 – Mengantuk sedang: kedipan mata melambat, durasi penutupan mata mulai terlihat lebih lama, wajah tampak kurang ekspresif.

4. Tingkat 4 – Sangat mengantuk: frekuensi kedipan tinggi, durasi mata tertutup >1 detik, terkadang kepala mulai terangguk.
5. Tingkat 5 – Tertidur sesaat (*microsleep*): mata tertutup lama, hilang kontak visual dengan jalan, kepala terangguk jelas.

Penelitian ini, tingkat kantuk pengemudi diklasifikasikan berdasarkan standar *Karolinska Sleepiness Scale* (KSS) yang telah dikelompokkan kembali guna menyesuaikan dengan karakteristik data visual. Skor KSS yang semula memiliki rentang satu hingga sembilan didistribusikan ke dalam empat kategori kelas utama berdasarkan kesamaan kondisi klinis dan perilaku yang ditunjukkan oleh subjek. Pendekatan pengelompokkan ini diadaptasi dari metode yang diusulkan oleh Ghoddoosian (Ghoddoosian dkk., 2019), namun dilakukan pengembangan lebih lanjut dengan menyertakan seluruh skala guna menangkap gradasi transisi kondisi pengemudi secara lebih detail. Pembagian empat kelas ini meliputi: kondisi terjaga penuh (Kelas 1), penurunan kewaspadaan atau fase transisi (Kelas 2), kondisi mengantuk (Kelas 3), hingga fase kantuk berat yang kritis (Kelas 4). Klasifikasi hierarkis ini bertujuan untuk melatih model ViViT agar mampu mengenali perubahan fitur spasio-temporal pada wajah pengemudi secara lebih presisi di setiap fasenya.

2.3.3 Pengolahan Citra Digital Wajah Dalam Sistem Komputer

Pengolahan citra wajah dimulai dengan deteksi wajah, yaitu proses otomatis yang melokalisasi area wajah dalam citra. Beberapa metode klasik, seperti algoritma Viola–Jones, menggunakan fitur Haar dan AdaBoost untuk mendeteksi wajah secara real-time pada citra (Viola & Jones, 2001) . Penelitian modern telah mengintegrasikan deep learning dalam tahap deteksi, seperti pendekatan Faster R-CNN yang diperkuat dengan PCA untuk meningkatkan akurasi pada dataset nyata(Muhammad dkk., 2021). Tahapan ini memiliki peran penting sebagai dasar untuk proses selanjutnya berupa prapemrosesan dan ekstraksi fitur. Tahap ekstraksi fitur menjadi inti sistem pengenalan wajah. Metode awal seperti PCA (eigenfaces) dan LDA (fisherfaces) berfokus pada reduksi dimensi dan representasi statistik wajah dalam ruang eigen, yang telah

diuji dalam berbagai dataset seperti ORL dan Yale-B(Eleyan & Demirel, 2006) . Teknik tekstur seperti LBP juga digunakan untuk menangkap pola lokal pada wajah, dan dikombinasikan dengan CNN untuk menambah kekokohan terhadap variasi pencahayaan dan ekspresi(J. Tang dkk., 2020) . Tren terkini mengandalkan arsitektur deep CNN, seperti FaceNet dan DeepFace, yang menghasilkan embedding vektor berdimensi tinggi (128 atau 4096) dengan teknik loss seperti triplet loss dan margin-based softmax, mencapai performa hingga ~99 % akurasi pada LFW dan YouTube Faces (William dkk., 2019) .

Setelah fitur diekstraksi tahap matching dan pengenalan wajah dilakukan dengan membandingkan embedding dengan basis data menggunakan metric seperti jarak Euclidean atau cosine similarity. Kualitas sistem sangat dipengaruhi oleh prapemrosesan seperti normalisasi pose (face alignment) dan super-resolution citra rendah . Selain itu, aspek keamanan seperti anti-spoofing menjadi penting untuk mencegah pemalsuan (misalnya menggunakan foto atau deepfake), yang ditangani melalui analisis tekstur serta pola frekuensi temporal oleh model berbasis CNN . Kombinasi komponen-komponen ini deteksi, praproses, ekstraksi fitur, matching, dan keamanan membentuk sistem pengenalan wajah modern yang andal dan adaptif terhadap tantangan dunia nyata. Tahap ekstraksi fitur merupakan inti dari pengolahan citra digital wajah. Pendekatan klasik seperti PCA (eigenface), LDA (fisherface), HOG, atau LBP menganalisis struktur statistik dan tekstur wajah dalam bentuk vektor berdimensi rendah . Namun, sejak tahun 2014, penggunaan deep learning, khususnya CNN, telah merevolusi ekstraksi fitur secara otomatis dan hierarkis, memberikan representasi wajah yang jauh lebih kaya dan efektif.

Teknologi digitalisasi dapat membantu dalam bidang transportasi modern mengoptimalkan kontribusi ekonomi, mengurangi pekerjaan, serta mengatasi solusi untuk area yang terisolasi. Sekarang ada metode yang mengurangi eksperimen kontak fisik dengan hewan untuk meningkatkan kenyamanan hewan dan menghindari wabah penyakit. Metode ini mulai banyak digunakan, menggunakan sensor, computer vision data video, big data,

dan teknologi blockchain untuk keuntungan bersama antara produsen ternak, konsumen, dan hewan ternak itu sendiri(Neethirajan & Kemp, 2021). Citra digital dikirimkan secara elektronik, file ditransfer ke media penyimpanan fisik, identifikasi obyek berdasarkan foto dan video adalah salah satu layanan yang bisa digunakan(Taoufik dkk., 2022). Proses ekstraksi citra digital ada beberapa tahap, diantaranya meliputi citra digital diambil setiap frame, kecepatan pengambilan adalah 30 frame per detik dan setiap frame dibaca oleh algoritma yang dikembangkan, rubah citra Red, Green, Blue (RGB) ke Hue Saturation Value (HSV). Citra yang ditangkap oleh webcam dalam format piksel RGB, Ambang citra HSV digunakan untuk mendapatkan ambang batas, merupakan metode untuk memisahkan daerah yang lebih tinggi atau lebih rendah dari nilai ambang batas yang telah ditetapkan berdasarkan nilai batas atas dan batas bawah. Ambang batas diterapkan untuk semua piksel citra dengan mengubah citra HSV menjadi citra biner. Citra biner merupakan satu set array yang memiliki nilai 0 dan 1, tanda bitwise sangat berguna saat mengekstrak bagian mana pun dari citra, yang termasuk operasi bitwise adalah and, or, not dan xor. Citra kontur dengan warna hijau sebagai ambang batas, kontur merupakan kurva yang menghubungkan semua titik secara kontinyu, yang memiliki warna atau intensitas yang sama. Dalam menghitung piksel per rasio matrik, untuk menentukan sifat geometris objek pada suatu citra, terlebih dahulu kita perlu melakukan kalibrasi menggunakan objek referensi. Objek referensi tidak lain adalah objek yang tinggi dan lebarnya akan dihitung (Gupta dkk., 2023)

Pengesahan wajah dimulai dengan tahap deteksi wajah, yakni identifikasi dan pelokalan area wajah dalam citra. Metode awal menggunakan teknik berbasis fitur seperti warna kulit, tepi, dan bentuk wajah, atau pola frekuensi seperti DCT dan morfologi matematika . Tahap ini bertujuan mengekstrak *region of interest* (ROI) yang menjadi input untuk serangkaian proses selanjutnya. Citra wajah yang telah dideteksi diproses pada fase prapemrosesan untuk memperbaiki kondisi citra. Teknik seperti normalisasi pencahayaan, pengurangan noise, dan perataan pose 2D–3D (frontalization)

kerap digunakan; contohnya metode 2D/3D alignment pada DeepFace yang mampu mengubah pose ke arah frontal secara akurat.

Dalam *face recognition*, CNN tidak hanya dipakai sebagai ekstraktor fitur, melainkan juga sebagai model end-to-end untuk klasifikasi wajah. Model seperti DeepFace dengan sembilan lapisan neural network menghasilkan vektor fitur (dimension 4096) yang selanjutnya dicocokkan menggunakan jarak Euclidean atau loss berbasis softmax terdiscriminatif(M. Wang & Deng, 2018) . Arsitektur dan fungsi loss modern semakin berkembang, misalnya margin-based softmax dan regularisasi, guna meningkatkan kapasitas diskriminatif sistem .

