

BAB I

PENDAHULUAN

1.1 Latar Belakang

Katarak adalah suatu kondisi dimana lensa mata mengalami kekeruhan sehingga menyebabkan menurunnya kemampuan penglihatan sampai kebutaan. Sebagian besar katarak disebabkan oleh proses penuaan. Faktor risiko yang dapat meningkatkan katarak adalah angka harapan hidup yang meningkat, riwayat trauma pada mata, infeksi saat kehamilan, konsumsi atau penggunaan tetes mata steroid dalam waktu lama, penyakit seperti kencing manis dan hipertensi, merokok, alkohol, paparan sinar matahari berlebih, paparan racun, riwayat keluarga, riwayat operasi pada mata, dan penyakit mata lainnya seperti glaukoma atau uveitis. Gejala katarak adalah pandangan kabur seperti berkabut, melihat lingkaran di sekeliling cahaya, penglihatan ganda, penurunan penglihatan di malam hari, silau saat melihat cahaya, dan sering mengganti ukuran kacamata. Sekitar 1 milyar penduduk dunia mengalami gangguan penglihatan berat atau kebutaan dan salah satu yang menyebabkan gangguan penglihatan atau kebutaan adalah katarak yaitu sebanyak 94 juta penduduk dunia (World Health Organization, 2023). Di Indonesia sendiri, terdapat sekitar 8 juta penduduk yang mengalami gangguan penglihatan sedang hingga berat dan 1,6 juta penduduk mengalami kebutaan. Sekitar 1,7% hingga 4,4% penduduk diatas usia 50 tahun mengalami kebutaan. Penyebab utama kebutaan di Indonesia adalah katarak sebesar 81,2% (Direktorat Jenderal Kesehatan Lanjutan, 2022). Secara umum, katarak dapat diklasifikasikan menjadi 2 stadium yaitu stadium awal (*incipient* dan *immature*) dan stadium lanjut (*mature* dan *hypermature*) (Kenali 4 Stadium Katarak Dan Cara Pencegahannya, 2024).

Telemedicine adalah salah satu sistem yang berkembang pesat dalam dunia medis, menawarkan solusi untuk berbagai masalah akses kesehatan, terutama di daerah terpencil atau dengan keterbatasan sumber daya. *Telemedicine* memberikan manfaat besar, seperti peningkatan efisiensi dalam memberikan perawatan, penurunan biaya kesehatan, serta peningkatan akses pasien terhadap spesialis. *Telemedicine* telah mengalami perkembangan signifikan sejak awal penggunaannya, didorong oleh kebutuhan untuk meningkatkan akses layanan kesehatan di daerah yang sulit dijangkau, mengurangi biaya, serta mempercepat diagnosis dan pengobatan melalui interaksi jarak jauh antara dokter dan pasien. Terkait penggunaan kamera *smartphone* dalam *telemedicine*, khususnya untuk keperluan klinis seperti oftalmologi (pengobatan mata) memiliki tantangan terkait keterbatasan kualitas gambar, pencahayaan dan fokus (Stoltzfus et al., 2023).

Vision Transformer (ViT) dan variannya telah diperkenalkan untuk berbagai tugas visi komputer misalnya, klasifikasi gambar, deteksi objek, segmentasi semantik serta pemrosesan gambar medis dan mendapatkan perhatian yang meningkat dari masyarakat. ViT dirancang untuk bekerja dengan data berkualitas tinggi, terutama karena model ini mengandalkan self-attention untuk mengekstrak detail dan konteks

dari gambar. Namun, gambar mata umumnya memantulkan spekular. Spekular pada gambar mata dapat menyebabkan diagnosis katarak menjadi tidak akurat karena pantulan cahaya yang berlebihan pada permukaan kornea atau lensa mata dapat mengaburkan atau menciptakan artefak pada area yang seharusnya diamati dengan jelas, sehingga dapat berpengaruh pada performa model. Mata bisa memantulkan spekular karena beberapa alasan yaitu permukaan yang halus dan mengkilap, lapisan air mata, indeks bias yang berbeda dan bentuk cembung (Faizal et al., 2023).

Standar pencahayaan ruang pemeriksaan pasien berkisar antara 100-200 lux (Kepmenkes RI No 1204, 2004), serta permukaan mata, terutama kornea, memiliki karakteristik reflektif tinggi dan melengkung, sehingga setiap sumber cahaya langsung yang digunakan untuk pencitraan mata cenderung dipantulkan ke arah kamera atau alat visualisasi, terlebih jika jarak pengambilan gambar yang begitu dekat dan posisi pengambilan gambar sejajar antara mata dan kamera yang tidak memungkinkan untuk melakukan teknik-teknik untuk mengatasi pantulan spekular sehingga menghindari pantulan spekular pada mata sangat sulit dilakukan.

Berbagai penelitian telah mengeksplorasi penggunaan ViT untuk melakukan diagnosis penyakit melalui citra medis baik itu yang diambil menggunakan alat elektromedik maupun kamera digital. Kumar et al., 2023 telah menunjukkan kemampuan ViT dalam mengidentifikasi penyakit katarak menggunakan data citra mata dengan tingkat akurasi sebesar 70%. Kemudian Li et al., 2024 menyajikan sistem klasifikasi citra medis dengan melakukan peningkatan ViT melalui token-wise perturbation dengan mengusulkan BOLT sebagai metode self-supervised learning dan menunjukkan tingkat akurasi sebesar 81.5%

(Hisham et al., 2023) juga menyajikan system klasifikasi penyakit mata pada dataset berupa citra fundus dengan menggunakan algoritma Swin Transformer dan Data Efficient Image Transformer (DeiT) yang merupakan varian dari ViT. Sistem ini menunjukkan tingkat akurasi tertinggi pada varian Swin Transformer sebesar 93.1%.