Kendala nyata dalam algoritma pengenalan wajah antara lain meliputi variasi pencahayaan, pose (sampai pandangan samping), ekspresi, occlusion, dan resolusi rendah. Survey terbaru menunjukkan adanya peningkatan algoritma super-resolution wajah (*face super-resolution*) untuk memperbaiki kualitas citra sebelum ekstraksi fitur . Selain itu, metode pose-invariant recognition dirancang agar sistem tetap efektif walau wajah tidak tampak frontal . Evaluasi kualitas citra (*face image quality assessment*) juga menjadi aspek penting karena kualitas data citra sangat mempengaruhi performa akhir sistem. Penelitian terbaru menyoroti kebutuhan model yang mampu menilai utilitas biometrik citra wajah secara otomatis dan integratif ke dalam pipeline pengenalan(Schlett dkk., 2022). Di samping itu, aspek anti-spoofing seperti pendeteksian wajah palsu atau manipulasi (*deepfake/forgery*) semakin mendapatkan perhatian. Teknik berbasis deep learning mendeteksi artefak tekstur, frekuensi, dan pola temporal untuk menghindari serangan identitas . Secara keseluruhan, rangkaian algoritma pengolahan citra wajah mulai deteksi, prapemrosesan, ekstraksi fitur, hingga matching dan keamanan menggunakan pendekatan klasik dan modern. Revolusi dimulai dari PCA/LDA dan berkembang pesat lewat CNN dan deep learning, memacu akurasi sistem ke level state-of-the-art, terutama di lingkungan dunia nyata yang menantang. Faktor-faktor seperti penyesuaian pose, kualitas citra rendah, dan ancaman

spoofing menjadi area riset aktif untuk meningkatkan keandalan sistem secara menyeluruh .

Pengolahan citra wajah dalam sistem komputer merupakan bagian dari bidang pengolahan citra digital yang berfokus pada analisis dan transformasi visual dari wajah manusia. Proses ini mencakup serangkaian langkah seperti peningkatan kualitas citra, penghilangan noise, normalisasi pencahayaan, dan koreksi geometri untuk menghasilkan citra wajah yang lebih representatif sebelum digunakan dalam sistem pengenalan atau analisis ekspresi wajah. Tahapan ini dapat dilakukan tanpa melibatkan identifikasi individu secara eksplisit, tetapi lebih ditujukan untuk meningkatkan kualitas visual data wajah itu sendiri agar sistem selanjutnya bekerja secara optimal (Jiang & Chen, 2008) .

2.3.4 Estimasi Kemiringan Kepala

Kemiringan kepala merupakan salah satu indikator visual penting yang digunakan dalam sistem deteksi kantuk karena secara fisiologis mencerminkan penurunan tingkat kewaspadaan seseorang. Dalam penelitian yang dilakukan oleh Vinay Shashidhar dan timnya (2022), kemiringan kepala dianalisis melalui sudut orientasinya terhadap sumbu vertikal dengan memanfaatkan 68 titik landmark wajah dari pustaka dlib. Penelitian ini secara khusus mengukur dua jenis kemiringan: *tilt*, yaitu kemiringan kepala ke samping kiri atau kanan, serta *nodding*, yaitu gerakan menunduk ke depan secara perlahan atau tiba-tiba. Kedua jenis gerakan ini dianggap sebagai gejala awal kantuk karena saat tingkat kewaspadaan menurun, kontrol otot leher melemah sehingga kepala secara tidak sadar cenderung terangguk atau condong ke satu sisi. Dalam sistem yang mereka rancang, diterapkan ambang batas sudut tertentu misalnya lebih dari 15 derajat untuk tilt dan lebih dari 20 derajat untuk nodding yang dijadikan parameter untuk mendeteksi kantuk. Ketika kepala melampaui sudut tersebut dalam durasi tertentu atau dengan frekuensi yang tinggi, sistem akan memberikan sinyal bahwa subjek berada dalam kondisi mengantuk. Pendekatan ini terbukti efektif dalam mendeteksi perubahan postur kepala secara real-time dan non-invasif, meskipun tetap memiliki keterbatasan

terhadap variabilitas posisi duduk, bentuk wajah, atau penggunaan aksesoris kepala yang dapat memengaruhi akurasi estimasi sudut (Shashidhar dkk., 2024) .

Sistem klasifikasi tahapan tidur berdasarkan sensor percepatan kepala (*head acceleration sensor*) tiga sumbu. Data dikumpulkan dari sensor yang dipasang di kepala pengguna selama tidur, lalu dianalisis menggunakan machine learning untuk mengklasifikasikan tahapan seperti bangun, *Rapid Eye Movement* (REM), tidur ringan, dan tidur dalam. Penelitian ini melakukan pendekatan berbasis getaran kepala (*ballistocardiogram*) yang ringan dan murah dibandingkan polisomnografi. Akan tetapi, metode ini kurang akurat dalam mendeteksi tahap REM, dan sensitif terhadap noise atau gerakan ekstrem kepala. Meskipun demikian, hasil penelitian cukup menjanjikan dengan akurasi keseluruhan mencapai 74,6%, membuka peluang aplikasi portable monitoring tidur di rumah (Yoshihi dkk., 2021).

2.3.5 Computer Vision

Computer vision adalah bidang ilmu interdisipliner yang membahas bagaimana model komputasi dapat digunakan untuk mendapatkan pemahaman dari citra atau video, sehingga membangun suatu sistem seperti yang dapat dilakukan oleh sistem visual manusia (Choi dkk., 2018). Kombinasi machine learning dan deep learning dengan computer vision mampu meningkatkan pemahaman komputer, seperti menyederhanakan proses deteksi dan prediksi. Computer Vision telah dikembangkan untuk secara otomatis mengekstrak fitur kompleks dan terus menerus belajar dari data training (Aziz dkk., 2021).

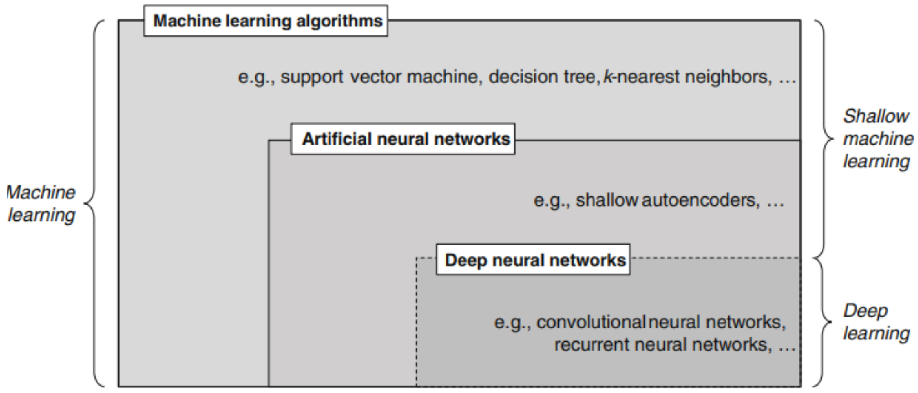
Keuntungan dari computer vision terletak pada pemantauan hewan noninvasif dan berbiaya rendah. Dengan demikian, sistem ini memungkinkan ekstraksi informasi dengan gangguan eksternal minimal (misalnya, penyesuaian sensor oleh manusia) dengan biaya yang terjangkau. Komponen utama dari sistem computer vision termasuk kamera, unit perekam, unit pemrosesan, dan model (Li, Huang, dkk., 2021).

Persepsi penglihatan manusia di adopsi dengan nama computer vision, persepsi visual pada dasarnya adalah tindakan mengamati dan mengenal pola serta 30objek melalui penglihatan atau input visual, secara sederhana sistem penglihatan manusia terdiri dari sensor atau mata untuk menangkap citra dan otak untuk memproses serta menafsirkan citra. Sistem yang ada di komputer merupakan refleksi dari sistem penglihatan visual yang ada pada manusia, untuk dapat meniru sistem tersebut di butuhkan dua komponen pokok yang pertama sensing device sebagai fungsi meniru mata pada manusia dan yang kedua interpreting device yang berfungsi sebagai otak.

2.3.6 Deep Learning

Deep Learning (DL), juga disebut pembelajaran hierarkis dan pembelajaran terstruktur dalam, adalah bagian dari machine learning yang mengeksploitasi penggunaan banyak lapisan pemrosesan informasi non-linear untuk ekstraksi dan transformasi fitur secara supervised maupun unsupervised, serta analisis pola dan klasifikasi (Purwins dkk., 2019).Dalam beberapa tahun terakhir, model Deep Learning telah mendapatkan momentum dalam masalah pencitraan seperti resolusi super gambar dan video, restorasi gambar atau pewarnaan ulang (López-Tapia dkk., 2021).

Deep Learning sangat berguna dalam domain dengan data berdimensi besar dan tinggi, itulah sebabnya jaringan saraf dalam mengungguli algoritma machine earning untuk sebagian besar aplikasi di mana data teks, gambar, video, ucapan, dan audio perlu diproses (Janiesch dkk., 2021). Konsep dan kelas deep learning dapat dilihat pada Gambar 2.3.