Penelitian yang disebutkan menunjukkan pentingnya terus mengembangkan dan menyesuaikan teknik pembelajaran ViT berdasarkan jenis data, lingkungan dan instruksi yang diberikan untuk mencapai tujuan yang optimal. Penelitian-penelitian ini secara kolektif menunjukkan potensi ViT dan modifikasinya dalam mencapai akurasi dan kinerja yang tinggi dalam tugas-tugas visi komputer. Karya-karya ini dapat dianggap sebagai dasar penting dalam literatur pembelajaran mendalam yang menunjukkan jalan menuju inovasi teknologi berkelanjutan di bidang medis.

Meskipun penelitian sebelumnya menunjukkan hasil yang menjanjikan, belum ada yang secara khusus meneliti bagaimana pantulan spekular pada gambar mata berbasis *smartphone* mempengaruhi kinerja Vision Transformer (ViT). Penelitian ini secara langsung menutup kesenjangan tersebut dengan berfokus pada kontribusi berikut:

1. Mengusulkan penerapan pra-pemrosesan berbasis image inpainting untuk mengurangi artefak pantulan spekular dan ViT yang dioptimalisasi dengan Local Window Attention (LWA) dan Region of Interest (ROI) Attention.
2. Melakukan perbandingan (*benchmarking*) terhadap sistem yang diusulkan dibandingkan ViT dasar.

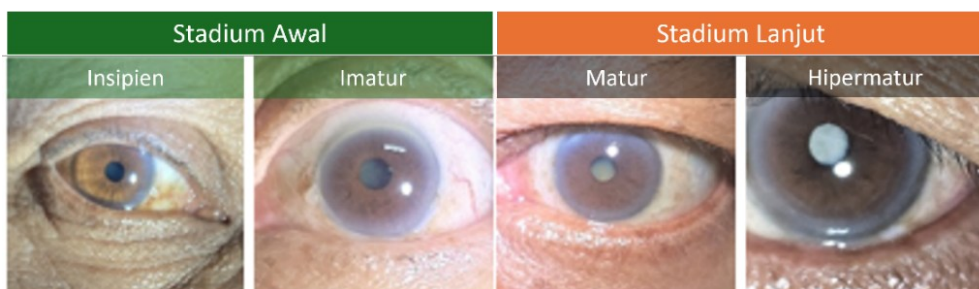
1.2 Landasan Teori

1.2.1 Katarak

Katarak adalah proses degeneratif berupa kekeruhan di lensa bola mata sehingga menyebabkan menurunnya kemampuan penglihatan sampai kebutaan. Kekeruhan ini disebabkan oleh terjadinya reaksi biokimia yang menyebabkan koagulasi protein lensa. Katarak bisa terjadi secara kongenital (katarak sejak lahir) namun pada umumnya, katarak terjadi karena proses degenerasi yang berhubungan dengan penuaan atau bisa juga diakibatkan karena trauma dan induksi dari obat-obatan serta komplikasi dari kondisi sistemik seperti diabetes mellitus atau penyakit mata seperti glaukoma dengan *uveitis*. Secara klinis, katarak dapat diklasifikasikan menjadi 4 stadium yaitu:

1. Insipien
Merupakan stadium satu (awal), dimana kekeruhan lensa mata masih sangat minim, bahkan terkadang sulit terlihat tanpa menggunakan alat bantu periksa.
2. Imatur
Katarak stadium immatur ditandai dengan kekeruhan lensa mata yang lebih tebal tetapi belum mengenai seluruh lensa sehingga masih terdapat bagian-bagian yang jernih pada lensa.
3. Matur
Katarak stadium tiga (tahap lanjut), dimana kekeruhan telah mengenai seluruh massa lensa. Kekeruhan ini terjadi akibat deposit ion Ca yang menyeluruh.
4. Hiper matur
Merupakan stadium empat (tahap sangat lanjut), dimana kekeruhan telah mengenai seluruh massa lensa dan telah mengalami proses degenerasi lanjut, dapat menjadi lembek dan mencair (Kenali 4 Stadium Katarak Dan Cara Pencegahannya, 2024).

Berdasarkan jarak pandang, keempat stadium katarak dapat dikelompokkan ke dalam 2 kategori yaitu stadium awal (insipien dan imatur) serta stadium lanjut (matur dan hiper matur) (Nizami et al., 2024). Jenis-jenis stadium katarak secara detail dapat dilihat pada gambar 1 berikut.



Gambar 1. Jenis-jenis stadium katarak

1.2.2 Telemedicine

Telemedicine adalah penggunaan teknologi komunikasi, seperti video, panggilan telepon dan internet, untuk memberikan layanan Kesehatan jarak jauh antara pasien dan penyedia layanan kesehatan. *Telemedicine* memungkinkan dokter dan pasien berkomunikasi secara real-time meskipun berada di lokasi yang berbeda.

Telemedicine sering kali mencakup konsultasi medis, diagnose awal, pemantauan kondisi kronis, dan bahkan penanganan darurat dalam situasi tertentu, seperti tele-stroke. Selain itu, telemedicine memberikan Solusi penting dalam memperluas akses layanan kesehatan ke daerah terpencil atau bagi pasien dengan keterbatasan mobilitas. *Telemedicine* bukan hanya alat sementara, melainkan bagian yang semakin penting dari layanan kesehatan modern, terutama dalam memperluas akses perawatan dan meningkatkan efisiensi sistem kesehatan global (Stoltzfus et al., 2023).

1.2.3 Pantulan spekular

Pantulan spekular adalah jenis pantulan cahaya di mana cahaya yang mengenai suatu permukaan dipantulkan secara teratur dan arah pantulannya sama dengan sudut datangnya cahaya. Fenomena ini terjadi pada permukaan yang sangat halus dan mengkilap, seperti cermin atau logam yang dipoles, di mana permukaan tersebut tidak menyebarkan cahaya ke berbagai arah, tetapi memantulkannya dalam satu arah tertentu.