Gambar 2.1 Konsep dan Kelas *Deep Learning* (Janiesch dkk., 2021)

2.3.7 Transformer

Transformer merupakan arsitektur model yang pertama kali diperkenalkan dalam artikel Attention Is All You Need (Vaswani dkk., 2017). Model ini mengubah paradigma dalam pengolahan urutan data (*sequence processing),* terutama dalam penerjemahan mesin (*machine translation*) dan aplikasi Natural Language Processing (NLP). Sebelumnya, model-model berbasis Recurrent Neural Networks (RNN), seperti LSTM (Long Short-Term Memory) dan GRU (Gated Recurrent Unit), serta Convolutional Neural Networks (CNN) lebih dominan dalam pengolahan urutan data. Transformer memperkenalkan pendekatan yang sepenuhnya bergantung pada self-attention untuk menangkap ketergantungan dalam urutan data. Sebelum adanya Transformer, arsitektur utama yang digunakan dalam *sequence transduction* (mengubah satu urutan menjadi urutan lainnya) adalah model Recurrent Neural Networks (RNN). RNN menggunakan mekanisme rekursif di mana output dari setiap waktu bergantung pada input saat ini dan informasi tersembunyi dari waktu sebelumnya. Hal ini memungkinkan RNN untuk memproses urutan data secara bertahap, namun memiliki kelemahan dalam menangkap ketergantungan jarak jauh karena masalah vanishing gradient. Untuk mengatasi masalah ini, LSTM (Long Short-Term Memory) (Hochreiter & Schmidhuber, 1997). LSTM mengatasi masalah vanishing gradient dengan memperkenalkan gated mechanisms yang memungkinkan memori untuk mempertahankan informasi lebih lama. Meskipun LSTM banyak digunakan dalam aplikasi NLP

dan penerjemahan mesin, mereka masih memerlukan proses sekuensial, yang membatasi paralelisasi dalam pelatihan dan menyebabkan waktu pelatihan yang lebih lama.

Salah satu ide yang membawa kita lebih dekat ke Transformer adalah attention mechanism. Attention digunakan untuk memfokuskan model pada bagian tertentu dari input yang relevan, terlepas dari jarak posisional. Memungkinkan model untuk menangkap ketergantungan panjang lebih baik daripada model berbasis RNN(Israr dkk., 2023), yang memperkenalkan cara baru untuk menangkap ketergantungan dalam penerjemahan mesin. Meskipun model berbasis RNN + Attention memberikan hasil yang lebih baik daripada model murni RNN, masih ada kekurangan dalam hal paralelisasi dan efisiensi komputasi.

2.3.8 Vision Transformer

Transformer yang awalnya dikembangkan untuk tugas pemrosesan bahasa alami (NLP), telah terbukti sangat efektif dalam pengolahan data visual, seperti pengenalan citra dan deteksi objek. Konsep dasar dari Transformer adalah mekanisme *self-attention*, yang memungkinkan model untuk memproses setiap bagian data secara paralel dan menangkap hubungan jangka panjang antara elemen-elemen dalam data input. Ini berbeda dengan pendekatan berbasis konvolusi yang biasanya memproses data secara lokal dalam grid yang lebih kecil. Dalam konteks visual, kemampuan Transformer untuk menangkap dependensi global pada citra sangat berharga, terutama pada citra berresolusi tinggi dan dalam aplikasi yang membutuhkan pemahaman tentang konteks yang lebih luas dalam image.

Self-attention merupakan inti dari arsitektur Transformer. Mekanisme ini bekerja dengan memberi bobot pada setiap bagian input untuk menentukan seberapa penting informasi tersebut dalam konteks seluruh input. Sebagai contoh, dalam pemrosesan citra, Transformer memecah citra menjadi *patches* atau potongan-potongan kecil dan memperlakukan setiap *patch* sebagai token yang diolah dalam urutan. Setiap *patch* ini kemudian "berkomunikasi" dengan

patch lainnya melalui mekanisme self-attention untuk menghasilkan representasi yang lebih kompleks dari citra tersebut. Transformer dapat menangkap informasi lokal dan global dalam satu model, yang menjadikannya sangat unggul dibandingkan dengan model konvolusional (CNN) dalam tugas yang melibatkan pemahaman citra secara lebih mendalam (Pan dkk., 2021).

Vision Transformer (ViT) diperkenalkan sebagai penerapan langsung dari Transformer pada tugas pengolahan citra. ViT memecah citra menjadi *patches* yang lebih kecil dan memperlakukan mereka sebagai urutan token, mirip dengan bagaimana Transformer diterapkan pada teks. Keuntungan utama dari ViT adalah kemampuannya untuk menangkap dependensi global dalam citra yang lebih kompleks, yang sangat berguna untuk tugas-tugas seperti klasifikasi citra dan deteksi objek. Namun, penerapan ViT juga menghadapi tantangan, terutama dalam hal kebutuhan akan dataset besar untuk pelatihan. ViT membutuhkan lebih banyak data dan waktu pelatihan dibandingkan dengan CNN tradisional untuk dapat bersaing dalam akurasi, terutama pada dataset yang lebih kecil. Meski demikian, dengan pre-training pada dataset besar seperti ImageNet atau JFT-300M, ViT menunjukkan hasil yang sangat kompetitif, bahkan mengungguli CNN pada banyak tugas pengolahan citra (Bin Mohamad Azmi & Kamaru Zaman, 2024) .

Penggunaan Transformer dalam pengolahan data EEG untuk mendeteksi kelelahan pengemudi. Model yang diusulkan TFormer, menggabungkan kemampuan Transformer untuk menangkap pola global dalam domain waktu dan frekuensi. TFormer menggunakan tiga komponen utama: konvolusi untuk penyematan input, perhatian silang multi-head (TF-MCA) untuk mengintegrasikan pola domain waktu ke dalam titik frekuensi, dan perhatian diri untuk mempelajari pola global lebih lanjut. Hasil eksperimen menunjukkan keunggulan TFormer dibandingkan dengan metode yang ada dalam mendeteksi kelelahan pengemudi menggunakan data EEG, terutama dalam pengaturan lintas subjek (R. Li dkk., 2024). Metode berbasis Transformer untuk estimasi posisi kepala yang efisien. Pendekatan ini menggabungkan pembelajaran token orientasi yang dapat menangkap

hubungan antar bagian wajah dalam citra meskipun terjadi occlusion (penutupan sebagian wajah). Dengan menggunakan Transformer, model ini belajar hubungan geometris antar bagian wajah dan dapat mengatasi masalah perubahan besar dalam orientasi kepala. TokenHPE menghasilkan kinerja terbaik dalam mengklasifikasikan orientasi kepala di berbagai dataset dengan akurasi tinggi (C. Zhang dkk., 2023) .

Penggunaan Depth and Visual Concepts Aware Transformer (DEVICE) untuk tugas captioning citra berbasis OCR. DEVICE memanfaatkan informasi kedalaman untuk memperbaiki pemahaman tiga dimensi dalam pengenalan teks pada citra. Metode ini mengintegrasikan dua modul utama, yaitu Depth-enhanced Feature Updating yang memperbarui fitur OCR dengan kedalaman, dan Semantic-guided Alignment, yang mengarahkan penggunaan informasi visual untuk meningkatkan akurasi dalam menghasilkan deskripsi teks. Dengan menggabungkan kedalaman dan konsep visual, model ini dapat menghasilkan caption yang lebih lengkap dan akurat. DEVICE berhasil mengatasi kelemahan dalam captioning citra OCR yang hanya mengandalkan hubungan dua dimensi, seperti yang dilakukan model-model sebelumnya. Dengan menambahkan kedalaman dan pemahaman konsep visual, DEVICE mampu membangun hubungan spasial dan geometris tiga dimensi antara objek dan teks, yang meningkatkan akurasi caption. Pada pengujian menggunakan dataset TextCaps, DEVICE menunjukkan peningkatan signifikan dalam performa, mengungguli model-model sebelumnya dalam menghasilkan caption yang lebih tepat dan komprehensif (D. Xu dkk., 2025).

Metode FFTran suatu jaringan berbasis Transformer untuk tugas klasifikasi citra multi-label. FFTran memanfaatkan dua teknik utama: Multi-level Scale Information Integration Mechanism (MSIIM) untuk mengintegrasikan fitur dari berbagai skala dan Intra-Image Spatial-Channel Semantic Mining (ISCM) untuk mengekstrak informasi spasial dan kanal yang relevan. Dengan menggabungkan kedua teknik ini, FFTran dapat menangani objek kecil yang sulit dikenali, yang sering menjadi tantangan pada klasifikasi citra multi-label. Teknik MSIIM memungkinkan fusi fitur multi-skala, yang

membantu dalam menangani objek dengan ukuran yang sangat berbeda dalam citra. Sementara itu, ISCM membantu meningkatkan pemahaman model terhadap hubungan spasial dan kanal yang ada dalam citra, memungkinkan klasifikasi yang lebih akurat. FFTran menunjukkan peningkatan yang signifikan dalam akurasi klasifikasi citra, terutama pada objek-objek kecil, yang sering kali terabaikan oleh model-model klasifikasi tradisional. Pada dataset yang terkenal, seperti VOC2007 dan COCO2014, FFTran mengungguli model-model state-of-the-art yang ada, membuktikan keunggulannya dalam menangani klasifikasi multi-label yang kompleks. Model ini memperlihatkan kemampuan luar biasa dalam mengatasi variasi ukuran objek yang ada di citra, serta memperbaiki kinerja pada situasi di mana objek kecil atau tersembunyi sering kali tertinggal dalam klasifikasi (P. Liu dkk., 2025) .

Temporal-Channel Visual Transformer (TCVT), sebuah model yang dirancang untuk tugas video summarization. TCVT mengintegrasikan informasi spasial dan temporal melalui dua komponen utama: dual-stream embedding module dan inter-frame encoder. Dual-stream embedding module memungkinkan model untuk menggabungkan fitur visual dan gerakan optik, sedangkan inter-frame encoder menangani korelasi antar-frame dengan menggunakan beberapa modul perhatian temporal dan kanal. Pendekatan ini memungkinkan TCVT untuk memahami hubungan temporal dan spasial antara frame-frame dalam video, serta menangkap informasi penting dari video yang mungkin terlewat oleh metode konvensional. Eksperimen yang dilakukan pada dataset SumMe dan TVSum menunjukkan bahwa TCVT berhasil mengungguli model-model sebelumnya dalam hal mean Average Precision (mAP) dan menghasilkan ringkasan video yang lebih representatif. TCVT menunjukkan kinerja terbaik pada tugas video summarization, baik dalam hal akurasi maupun ukuran model, dibandingkan dengan teknik-teknik lain yang ada. Dengan menggunakan pendekatan temporal dan kanal yang disesuaikan, TCVT mampu menghasilkan ringkasan video yang lebih akurat, yang memperhitungkan baik perubahan visual maupun gerakan dalam video (Tian dkk., 2025) .

ASAFormer, sebuah metode pelacakan visual yang menggabungkan Vision Transformer (ViT) dengan Asymmetric Selective Attention (ASA) untuk meningkatkan efisiensi komputasi dalam pelacakan objek. ASAFormer mengatasi masalah redundansi dalam perhatian dengan memilih token representatif, yang mengurangi kompleksitas komputasi tanpa mengurangi akurasi pelacakan. Selain itu, ASAFormer menggunakan Dynamic Masked Attention (DMA) untuk menyesuaikan interaksi antara template objek dan cabang pencarian secara adaptif. Hal ini memungkinkan model untuk lebih fokus pada bagian-bagian penting dari citra, meningkatkan ketepatan pelacakan, dan mengurangi gangguan dari latar belakang yang tidak relevan. Dalam penelitian ini, ASAFormer diuji pada beberapa dataset benchmark pelacakan objek, termasuk TrackingNet dan UAV123. Hasil eksperimen menunjukkan bahwa ASAFormer memberikan kinerja pelacakan yang superior dibandingkan dengan metode-metode pelacakan visual lainnya, dengan akurasi yang lebih tinggi dan efisiensi komputasi yang lebih baik. Model ini berhasil menangani tantangan-tantangan seperti occlusion, perubahan skala, dan variasi bentuk objek dengan lebih efektif. Metode utama yang digunakan dalam ASAFormer adalah Asymmetric Selective Attention (ASA), yang memanfaatkan teknik pemilihan token representatif untuk mengurangi redundansi dalam operasi perhatian. Dengan mengurangi jumlah token yang terlibat dalam perhatian, model ini berhasil mengurangi beban komputasi. Selain itu, Dynamic Masked Attention (DMA) digunakan untuk menyesuaikan interaksi antara template dan cabang pencarian, memungkinkan model untuk berfokus pada area yang paling relevan dalam citra atau video yang meningkatkan akurasi pelacakan (Gong dkk., 2024).

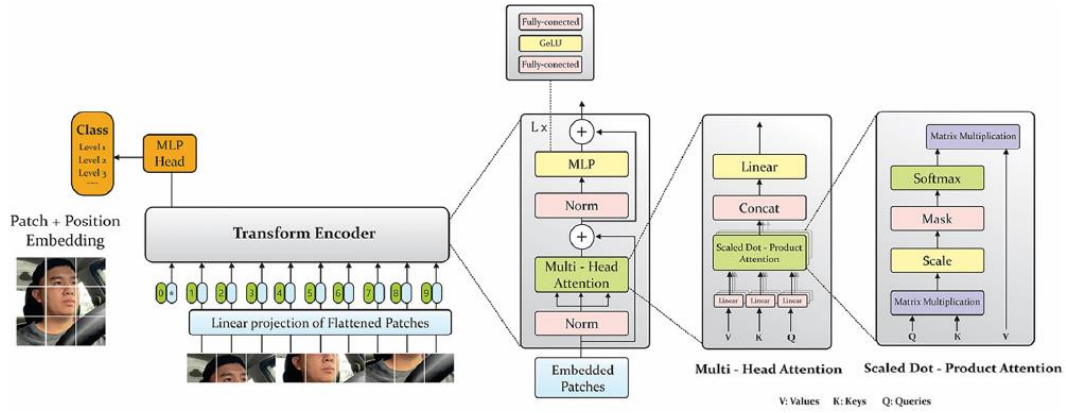
HuntFormer suatu metode pelacakan objek visual yang terinspirasi dari prinsip berburu alami, dengan menggunakan dua modul utama: Predictor dan Verifier. Predictor menggunakan particle filter untuk memprediksi lokasi objek dengan mengandalkan informasi pergerakan dan korelasi target, memungkinkan model untuk fokus pada objek yang terdeteksi dan mengurangi ruang pencarian. Verifier, di sisi lain, berfungsi untuk menilai keandalan hasil

pelacakan dengan menggunakan model ketidakpastian untuk memutuskan apakah template objek perlu diperbarui berdasarkan tingkat ketidakpastian yang terukur. Metode ini mengatasi masalah drift pelacakan dengan lebih efisien, memungkinkan HuntFormer untuk mendeteksi dan memperbaiki pelacakan yang hilang dengan cepat. Penelitian ini menunjukkan bahwa HuntFormer memiliki kinerja yang lebih unggul dibandingkan dengan model-model pelacakan lainnya pada dataset benchmark yang digunakan, termasuk pada kondisi pelacakan yang kompleks seperti occlusion dan objek yang keluar dari frame. Dalam eksperimen, HuntFormer berhasil mendeteksi kembali objek yang hilang lebih cepat dan lebih akurat daripada model pelacakan lainnya, yang menunjukkan keunggulannya dalam menghadapi perubahan besar dalam posisi objek. HuntFormer menggunakan metode berbasis particle filter untuk memperkirakan posisi objek dengan lebih baik, memanfaatkan motion trajectory sebagai pedoman untuk mendeteksi perubahan posisi objek. Verifier yang mengukur ketidakpastian dalam proses pelacakan membantu memperbarui template objek hanya jika diperlukan, meningkatkan efisiensi tanpa membebani komputasi. Metode ini menunjukkan bahwa menggabungkan prediksi berbasis filter dan pengecekan ketidakpastian dapat meningkatkan keandalan dan kecepatan dalam pelacakan objek (Z. Zhang dkk., 2024).

IAC-tracker merupakan pelacak objek visual berbasis Transformer yang berfokus pada penanganan perubahan penampilan objek secara langsung (*immediate appearance change*) menggunakan dua komponen utama: *Spatial Information Extractor* (SIE) dan *Temporal Information Extractor* (TIE). SIE bertanggung jawab untuk mengekstrak informasi spasial dari template objek dan wilayah pencarian, sementara TIE memproses informasi temporal dengan memperhatikan perubahan penampilan objek dalam urutan video. Gabungan dari kedua komponen ini memungkinkan model untuk lebih baik menangani perubahan skala dan deformasi objek, yang sering menjadi tantangan dalam pelacakan objek visual. Eksperimen yang dilakukan pada beberapa dataset benchmark, termasuk OTB100, menunjukkan bahwa IAC-tracker

mengungguli model pelacakan lainnya dalam hal *Area Under Curve* (AUC) dan mean Average Precision (mAP). Hasil ini menunjukkan bahwa IAC-tracker mampu menangani berbagai skenario pelacakan yang melibatkan objek dengan perubahan bentuk yang signifikan dan kondisi video yang kompleks seperti motion blur, occlusion, dan fast motion. Kelebihan lain dari IAC-tracker adalah kemampuannya untuk berjalan secara real-time, menjadikannya solusi yang efisien untuk aplikasi yang membutuhkan kecepatan dan akurasi. IAC-tracker menggunakan Transformer sebagai dasar arsitekturnya, yang memberikan kemampuan untuk menangkap ketergantungan jangka panjang antara objek dan latar belakang. Model ini menggabungkan SIE untuk ekstraksi fitur spasial dan TIE untuk menangkap dinamika temporal dalam video. Teknik spatial-temporal fusion digunakan untuk mengintegrasikan kedua informasi ini secara bersamaan, memungkinkan model untuk memahami dan melacak objek dengan lebih baik meskipun ada perubahan dalam penampilannya di berbagai frame video (Y. Li dkk., 2024).