Pantulan spekular pada mata dapat terjadi ketika cahaya mengenai permukaan halus mata, seperti kornea atau lensa, dan dipantulkan langsung ke mata. Berikut beberapa faktor yang menyebabkannya:

1. Permukaan yang halus dan mengkilap
Kornea ini memiliki permukaan yang halus dan berkilau, sehingga berfungsi seperti cermin yang memantulkan cahaya secara langsung.
2. Lapisan air mata
Di atas kornea terdapat lapisan tipis air mata yang menjaga permukaan kornea tetap lembap dan licin. Lapisan air ini juga menambah kemampuan mata untuk memantulkan cahaya dengan baik, karena air meningkatkan kemampuan permukaan untuk memantulkan cahaya secara spekular.
3. Indeks bias yang berbeda
Transisi antara udara, lapisan air mata, dan kornea yang memiliki indeks bias berbeda-beda menyebabkan cahaya dipantulkan secara efisien. Kombinasi ini membuat mata mampu memantulkan pantulan spekular yang sangat jelas.
4. Bentuk cembung
Kornea memiliki bentuk cembung yang membantu memfokuskan dan memantulkan cahaya. Sifat optik ini mirip dengan lensa yang membuat pantulan spekular terlihat dengan jelas saat terkena cahaya dari sudut tertentu (Faizal et al., 2023).

1.2.4 Pengolahan citra

Pengolahan citra digital adalah teknologi yang menerapkan sejumlah algoritma komputer untuk memproses citra digital. Hasil keluaran dari proses tersebut dapat berupa gambar atau karakteristik yang merepresentasikan citra. Menurut (Russakovsky et al., 2014), pengolahan citra adalah istilah umum untuk berbagai teknik yang keberadaannya untuk memanipulasi dan memodifikasi citra dengan berbagai cara. Biasanya gambar dua dimensi yang bisa diolah dengan mudah adalah foto, setiap foto dalam bentuk citra digital bisa diolah dengan melalui perangkat lunak tertentu.

Pengolahan citra juga dilakukan agar informasi yang terkandung di dalam citra tersebut dapat tersampaikan dengan baik dan jelas kepada pengguna (user). Sebuah citra dapat mengandung banyak informasi, namun seringkali citra yang kita miliki mengalami penurunan kualitas ataupun penurunan mutu (degradasi), misalnya mengandung cacat (noise), warna yang terlalu kontras, kurang tajam, ataupun kabur (blur). Hal semacam ini menyebabkan citra menjadi sulit untuk diinterpretasikan karena informasi yang disampaikan oleh citra tersebut menjadi berkurang.

1.2.5 Teknik morfologi

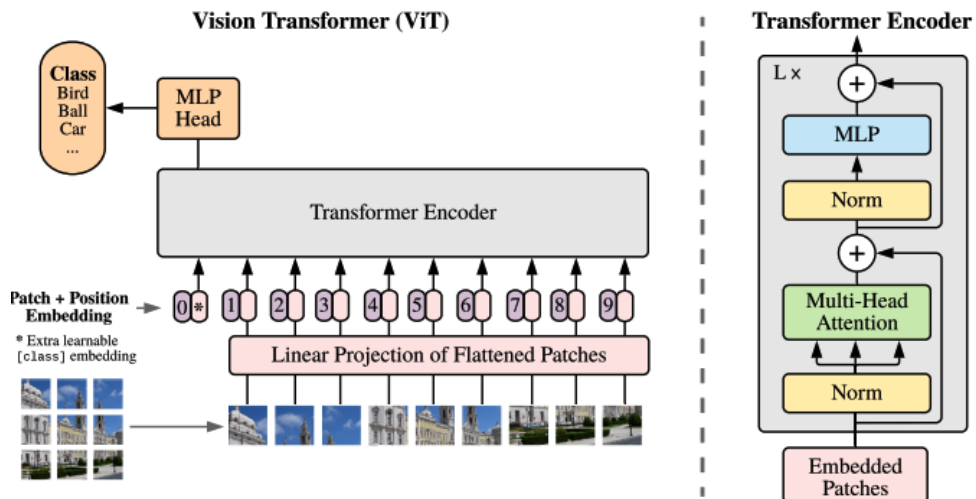
Teknik morfologi dalam pengolahan citra berawal dari teori himpunan dan dirancang untuk mengoperasikan struktur geometris citra melalui penggunaan elemen struktural. Teknik dasar yang paling banyak digunakan adalah erosi dan dilasi. Erosi akan mengecilkan area objek dengan menghapus piksel tepi sedangkan dilasi akan memperluas area objek. Dari kombinasi kedua operasi ini dibentuk operasi *opening* (erosi lalu dilasi) untuk menghilangkan noise kecil atau protusi, dan *closing* (dilasi lalu erosi) untuk menutup celah kecil atau lubang pada objek (Beyerer Jürgenand Puente León, 2016).

Selanjutnya, morfologi juga menyediakan operator lebih kompleks dan aplikasi spesifik di bidang segmentasi, ekstraksi bentuk, dan pemrosesan medis. Misalnya, transformasi top-hat atau bottom-hat digunakan untuk memperkuat fitur tertentu atau latar belakang lokal, sedangkan teknik skeletonization dan operator “connected” (penghubung) bekerja pada representasi region (zona datar) untuk mempertahankan kontur tanpa menciptakan kontur baru (Lotufo et al., 2023). Dalam aplikasi medis, metode morfologi kadang digunakan sebagai tahap preprocessing untuk menghapus artefak, memperhalus tepi objek biologis, atau memperkuat struktur yang relevan (misalnya meningkatkan kontras pada citra jaringan) agar tahap segmentasi dan klasifikasi menjadi lebih *robust* (Hassanpour et al., 2015).

1.2.6 Vision transformer (ViT)

Salah satu inovasi terbaru di bidang deep learning adalah penerapan model transformer, yang awalnya dikembangkan untuk tugas pemrosesan bahasa alami, kini dikembangkan ke dalam domain pengolahan citra. Baru-baru ini, Vision Transformer telah menunjukkan kinerja yang sangat menjanjikan pada tugas-tugas visi komputer, seperti klasifikasi gambar, deteksi objek, dan segmentasi. (Hisham et al., 2023).

Secara khusus, Vision Transformer (ViT) memperlakukan seluruh gambar input sebagai urutan tambalan dan memanfaatkan mekanisme self-attention untuk mengeksplorasi informasi spasial global gambar untuk tugas-tugas yang berhubungan dengan visi komputer. Secara umum, Vision Transformer terdiri dari blok multihead self-attention (MSA) dan multi-layer perceptron (MLP). Adapun arsitektur Vision Transformer dapat dilihat pada gambar 2 berikut.



Gambar 2. Arsitektur Vision Transformer (ViT)

Cara kerja Vision Transformer secara garis besar dapat dibagi menjadi beberapa tahap yaitu (Dosovitskiy et al., 2020):

1. Pembagian gambar menjadi patch

Dalam ViT, gambar 2D (misalnya, gambar 224×224 piksel) dibagi menjadi beberapa patch kecil berukuran tetap (misalnya, patch 16×16 piksel). Setiap patch kemudian diubah menjadi vektor linier untuk mewakili informasi dalam patch tersebut. Jika kita mengambil gambar berukuran 224×224 dengan patch 16×16 , kita akan mendapatkan $14 \times 14 = 196$ patch.

2. Embedding patch

Setelah gambar dibagi menjadi patch, setiap patch di-embed menjadi vektor berdimensi tetap. Embedding ini juga mencakup penambahan embedding posisi (positional embedding) untuk menjaga urutan spasial informasi dari gambar.

3. Transformer encoder

Vektor embedding patch kemudian dimasukkan ke dalam Transformer encoder (seperti dalam Transformer NLP) yang terdiri dari lapisan multi-head self-attention dan lapisan feed-forward. Encoder ini menangani hubungan global antar patch, sehingga memungkinkan model untuk "memahami" citra dengan mempertimbangkan seluruh gambar.

4. Klasifikasi

Setelah proses melalui Transformer encoder, vektor hasil akhirnya diambil dari token klasifikasi yang kemudian digunakan untuk prediksi atau klasifikasi pada gambar tersebut.

1.2.7 *Image inpainting*

Pantulan spekular menyebabkan terbentuknya bitnik-bintik terang pada gambar mata. Hal ini terutama disebabkan oleh pantulan cahaya yang dipancarkan dari senter perangkat dari permukaan mata. Pantulan spekular mengganggu dan memanipulasi fitur gambar yang dapat berpotensi menyebabkan diagnosis yang salah. Hal ini disebabkan oleh ekstraksi fitur yang salah yang diperoleh sebagai hasil dari penerapan thresholding adaptif pada gambar. Thresholding adaptif sangat penting untuk memisahkan dan mengisolasi wilayah yang diminati dari latar belakang lainnya berdasarkan perbedaan intensitas piksel masing-masing wilayah.

Dalam hal ini, gambar RGB pertama-tama dikonversi ke bentuk abu-abu yang diikuti dengan mengubahnya ke dalam bentuk biner. Daerah dengan intensitas tertinggi terlihat jelas dalam gambar biner dan dapat dipisahkan setelah memeriksa nilai ambang batas. Pelebaran morfologi kemudian diterapkan pada gambar biner yang diperoleh. Hal ini mengarah pada identifikasi batas luar daerah pantulan spekular. Bagian yang dibatasi oleh batas ini kemudian diganti dengan piksel yang berdekatan untuk menyamakan sesuai dengan latar belakangnya (Faizal et al., 2023). Teknik ini disebut sebagai *image inpainting* seperti yang ditunjukkan pada gambar 3 berikut.



Gambar 3. Proses penghapusan pantulan spekular dengan *image inpainting*

1.2.8 *Confusion matrix*

Confusion matrix adalah sebuah matriks yang menunjukkan jumlah kejadian antar dua nilai label, yaitu label aktual dan label prediksi. Sebuah *confusion matrix* memiliki tabel berukuran $N \times N$, dimana N adalah jumlah nilai label atau kelas yang berbeda (Kohavi, n.d.). Pada multi-class *confusion matrix*, kelas-kelas diurutkan dalam urutan yang sama dalam baris dan kolomnya. Oleh karena itu, elemen yang diklasifikasi benar (True Positive) terletak pada diagonal utama dari kiri atas ke kanan bawah dan merupakan titik temu antara prediksi dan aktual (Grandini et al., 2020). Berikut merupakan contoh multi-class *confusion matrix* yang ditunjukkan pada gambar 4.

		Predicted		
		Awal	Lanjut	Normal
Actual	Awal	TP	FN	FN
	Lanjut	FP	TP	FN
	Normal	FP	FP	TP

Gambar 4. *Confusion matrix multiclass*

Metrik evaluasi digunakan adalah *precision*, *recall*, dan *F1 Score*.

1. *Precision* menggambarkan tingkat keakuratan model dalam memberikan prediksi yang sesuai dengan data yang diharapkan. *Precision* dihitung sebagai rasio antara jumlah prediksi positif (TP) yang benar dengan total data yang diprediksi sebagai positif. Dengan kata lain, *precision* menunjukkan seberapa banyak dari prediksi positif yang benar – benar merupakan data positif. Nilai *precision* dapat diperoleh dengan persamaan (1).

$$Precision = \frac{True\ Positif(TP)}{True\ Positif(TP) + False\ Positif(FP)} \quad (1)$$

2. *Recall* menunjukkan kemampuan model dalam mengidentifikasi Kembali informasi yang relevan. *Recall* dihitung sebagai rasio antara jumlah prediksi positif yang benar terhadap total data yang benar – benar positif. Dengan kata lain, *recall* mengukur sejauh mana model berhasil mendeteksi data positif secara akurat. Nilai *recall* dapat diperoleh dengan persamaan (2).

$$Recall = \frac{True\ Positif(TP)}{True\ Positif(TP) + False\ Negatif(FN)s} \quad (2)$$

3. *F1-score* merupakan metrik evaluasi yang mengukur keseimbangan antara *precision* dan *recall*. Nilai *F1-score* memberikan gambaran seberapa baik model dalam mengintegrasikan kemampuan *precision* dan *recall*, sehingga dapat menilai efektivitas model dalam melakukan kalsifikasi secara akurat. Nilai *F1-score* dapat diperoleh dengan persamaan (3).

$$F1\text{-score} = 2 \times \frac{(Precision \times Recall)}{(Precision + Recall)} \quad (3)$$