Vision Transformer (ViT) diperkenalkan oleh sekelompok peneliti (Dosovitskiy dkk., 2020). Arsitektur ini merupakan jaringan saraf intensif yang dirancang untuk tugas pengenalan citra, dan dibangun berdasarkan arsitektur Transformer (Vaswani dkk., 2017). Gagasan utamanya adalah memperlakukan potongan-potongan citra (*image patches*) sebagai urutan token yang dapat digunakan sebagai masukan ke dalam model Transformer. Arsitektur Transformer telah berhasil diadaptasi dan diterapkan secara efektif dalam berbagai tugas, seperti restorasi citra dan deteksi objek (Y. Liu dkk., 2019). Dalam Vision Transformer, sebuah citra dibagi menjadi grid patch yang tidak saling tumpang tindih, di mana setiap patch kemudian diproyeksikan secara linear. Selanjutnya, urutan hasil embedding linear dari patch tersebut diberikan sebagai input ke encoder Transformer standar. Proses ini memungkinkan model untuk mempelajari informasi global maupun lokal dari suatu *image*.



Gambar 2.2 Vision Transformer Model (T.-C. Phan dkk., 2025a)

Gambar 2.2 menunjukkan gambaran umum dari model ViT, yang terdiri atas lapisan embedding, lapisan encoder, dan lapisan klasifikasi. Langkah pertama adalah membagi citra $\mathcal{X} \in \mathbb{R}^{CxHxW}$ dari dataset pelatihan menjadi patch $x_p \in \mathbb{R}^{Nx(CxPxp)}$, di mana $(H\ x\ W)$ adalah resolusi dari citra x, c adalah jumlah saluran channel, N adalah jumlah total patch, dan $(P\ x\ P)$ (Ayana & Choe, 2022) adalah resolusi dari tiap patch χp. Ukuran patch p umumnya ditetapkan sebesar 32 × 32 atau 16 × 16. Semakin kecil ukuran patch, maka performa model cenderung semakin baik, namun juga meningkatkan beban komputasi. Oleh karena itu, ukuran patch 16 × 16 dipilih karena dianggap mampu memberikan performa yang baik tanpa meningkatkan kompleksitas komputasi secara signifikan.

Transformer melakukan setiap patch sebagai token individu dan memetakannya ke dalam dimensi D menggunakan proyeksi linear E yang dapat dilatih. Proyeksi embedding ini dikaitkan dengan token kelas (*class token*) X_{class} yang dapat dilatih, yang berperan penting dalam menyelesaikan proses klasifikasi, serupa dengan yang dilakukan pada model BERT. Untuk melacak informasi lokasi, digunakan embedding posisi E_{pos} guna mempertahankan susunan spasial patch sesuai dengan posisi aslinya dalam citra. Hubungan terenkripsi dari patch ke token awal Z_0 direpresentasikan dalam suatu persamaan (Eq).

$$Z_0 = \left[x_{class}; X_p^1\ E; \ldots; X_p^N E\right] + Epos \qquad (2.1)$$

dimana :

Z_0 = Matriks input pertama (initial embedding)

Epos = Position Embedding

Vision Transformer terdiri atas *L* blok pengkode (*encoding blocks*) dan setiap blok dapat dibagi menjadi dua sub-komponen yang terdiri dari blok multi-head self-attention (MHA) dan blok multilayer perceptron (MLP). Blok MLP bertugas untuk mengalirkan nilai z_0 ke dalam pengkode transformer. Di dalam encoder MHA merupakan komponen utama dari transformer yang mencakup lapisan self-attention dan lapisan *concatenation* (penggabungan).

Self-attention nilai masukan $X = x_1, x^2, \ldots xn$ transformer melakukan operasi atensi secara simultan pada himpunan query set (Q) terhadap seluruh key (K) dan value (V) seperti ditunjukkan pada Persamaan (2.2). Matriks bobot W^Q,W^K,W^V adalah matriks bobot yang dapat dipelajari, yang didasarkan pada perhitungan hasil dot product antara query dan seluruh key, kemudian diskalakan dengan akar kuadrat dari dimensi D dan selanjutnya diterapkan fungsi softmax.

$$Atten(Q,K,V) = softmax\left(\frac{QK^T}{\sqrt{D}}\right)V \qquad (2.2)$$

Dimana :

Atten = *Attention*

Q (Query) = Matriks yang mewakili model pada setiap posisi dalam video

K (Key) = Matriks yang berisi indeks atau informasi kunci dari setiap posisi video untuk dicocokkan dengan *Query*

V (Value) = Matriks yang berisi informasi fitur yang sebenarnya

Softmax = Fungsi aktivasi yang mengubah skor kecocokan antara Q dan K menjadi bobot

Lapisan concatenation merupakan kerangka kerja transformer menjalankan mekanisme Multi-Head melalui scaled dot-product attention secara berulang kali dengan bobot pembelajaran yang berbeda-beda. Selanjutnya, semua kepala atensi (*attention heads*) digabungkan untuk

menghasilkan keluaran akhir, seperti ditunjukkan pada Persamaan (2.3), di mana W_i^Q, W_i^K, W_i^V, W^o adalah matriks parameter. Keluaran akhir pada kelas ι dari blok MHA ditentukan oleh Persamaan (2.4).

$$MHA(Q,K,V) = Concat(Atten1, Atten\ 2, \dots, Atten\ h)W^0 \quad (2.3)$$

dimana :

MHA = *Multi-Head Attention*

h = jumlah *head* atau kepala perhatian dalam model

W^0 = Matriks bobot proyeksi akhir

$$Z_l^{\backprime} = MHA\ (LN_{(z\iota-1)} + Z_\iota - 1, dimana\ l\ = 1, \dots, L \quad (2.4)$$

dimana :

Z_l^1 = Hasil keluaran (*output*)

$LN_{(z\iota-1)}$ = Layer Normalization

$Z_\iota - 1$ = Input dari lapisan sebelumnya

$L =$ Indeks Lapisan

MLP terdiri atas dua lapisan fully connected dan satu aktivasi GeLU di tengahnya, dengan keluaran pada kelas *l* didefinisikan dalam Persamaan (2.5). Elemen pertama dari Z_l^0 diambil dan dikirimkan ke *head classifier* untuk memprediksi kelas akhir dari encoder terhadap label sebagaimana didefinisikan dalam Persamaan (2.6).

$$z_1 = MLP\ (LN_{(z\iota-1)} + Z_l^{\backprime} - 1\ , dimanal = 1, \dots, L \quad (2.5)$$

dimana :

MLP = Multilayer Perceptron

z_1 = Representasi fitur akhir dari lapisan ke-l

$$y = LN(Z_l^0) \quad (2.6)$$

dimana :

y = Vektor fitur final

LN = Layer Normalization

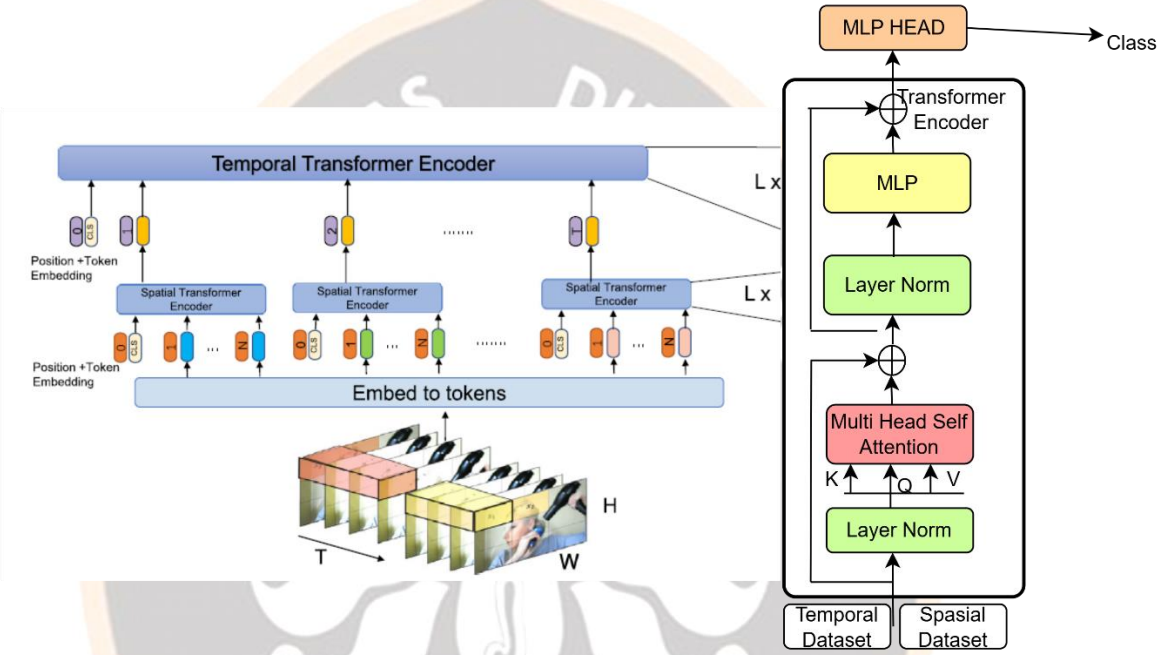
Z_l^0 = token pertama indeks ke-0

2.3.9 Video Vision Transformer (ViViT)

Sebagai salah satu terobosan dalam bidang pengenalan citra, Vision Transformer (ViT) telah memperkenalkan pendekatan baru yang mengadaptasi mekanisme perhatian (*attention mechanism*) dari arsitektur Transformer yang awalnya dikembangkan untuk pemrosesan bahasa alami ke dalam domain pengolahan citra. Tidak seperti Convolutional Neural network (CNN) yang mengandalkan operasi konvolusi lokal, Vision Transformer (ViT) membagi citra menjadi patch-patch kecil yang diperlakukan sebagai token, mirip seperti kata dalam kalimat, dan kemudian memprosesnya secara global menggunakan self-attention. Pendekatan ini memungkinkan model untuk menangkap hubungan spasial yang lebih luas dan kompleks antarpixel tanpa tergantung pada struktur lokal. Keunggulan ViT dalam memahami representasi visual inilah yang menjadi dasar bagi berbagai pengembangan arsitektur berbasis transformer, termasuk Video Vision Transformer (ViViT), yang diperluas untuk menangani data video yang mengandung informasi spasial dan temporal secara simultan.