4. *Macro-F1* merupakan metrik evaluasi yang mengukur performa model klasifikasi dengan lebih dari dua kelas secara seimbang. Tidak seperti akurasi yang dapat bias pada kelas mayoritas, *macro-F1* memberikan gambaran menyeluruh dengan menghitung nilai *F1-Score* untuk setiap kelas terlebih dahulu, kemudian mengambil rata-ratanya. Nilai *macro-F1* dapat dihitung dengan persamaan (4).

$$Macro\text{-}F1 = \frac{1}{N} \sum_{i=1}^N F1_i \quad (4)$$

1.3 Rumusan Masalah

Berdasarkan latar belakang penelitian adapun rumusan masalah dalam penelitian ini adalah sebagai berikut:

1. Bagaimana meningkatkan akurasi klasifikasi stadium katarak dari gambar mata yang diambil menggunakan smartphone pada telemedicine?
2. Bagaimana performa Image Inpainting dan Vision Transformer yang telah dioptimalisasi dalam klasifikasi stadium penyakit katarak dari gambar mata yang diambil menggunakan smartphone pada telemedicine?

1.4 Tujuan dan Manfaat

1.4.1 Tujuan

Berdasarkan masalah yang telah dirumuskan, adapun tujuan dari penelitian ini adalah sebagai berikut:

1. Menerapkan metode Image Inpainting dan melakukan optimalisasi ViT dalam melakukan klasifikasi stadium katarak dari gambar mata yang diambil menggunakan *smartphone*.
2. Menganalisis performa image inpainting dan algoritma ViT yang telah dioptimalisasi dalam melakukan klasifikasi stadium katarak dari gambar mata yang diambil menggunakan *smartphone*.

1.4.2 Manfaat

Manfaat yang dapat diperoleh dari penelitian yaitu sebagai berikut.

1. Menambah referensi terkait pendekatan deep learning dengan menggunakan metode Vision Transformer terhadap klasifikasi stadium penyakit katarak.
2. Menambah wawasan dan pengetahuan baru terkait metode klasifikasi stadium penyakit katarak menggunakan ViT yang telah dioptimalkan.
3. Sebagai referensi dalam pengembangan penelitian topik terkait *telemedicine*.

1.5 Ruang Lingkup Penelitian

Ruang lingkup penelitian antara lain sebagai berikut:

1. Penelitian ini mencakup proses akuisisi data, pra-pemrosesan gambar (*preprocessing*), dan klasifikasi gambar mata untuk menentukan stadium katarak.
2. Pengambilan data menggunakan kamera ultra-wide Samsung S24 Ultra resolusi 12 megapiksel, pengaturan 3x optical zoom, dan posisi kamera sejajar mata pasien pada jarak 10-15 cm, dengan bantuan pencahayaan flashlight untuk memastikan pencahayaan yang cukup untuk melihat kondisi pupil.
3. Model dilatih menggunakan dataset hasil pra-pemrosesan untuk klasifikasi stadium katarak (awal, lanjut dan normal).

BAB II

METODE PENELITIAN

2.1 Waktu dan Lokasi Penelitian

Penelitian ini dimulai pada bulan Desember 2024 – Juli 2025. Lokasi penelitian dilaksanakan di:

1. Laboratorium Kecerdasan Buatan, Departemen Teknik Informatika Universitas Hasanuddin, Jalan Poros Malino, Kelurahan Borongloe, Kecamatan Bontomarannu, Kabupaten Gowa, Provinsi Sulawesi Selatan.
2. RS Mata Makassar, Jalan Jalur Lingkar Barat, Tallasa City, Tamalanrea, Makassar, Sulawesi Selatan.

Adapun alasan pemilihan lokasi mencakup:

1. Fasilitas perangkat keras dan lunak yang mendukung eksperimen komputasi.
2. Akses terhadap data primer pasien penderita katarak
3. Kolaborasi dengan dokter spesialis mata dan optisi refraksi.

2.2 Perangkat Penelitian

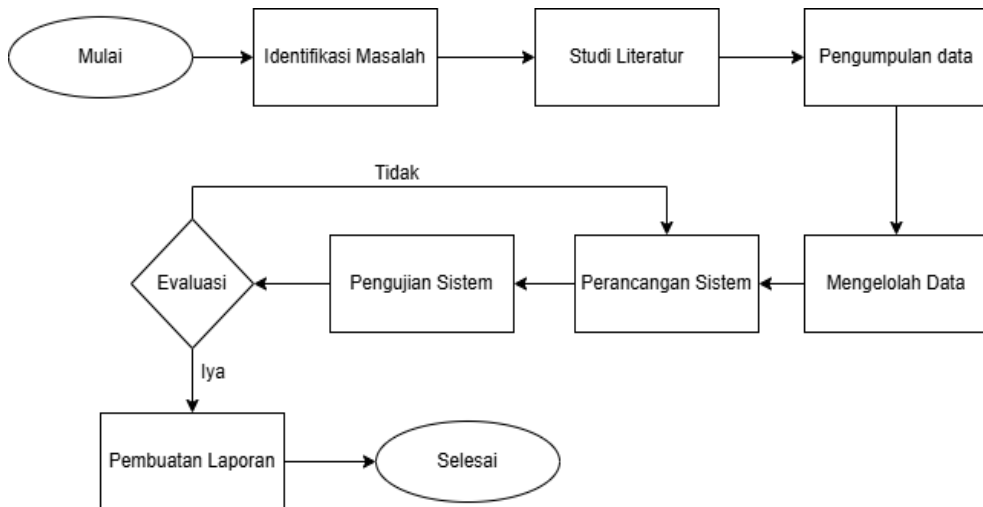
Adapun perangkat-perangkat yang digunakan dalam menunjang penelitian ini yaitu sebagai berikut:

1. Perangkat Keras (Hardware)
 - a. Laptop Intel® Core™ i7 processor, 16 GB RAM, 6 GB NVIDIA GeForce RTX 3070
 - b. Samsung S24 Ultra
 - c. *Flashlight* (Sender)s
2. Perangkat Lunak (Software)
 - a. Sistem Operasi Windows 11-64 bit
 - b. Bahasa Pemrograman Python
 - c. Visual Studio Code
 - d. Roboflow
 - e. Microsoft Office

2.3 Tahapan Penelitian

Penelitian ini merupakan penelitian dengan jenis eksperimental yang dilakukan melalui serangkaian tahapan terstruktur untuk menguji dan menganalisis performa sistem yang dikembangkan. Pengumpulan data dilakukan melalui studi literatur dari berbagai jurnal dan artikel ilmiah yang relevan, sehingga diperoleh pemahaman mendalam mengenai ruang lingkup permasalahan serta dasar teori yang mendukung penelitian ini. Dari hasil kajian tersebut, dirumuskan pendekatan metodologis dan rancangan sistem yang menjadi dasar eksperimen.