Video Vision Transformer (ViViT) adalah sebuah arsitektur deep learning berbasis Transformer yang dirancang untuk memahami informasi spasial dan temporal dalam data video secara simultan. Tidak seperti Vision Transformer (ViT) yang hanya menangani data citra statis, ViViT memperluas prinsip Transformer ke domain video dengan membagi video menjadi potongan tiga dimensi yang disebut *tubelets*, yaitu patch citra yang mencakup dimensi ruang dan waktu. Setiap tubelet diubah menjadi representasi vektor dan diproses melalui rangkaian blok Transformer, yang mengandalkan mekanisme multi-head self-attention untuk menangkap hubungan antar patch secara fleksibel, tanpa ketergantungan pada operasi konvolusi atau jaringan berulang. ViViT mendukung berbagai pendekatan arsitektur, seperti *joint spatio-temporal attention* maupun *factorized attention* (pemrosesan spasial dan temporal dilakukan secara terpisah), yang masing-masing memiliki keunggulan dalam hal akurasi dan efisiensi komputasi. Berkat desainnya yang

modular dan skalabel, ViViT terbukti unggul dalam berbagai tugas pemahaman video, termasuk klasifikasi aksi, deteksi aktivitas, dan aplikasi real-time lainnya seperti deteksi kantuk berbasis video (Arnab dkk., 2021). Berikut merupakan gambaran ViViT .



Gambar 2.3 Arsitektur Video Vision Transfomer (Sun dkk., 2024)

Berbeda dengan ViT pada ViViT terjadi dari penyisipan klip video. Perluasan penyisipan ViT ke dalam video adalah ViViT, yang mengekstrak "tubelet" spasio-temporal yang tidak tumpang tindih dari video input dan melakukan proyeksi linear. Dibandingkan dengan frame n_t sampel seragam dari video, dimensi penyisipan tubelet

$$t\ x\ h\ x\ w,\ n_t = \left[\frac{T}{t}\right],\ n_h = \left[\frac{H}{h}\right],\ n_w = \left[\frac{W}{w}\right] \qquad (2.7)$$

Token-token yang diperoleh dari dimensi temporal, tinggi, dan lebar diproses secara terpisah. Kemudian, total token sebanyak $nt \cdot nh \cdot nw$ akan diteruskan ke dalam encoder Transformer. Dengan kata lain, penyematan tubelet dapat dianggap sebagai membangun blok 3D, informasi spatio-temporal digabungkan selama tokenisasi, seperti yang ditunjukkan di sisi kiri Gambar 2.3. *Multi-Headed Self-Attention* (MSA) dimasukkan dalam penyematan yang telah diubah pada model ViT dan ViViT setelah blok layer

normalization (LN). Blok multilayer perceptron (MLP) juga diperlukan dalam encoder Transformer setelah blok LN dan terdiri dari dua proyeksi linear yang dipisahkan oleh non-linearitas GELU.

Metode Video Vision Transformer (ViViT) yang digunakan dalam penelitian ini merupakan arsitektur yang berdiri di atas fondasi bertingkat, mulai dari konsep Deep Learning, arsitektur Transformer, hingga Vision Transformer (ViT). Sebagai bagian dari rumpun Deep Learning, ViViT memanfaatkan jaringan saraf tiruan yang dalam untuk mengekstraksi fitur kompleks dari data visual tanpa intervensi manual. Arsitektur ini mengadopsi mekanisme *self-attention* dari model Transformer, yang memungkinkan sistem memberikan bobot perhatian berbeda pada setiap bagian informasi. ViViT merupakan pengembangan langsung dari Vision Transformer (ViT), jika ViT mengadaptasi Transformer untuk pemrosesan citra spasial, maka ViViT memperluas kapabilitas tersebut ke dalam dimensi temporal. Dengan demikian, ViViT mengintegrasikan kekuatan ketiga unsur tersebut untuk menangkap hubungan spasio-temporal dalam video, sehingga perubahan kecil pada kondisi mata dan gerakan kepala pengemudi dapat dideteksi secara lebih akurat sebagai representasi dari kondisi kantuk.

2.3.10 Evaluasi Performa Model Deteksi Kantuk

Penelitian ini bertujuan untuk mengevaluasi performa model deteksi kantuk pada pengemudi dengan menggunakan pendekatan Brain-Computer Interface (BCI) yang berbasis sinyal EEG (*Electroencephalography*). Para penulis mengusulkan sistem klasifikasi kantuk yang tidak bergantung pada pengolahan citra wajah, melainkan menggunakan fitur neurofisiologis yang ditangkap langsung dari aktivitas otak. Pendekatan ini sangat relevan dalam konteks pengembangan sistem deteksi kantuk yang lebih objektif dan tidak terpengaruh oleh pencahayaan atau ekspresi wajah. Dataset yang digunakan berasal dari eksperimen mengemudi simulatif di mana peserta menjalani sesi panjang mengemudi monoton sambil dipantau oleh perangkat EEG. Setiap sesi diberi label berdasarkan kondisi sadar (alert) dan mengantuk (drowsy), yang dikonfirmasi secara subjektif maupun objektif melalui pengamatan gelombang

otak (terutama alpha dan theta) dan validasi perilaku. Hasil penelitian menunjukkan bahwa fitur-fitur EEG tertentu, seperti band daya pada frekuensi alpha dan theta, sangat relevan untuk deteksi kondisi kantuk. Model klasifikasi terbaik yang dihasilkan dalam studi ini adalah Random Forest, yang mampu mencapai F1-score hingga 79%, sementara SVM juga menunjukkan performa yang kompetitif. Selain itu, kombinasi kanal EEG tertentu (seperti O1 dan Fp2) menghasilkan akurasi yang lebih tinggi dibandingkan penggunaan semua kanal secara serempak, yang menunjukkan pentingnya pemilihan fitur kanal yang selektif(Hidalgo Rogel dkk., 2024) .

Beberapa penelitian menggabungkan pendekatan berbasis pengenalan citra wajah dengan algoritma machine learning untuk mendeteksi tanda-tanda kantuk seperti seringnya mata tertutup, frekuensi kedipan yang tinggi, serta ekspresi wajah yang menunjukkan kelelahan. Penelitian ini mengimplementasikan beberapa teknik klasifikasi seperti K-Nearest Neighbors (KNN), Support Vector Machine (SVM), Random Forest (RF), dan algoritma deep learning seperti Convolutional Neural Network (CNN) dan YOLOv5. Sebagai perbandingan, model juga diuji menggunakan Faster R-CNN dan YOLOv8 untuk mendeteksi wajah dan bagian mata secara presisi. Data yang digunakan dalam studi ini berasal dari tiga dataset publik yang sangat dikenal dalam penelitian deteksi kantuk, yaitu NTHU-DDD (National Tsing Hua University Drowsy Driver Detection), YawDD (Yawning Detection Dataset), dan UTA-RLDD (University of Texas at Arlington Real-Life Drowsiness Dataset). Ketiga dataset ini mencakup berbagai ekspresi wajah, kondisi pencahayaan, dan variasi gerakan kepala yang meniru kondisi nyata dalam berkendara, sehingga memungkinkan pengujian sistem secara lebih komprehensif. Hasil penelitian menunjukkan bahwa pendekatan berbasis deep learning, khususnya kombinasi CNN dan YOLOv5, mampu memberikan performa terbaik dalam mendeteksi kantuk dengan akurasi mencapai lebih dari 98% dalam pengujian pada dataset NTHU-DDD. YOLOv5 terbukti unggul dalam mendeteksi fitur wajah secara cepat dan akurat dalam skenario real-time. Selain itu, integrasi sistem klasifikasi berdasarkan pola mata dan kepala

menghasilkan peningkatan signifikan pada sensitivitas sistem dalam mengenali fase awal kantuk (Essahraui dkk., 2025).

Penelitian ini membahas pengembangan serta evaluasi performa dan penerimaan pengguna terhadap sistem deteksi kantuk berbasis wearable device, yang dirancang untuk memberikan solusi praktis dan nyaman bagi pengemudi dalam mendeteksi kelelahan selama berkendara. Sistem ini menggunakan perangkat pintar yang dikenakan di tubuh untuk memantau indikator fisiologis yang berkaitan dengan kondisi kantuk, seperti denyut jantung dan tingkat aktivitas tubuh. Fokus penelitian tidak hanya pada akurasi teknis sistem, tetapi juga pada bagaimana sistem ini diterima dan dirasakan oleh penggunanya dalam situasi nyata. Metode penelitian yang diterapkan mencakup dua tahap utama, yaitu pengujian teknis terhadap performa sistem dan survei evaluasi terhadap persepsi pengguna. Sinyal denyut jantung dan akselerasi gerakan digunakan sebagai parameter utama dalam mendeteksi peralihan dari kondisi waspada ke kondisi mengantuk. Data yang digunakan dalam penelitian ini dikumpulkan dari sejumlah peserta yang diminta menjalani sesi mengemudi panjang di lingkungan yang terkendali, baik dalam skenario mengemudi nyata maupun menggunakan simulator. Hasil penelitian menunjukkan bahwa sistem deteksi kantuk berbasis wearable ini mampu mencapai tingkat akurasi yang cukup tinggi dalam mengidentifikasi kondisi kantuk berdasarkan kombinasi sinyal denyut jantung dan gerakan tubuh. Dalam uji coba menggunakan 30 peserta dengan rentang usia 20–25 tahun dan 65–70 tahun, sistem deteksi kondisi kantuk berbasis deteksi denyut jantung (heart rate) berhasil mencapai akurasi keseluruhan 82,72% (Kundinger dkk., 2021).

Arsitektur Transformer yang diusulkan dalam FAViT (Feature-level Augmentation Vision Transformer) dirancang untuk mengatasi tantangan dalam tugas *video-based person re-identification* (ReID), khususnya dalam konteks efisiensi komputasi dan ketahanan terhadap gangguan visual. Penelitian ini bertujuan membangun model Vision Transformer yang mampu secara efektif mengekstraksi dan menggabungkan informasi spasial serta

temporal dari video, sembari menjaga kompleksitas model tetap rendah dibandingkan arsitektur Transformer konvensional yang cenderung sangat besar dan mahal secara komputasi. Dibandingkan dengan Vision Transformer konvensional, FAViT memiliki sejumlah keunggulan penting. Pertama, input token pada FAViT mencakup token latar belakang tambahan, yang tidak tersedia dalam ViT klasik. Kedua, mekanisme perhatian dalam FAViT dibedakan berdasarkan jenis token, sehingga model dapat memfokuskan perhatian hanya pada bagian-bagian penting dari video. Ketiga, augmentasi dilakukan pada tingkat fitur (bukan gambar), yang memungkinkan pelatihan model menjadi lebih fleksibel dan efisien. Selain itu, FAViT memperkenalkan dua sub-tugas pelatihan tambahan yang secara signifikan meningkatkan kemampuan generalisasi model. Yang terpenting, FAViT mampu mencapai performa tinggi dengan jumlah parameter yang jauh lebih rendah (~78 juta), sekitar 65% lebih ringan dibandingkan model-model Transformer sebelumnya(Kim dkk., 2025) .

Model Transformer berbasis perhatian berbobot tetangga terdekat (k-NN) yang disebut k-ViViT (k-NN Video Vision Transformer) untuk meningkatkan akurasi dan efisiensi dalam tugas pengenalan aksi pada video. Berangkat dari kelemahan model Vision Transformer (ViT) dan turunannya, seperti ViViT, yang menggunakan mekanisme self-attention sepenuhnya terhubung (fully-connected), model ini mencoba mengurangi kompleksitas komputasi dan dampak dari token yang tidak relevan atau mengandung noise. Untuk itu, penulis mengintegrasikan mekanisme k-NN attention, yang hanya memilih top-k token paling relevan, dalam arsitektur Transformer. Pendekatan ini memungkinkan proses pelatihan yang lebih cepat serta meningkatkan akurasi model dengan meminimalkan overfitting terhadap informasi tidak penting dalam frame video. Arsitektur yang dikembangkan terbagi menjadi dua varian: Uk-ViViT (Undecomposed) dan Dk-ViViT (Decomposed). Model Uk-ViViT mempertahankan arsitektur ViViT konvensional dengan integrasi k-NN attention secara langsung pada keseluruhan token spatio-temporal. Sementara itu, model Dk-ViViT memisahkan pemrosesan spasial dan temporal ke dalam

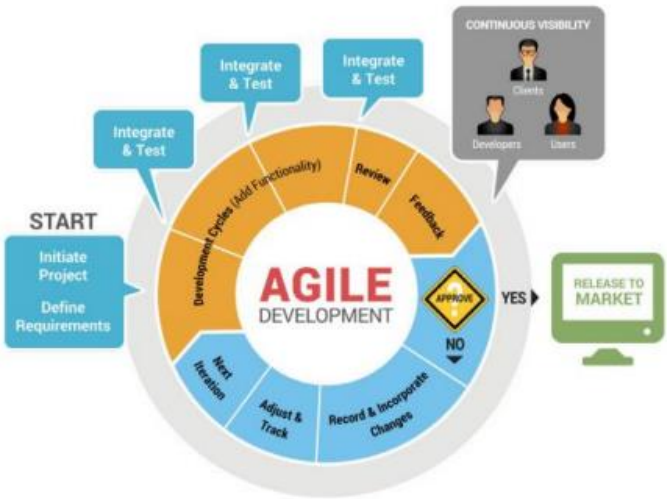
dua encoder Transformer yang berbeda, masing-masing dilengkapi dengan k-NN attention. Pemisahan ini memungkinkan interaksi antar-token yang lebih fokus dan mengurangi pencampuran informasi spasial dan temporal yang tidak berkaitan. Eksperimen dilakukan pada dua benchmark standar, yaitu UCF101 dan HMDB51, dengan menunjukkan peningkatan akurasi yang signifikan. Model Dk-ViViT mencatat Top-1 accuracy sebesar 94.2% pada UCF101 dan 82.5% pada HMDB51, mengungguli varian ViViT orisinal maupun model Uk-ViViT. Hasil ini menunjukkan peningkatan akurasi sebesar 7.5% dan 15.4% dibanding ViViT pada kedua dataset tersebut. Lebih lanjut, studi ablation menunjukkan bahwa penerapan k-NN attention secara terpisah pada domain spasial dan temporal masing-masing juga meningkatkan kinerja, meskipun implementasi kombinatif pada keduanya (Dk-ViViT) tetap menjadi yang paling unggul(Sun dkk., 2024).

Pendekatan Vision Transformer (ViT) yang diperkuat dengan modul Depth-Wise Convolutions (DWConv) sebagai jalur bypass untuk mengatasi kelemahan ViT dalam menangkap informasi lokal, khususnya pada dataset kecil. DWConv dirancang sebagai shortcut ringan yang memotong seluruh blok Transformer untuk mengembalikan representasi lokal tanpa menambah beban komputasi secara berarti, dan dapat diintegrasikan secara fleksibel tanpa mengubah struktur dasar model. Metode ini terdiri dari dua varian, yaitu bypass beberapa blok Transformer untuk efisiensi dan penggunaan paralel DWConv dengan berbagai kernel untuk menangkap detail lokal yang kaya. Implementasinya pada ViT-Tiny, ViT-Small, CaiT-xxs, dan Swin-Tiny menunjukkan peningkatan akurasi signifikan, seperti pada CIFAR-10 (dari 94.01% ke 96.41%) dan CIFAR-100 (dari 73.68% ke 78.05%), bahkan mengungguli model ViT yang lebih besar dengan parameter dan FLOPs lebih rendah. Selain itu, DWConv juga meningkatkan performa pada Tiny-ImageNet, ImageNet-1K, serta tugas deteksi dan segmentasi objek di COCO, dengan peningkatan skor mAP untuk Mask R-CNN dari 28.2 menjadi 30.6. Metode ini menunjukkan konvergensi pelatihan lebih cepat (~100 epoch dibanding 300 epoch), mampu menangkap objek kecil secara lebih akurat

(terlihat pada visualisasi Grad-CAM), dan menambah beban parameter serta komputasi <0.5%. Dengan demikian, DWConv menjadi solusi yang efisien dan efektif untuk meningkatkan performa ViT, terutama pada skenario pelatihan dari awal di dataset terbatas, tanpa memerlukan modifikasi besar pada arsitektur (T. Zhang dkk., 2025).