Tahapan penelitian meliputi proses pengumpulan dan persiapan data, penerapan metode pengolahan citra, pengembangan model Vision Transformer yang telah dioptimasi, serta evaluasi kinerja sistem. Secara lengkap, alur tahapan penelitian yang dilakukan dapat dilihat pada gambar 5, yang menggambarkan hubungan antara setiap proses mulai dari input data hingga tahap evaluasi hasil.



Gambar 5. Tahapan penelitian

2.3.1 Identifikasi masalah

Dalam penelitian ini, tantangan utama dalam meningkatkan akurasi klasifikasi stadium katarak terletak pada adanya pantulan cahaya (spekular reflection) pada permukaan mata. Pantulan ini muncul karena karakteristik mata yang halus, mengkilap, dan memiliki lapisan air mata, sehingga sering kali memunculkan area terang berlebihan pada citra. Kondisi tersebut dapat menimbulkan ambiguitas pada area pupil dan lensa, yang menjadi fitur utama dalam penentuan stadium katarak. Oleh karena itu, penelitian ini menerapkan metode penghapusan pantulan spekular (image inpainting) untuk memulihkan area refleksi berlebih pada citra mata, guna meningkatkan kejelasan fitur dan memperbaiki akurasi klasifikasi stadium katarak pada sistem berbasis Vision Transformer.

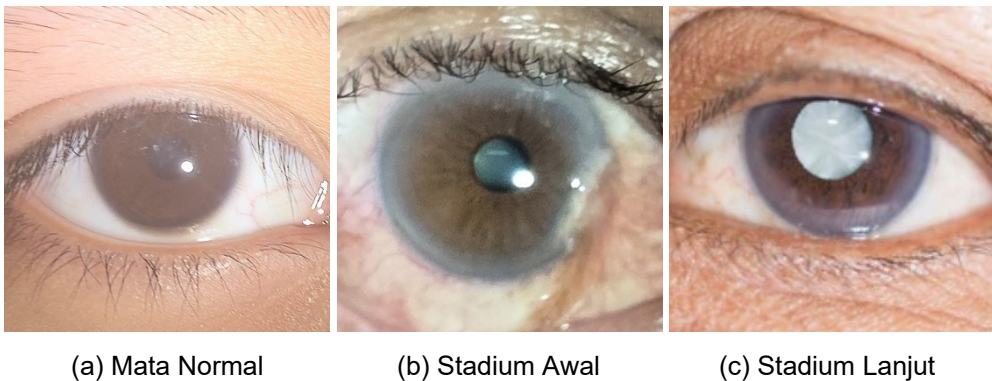
2.3.2 Studi literatur

Tahap awal penelitian ini adalah studi literatur, yang dilakukan dengan menelusuri berbagai sumber ilmiah seperti jurnal, artikel konferensi, buku, dan publikasi daring terkait computer vision, deep learning, serta penerapan Vision Transformer (ViT) dalam klasifikasi citra medis. Melalui tahap ini, juga dikaji metode dan teknik yang telah digunakan dalam klasifikasi penyakit katarak, termasuk inpainting, edge detection, ROI cropping, dan augmentasi data. Studi literatur ini juga digunakan

untuk mengidentifikasi research gap dan merumuskan kontribusi penelitian dalam peningkatan akurasi dan efisiensi klasifikasi stadium katarak berbasis telemedisin.

2.3.3 Pengumpulan data

Tahap ini dilakukan pengumpulan dataset yang digunakan dalam penelitian ini. Dataset yang dikumpulkan adalah data berupa foto mata pasien yang diambil menggunakan kamera ultra-wide Samsung S24 Ultra dengan resolusi 12 megapixel dan pengaturan 3x zoom optik dengan jarak 10-15 centimeter dari mata pasien menggunakan bantuan pencahayaan *flashlight*. Gambar mata pasien yang telah dikumpulkan, kemudian dikelompokkan ke dalam tiga kategori yaitu stadium Awal, stadium Lanjut, dan Normal. Dataset yang dikumpulkan nantinya digunakan untuk melatih model yang dibangun. Beberapa sampel gambar mata yang telah dikumpulkan dapat dilihat pada gambar 6 berikut:



Gambar 6. Sampel gambar mata masing-masing kelas

2.3.4 Mengelolah data

Dilakukan pengolahan data gambar mata untuk memperoleh informasi visual yang relevan dalam proses klasifikasi stadium penyakit katarak. Proses ini mencakup beberapa langkah utama yang bertujuan untuk meningkatkan kualitas gambar dan memperkuat kemampuan model dalam mengenali pola visual yang signifikan.

Tahapan awal yang dilakukan adalah pra-pemrosesan (*preprocessing*), yang meliputi kegiatan anotasi dan augmentasi data. Pada tahap anotasi, setiap gambar mata diberikan label sesuai dengan kategori stadium penyakit, yaitu normal, katarak awal, dan katarak lanjut berdasarkan hasil diagnosa dokter spesialis mata menggunakan peralatan medis standar.