2.3.11 Agile Software Development Lifecycle

Pengembangan rekayasa perangkat lunak ditawarkan untuk mengatasi keterbatasan dan mengurangi biaya dengen memberikan fleksibilitas perubahan di setiap tahap (Fitzgerald dkk., 2019). Trend terkini mengadopsi pengembangan cloud computing, big data, dan koordinasi di setiap tahapan. Ada banyak metode dan pendekatan pengembangan perangkat lunak tradisional, seperti pendekatan waterfall, pendekatan iteratif dan inkremental, pendekatan spiral dan pendekatan evolusioner. Metode pengembangan perangkat lunak tradisional membutuhkan proses yang sangat berat. *Agile software development lifecycle* (ASDL) merupakan metode pengembangan perangkat lunak modern, metode pengembangan sistem ini adalah kerangka kerja konseptual untuk rekayasa perangkat lunak yang dimulai dengan fase perencanaan awal dan mengikuti jalan menuju fase pengembangan dengan interaksi berulang dan inkremental sepanjang siklus hidup sistem. Nilai-nilai dan prinsip-prinsip ini memberikan dasar untuk memandu proses pengembangan perangkat lunak dan untuk memberikan karakteristik yang dapat dibedakan untuk metode apa pun dengan fitur kecepatan(Al-Saqqa dkk., 2020). Tahapan agile dapat dilihat pada Gambar 2.4, gambar tersebut menjelaskan setiap tahapan development *life cycle agile*.



Gambar 2.4 Agile Software Development Lifecycle (ASDL) (Fitzgerald dkk., 2019)

Tahapan Agile Development Software :

1. Konsep (*Consept)*

Perencanaan adalah tahap kunci dalam metodologi Agile yang memungkinkan tim untuk memahami tujuan proyek, mengidentifikasi kebutuhan, dan merencanakan langkah-langkah untuk mencapai hasil yang diinginkan. Dalam tahap perencanaan peneliti mengidentifikasi fitur-fitur yang akan dikembangkan dan menentukan prioritasnya. Pada Tahapan ini juga melibatkan pembuatan rencana kerja yang mencakup estimasi waktu dan sumber daya yang diperlukan untuk mengembangkan setiap fitur.

2. Lahir (*Inseption*)

Tahap perancangan dalam metodologi Agile adalah momen ketika peneliti mengembangkan rancangan rinci untuk produk yang akan dikembangkan. Desain ini mencakup aspek visual, antarmuka pengguna, dan struktur keseluruhan produk. Dalam tahap perancangan, peneliti menggabungkan kebutuhan pengguna dengan visi produk untuk menciptakan desain yang baik dan fungsional. Desain ini dapat dibagi menjadi iterasi yang lebih kecil untuk memastikan bahwa desain terus berkembang seiring berjalannya waktu.

3. Pengulangan (*Iteration*)

Pengulangan dilakukan secara teratur dalam siklus Agile dan dapat melibatkan pengujian produk, pemeriksaan kode, atau evaluasi desain. Umpan balik yang diterima selama tahap ini membantu peneliti untuk melakukan perbaikan dan penyesuaian yang diperlukan sebelum melanjutkan ke tahap selanjutnya.

4. Peluncuran (*Release*)

   Tahap peluncuran adalah saat produk akhirnya siap untuk dirilis ke pengguna akhir. Setelah melalui berbagai tahap pengembangan, pengujian, dan perbaikan, produk dianggap telah mencapai kualitas yang memadai untuk digunakan oleh pengguna. Peluncuran produk bisa bersifat perlahan atau sekaligus, tergantung pada preferensi tim pengembang dan jenis produk yang dibangun. Dalam tahap ini, produk telah melewati semua proses pengembangan dan siap untuk memberikan nilai kepada pengguna.

5. Pemeliharaan (*Maintance*)

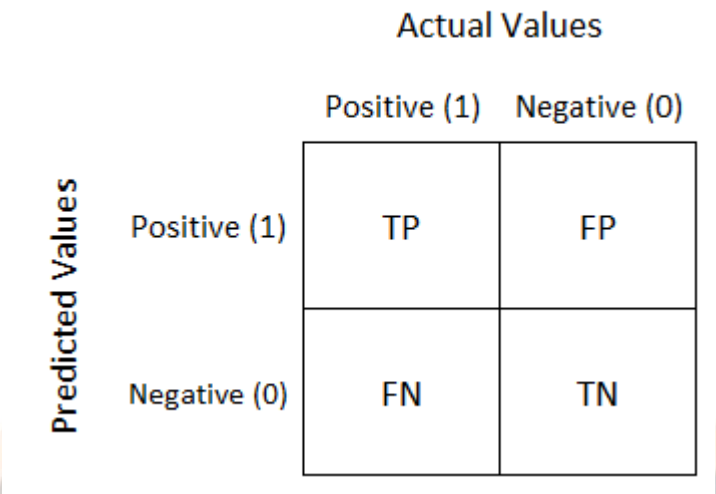
   Setelah produk dirilis dan digunakan oleh pengguna, tahap pemeliharaan dimulai. Tahap ini melibatkan pemantauan produk secara terus-menerus untuk memastikan kinerja yang baik dan mengatasi masalah yang mungkin muncul.

6. Pensiun (*Retired*)

   Tahap pensiun adalah akhir dari siklus hidup produk dalam metodologi Agile. Penting untuk mengelola pensiun produk dengan baik dan memberikan informasi kepada pengguna tentang perubahan yang akan terjadi. Dalam beberapa kasus, produk yang pensiun dapat digantikan dengan produk yang lebih baru dan lebih baik sesuai dengan kebutuhan pengguna dan pasar.

2.3.12 Metode Hasil Evaluasi

Matriks evaluasi diperlukan untuk memastikan bahwa model yang dikembangkan benar-benar mampu menyelesaikan permasalahan dalam mendeteksi tingkat kantuk secara akurat. Gambar 2.4 mengilustrasikan evaluasi model yang disebut dengan *confusion matrix*.



Gambar 2.5 Ilustrasi Confusion Matrix

Keterangan komponen pada *confusion matrix*:

- TP (True Positif): Kondisi kantuk yang berhasil terdeteksi dengan benar sesuai kenyataan.
- FP (False Positif): Model memprediksi kondisi kantuk padahal sebenarnya tidak terjadi.
- FN (False Negatif): Model gagal mendeteksi kondisi kantuk yang sebenarnya ada.
- TN (True Negatif): Model dengan benar memprediksi tidak ada kondisi kantuk sesuai kenyataan.

Dalam penelitian ini, analisis deteksi kantuk dilakukan berdasarkan empat metrik evaluasi utama untuk mengukur kinerja model, yaitu:

Analisis kumpulan data deteksi kantuk didasarkan pada empat metrik evaluasi utama yang digunakan untuk mengukur kinerja model secara menyeluruh. Metrik yang digunakan untuk analisis model deteksi ini meliputi Accuracy, Recall, Precision, dan F1-Score. Sedangkan metrik yang digunakan, yaitu : Inference Time (s)

Analisis dalam penelitian ini dilakukan berdasarkan eksperimen yang melibatkan keempat metrik tersebut untuk menguji kekuatan serta kelemahan metode yang digunakan. Setiap metrik dipilih untuk memberikan gambaran menyeluruh mengenai tingkat akurasi, ketepatan, sensitivitas, serta keseimbangan kinerja model dalam mendeteksi kondisi kantuk pada data uji. Metode metrik evaluasi deteksi kantuk dibuat sebagai berikut:

1. *Accuracy* adalah metrik evaluasi yang mengukur seberapa baik model membuat prediksi yang benar dari total prediksi yang dilakukan. Metrik *Accuracy* dapat dilihat pada persamaan 2.7

(2.8)

$$Accuracy = \frac{TP + TN}{TS}$$

dengan *True Positive* (TP) adalah jumlah prediksi positif benar, *True Negative* (TN) adalah jumlah prediksi negatif benar, *Total Sample* (TS) adalah jumlah dari sampel yang digunakan.

2. Precision adalah perbandingan antara True Positive (TP) dengan jumlah keseluruhan data yang diprediksi positif. Metrik *precision* dapat dilihat pada persamaan 2.8

(2.9)

$$Precision = \frac{TP}{TP + FP}$$

dengan *True Positive* (TP) adalah jumlah prediksi positif benar, *False Positive* (FP) adalah jumlah prediksi positif salah.

3. *Recall* adalah perbandingan antara True Positive (TP) dengan banyaknya data yang sebenarnya positif. Metrik *recall* dapat dilihat pada persamaan 2.9

(2.10)

$$Recall = \frac{TP}{TP + FN}$$

dengan *True Positive* (TP) adalah jumlah prediksi positif benar, *False Negative* (FN) adalah jumlah prediksi negatif salah.

4. *F1-Score* merupakan rata-rata harmonis dari precision dan recall yang memberikan gambaran seimbang antara kedua metrik tersebut. Metrik *F1-Score* dapat dilihat pada persamaan 2.10

(2.11)

$$F1 = 2 * \frac{1}{\frac{1}{precision} + \frac{1}{recall}}$$

dengan *precision* adalah perbandingan antara *True Positive* (TP) dengan banyaknya data yang diprediksi positif dan *recall* adalah perbandingan antara *True Positive* (TP) dengan banyaknya data yang sebenarnya positif.