Sementara itu, tahap augmentasi dilakukan untuk memperluas variasi data latih sehingga model menjadi lebih robust terhadap berbagai kondisi pengambilan gambar di dunia nyata. Beberapa teknik augmentasi yang diterapkan antara lain rotasi (*rotate*) untuk mensimulasikan perbedaan sudut pandang kamera, pembalikan horizontal (*flip horizontal*) untuk menyeimbangkan distribusi orientasi mata kiri dan kanan, serta penambahan efek blur dan noise untuk meniru variasi kualitas gambar akibat getaran tangan, pencahayaan yang tidak merata, atau fokus kamera yang kurang stabil.

2.3.5 Perancangan sistem

Tahap perancangan sistem dimulai dengan mengidentifikasi kebutuhan dan permasalahan utama dalam klasifikasi stadium katarak menggunakan gambar mata dari kamera *smartphone*. Selanjutnya dilakukan perancangan arsitektur model berbasis Vision Transformer yang dilengkapi Local Window Attention (LWA) dan Regional Of Interest Attention untuk meningkatkan fokus pada area paling relevan.

Tahap ini mencakup penentuan ukuran *patch*, dimensi *embedding*, jumlah *head*, dan kedalaman *encoder*. Strategi pelatihan seperti pemilihan optimizer AdamW, pengaturan learning rate, early stopping, dan augmentasi juga dirancang untuk memastikan model dapat belajar secara stabil. Metrik evaluasi dipilih berupa akurasi, *precision*, *recall*, dan *macro-F1* score agar kemampuan model dalam membedakan setiap stadium dapat dinilai secara seimbang dan valid.

2.3.6 Pengujian dan analisis

Tahap pengujian dilakukan dengan menggunakan dataset uji terpisah untuk mengevaluasi kinerja model. Setiap gambar dari dataset uji dianalisis untuk klasifikasi stadium penyakit katarak. Hasilnya dievaluasi menggunakan metrik seperti *precision*, *recall*, dan *F1-score* untuk memvalidasi keandalan sistem dalam memberikan prediksi klasifikasi yang optimal.

2.3.7 Evaluasi

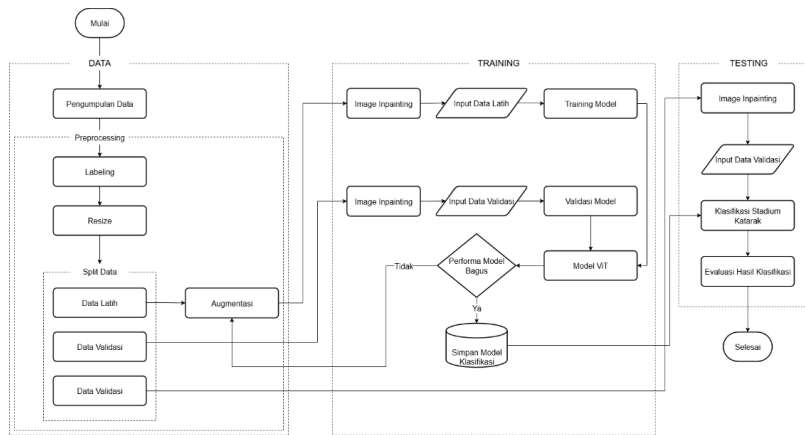
Evaluasi sistem ini dilakukan untuk mengukur sejauh mana sistem berhasil melakukan klasifikasi stadium penyakit katarak secara akurat dan konsisten. Proses evaluasi mencakup pengujian model terhadap data uji (testing dataset) yang tidak digunakan selama tahap pelatihan, sehingga dapat menggambarkan kemampuan generalisasi model terhadap data baru. Evaluasi dilakukan dengan membandingkan performa model Vision Transformer, Vision Transformer dengan Image Inpainting, dan Optimalisasi Vision Transformer dengan Image Inpainting dengan melihat nilai akurasi dan *macro-F1*.

2.3.8 Pembuatan laporan

Bagian ini membahas hasil eksperimen dan analisis yang dilakukan selama proses penelitian, dengan tujuan untuk menginterpretasikan temuan-temuan yang diperoleh serta menilai kinerja sistem klasifikasi yang telah dikembangkan. Analisis dilakukan secara kuantitatif berdasarkan hasil pengujian model terhadap dataset uji, dan secara kualitatif melalui pengamatan terhadap pola prediksi yang dihasilkan model pada masing-masing kelas stadium katarak. Hasil yang diperoleh kemudian dibandingkan dengan penelitian-penelitian terdahulu guna menilai sejauh mana metode yang diusulkan mampu memberikan peningkatan performa.

2.4 Perancangan Sistem

Alur perancangan sistem pada penelitian ini ditunjukkan pada gambar 7 berikut:



Gambar 7. Flowchart rancangan sistem

2.4.1 Akuisisi Data

1. Pengumpulan Data

Pengumpulan data pada penelitian ini dilakukan dengan cara mengambil foto mata pasien secara langsung di lapangan menggunakan kamera *ultra-wide smartphone* resolusi 12 megapixel dengan pengaturan 3x optical zoom dan posisi kamera sejajar mata pasien dengan jarak 10-15 cm dan dengan bantuan pencahayaan *flashlight*. Semua gambar disimpan dalam format .jpg. Pada proses pengumpulan data menghasilkan jumlah data sebanyak 648 gambar. Teknik pengambilan data dapat dilihat pada gambar 8 berikut.



Gambar 8. Teknik pengambilan gambar mata

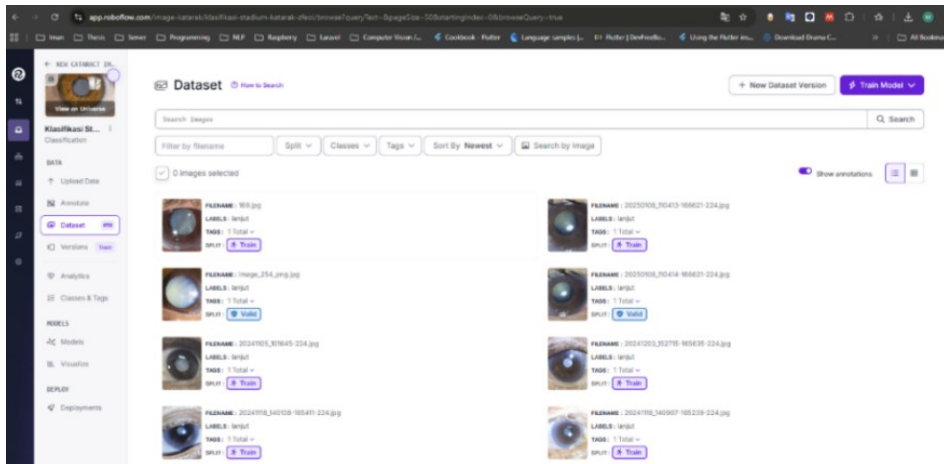
2. Preprocessing

Setelah proses pengumpulan data dilakukan, dilanjutkan ke tahap pra-pemrosesan (*preprocessing*), untuk mempersiapkan gambar sebelum diproses lebih lanjut

dengan tujuan untuk memperoleh hasil yang optimal. Tahapan dalam proses ini meliputi:

a. Labeling

Pada tahap ini dilakukan pelabelan pada dataset menggunakan aplikasi Roboflow. Data dikelompokkan menjadi 3 kategori yaitu Awal, Lanjut dan Normal. Proses pelabelan data dapat dilihat pada gambar 9 berikut.



Gambar 9. Proses pelabelan data

b. *Resize*

Proses *resize* dilakukan dengan mengubah resolusi gambar asli 1172 x 1172 menjadi 224 x 224 piksel. Tujuan dari perubahan ukuran ini adalah untuk menyamakan dimensi seluruh gambar agar sesuai dengan standar input yang direkomendasikan untuk arsitektur ViT-Base. Selain itu, penyesuaian ukuran ini membantu mengurangi kompleksitas data, mempercepat proses pelatihan model, serta mengoptimalkan penggunaan memori selama komputasi.

c. Split Data

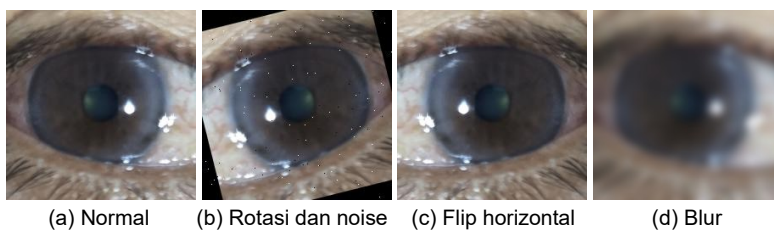
Setelah dilakukan *resize*, selanjutnya dilakukan split data untuk membagi dataset ke dalam tiga kelompok, yaitu data training, data validation, dan data testing. Pembagian data ini bertujuan untuk melatih, memvalidasi, dan menguji kinerja model melakukan klasifikasi stadium katarak. Adapun pembagian data dilakukan dengan proporsi 70% data training, 10% data validasi, dan 20% data testing. Pembagian data secara detail dapat dilihat pada Tabel 1 berikut.

Tabel 1. Proporsi data training, validasi dan testing

Kelas	Training	Validasi	Testing	Total
Awal	154	22	42	218
Lanjut	147	21	43	211
Normal	153	23	43	219

d. Augmentasi Data

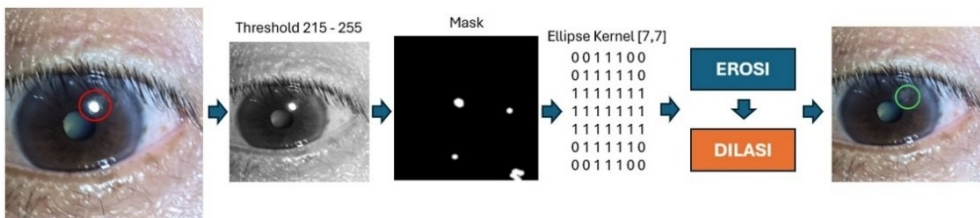
Augmentasi merupakan teknik pengolahan gambar pada machine learning yang digunakan untuk memperbanyak data pelatihan dengan memodifikasi gambar asli (Maharana et al., 2022). Pada penelitian, ini Augmentasi hanya dilakukan pada data training karena tujuannya adalah memperkaya variasi data agar model dapat belajar lebih baik dan menjadi lebih tangguh terhadap berbagai kondisi input. Sebaliknya, data validasi dan testing harus tetap dalam kondisi asli agar dapat memberikan evaluasi yang objektif dan akurat terhadap performa model pada data nyata yang belum pernah dilihat sebelumnya. Data training pada penelitian ini dilakukan teknik augmentasi rotasi berkisar -15° dan $+15^\circ$, blur 0.5px, flip horizontal dan noise 0.5-1 piksel. Data latih yang dihasilkan setelah melalui proses augmentasi sebanyak 1096 gambar. Hasil augmentasi data dapat dilihat pada gambar 10 berikut:



Gambar 10. Sampel hasil augmentasi data

2.4.2 Image Inpainting

Proses *image inpainting* dilakukan untuk menghapus pantulan spekular pada gambar mata yang disebabkan oleh cahaya kilat (*flashlight*). Pantulan tersebut dapat menyulitkan dalam menentukan stadium katarak karena bintik putih terang dari pantulan cahaya sering kali menyerupai kekeruhan lensa (*lens opacity*), yaitu ciri klinis dari katarak. Akibatnya, pantulan tersebut bisa disalahartikan sebagai kekeruhan, padahal sebenarnya hanyalah efek silau. Selain itu, pantulan juga dapat menutupi area katarak yang sebenarnya. Kondisi ini dapat mengganggu distribusi intensitas piksel dan fitur tekstur pada area pupil. Akibatnya, model cenderung lebih fokus pada pantulan bercahaya tinggi daripada pada kekeruhan halus yang penting untuk membedakan antara stadium awal dan stadium lanjut katarak. Proses penghapusan pantulan spekular pada mata dapat dilihat pada gambar 11 berikut.



Gambar 11. Proses penghapusan pantulan spekular

Langkah-langkah penghapusan pantulan spekular dijelaskan sebagai berikut:

1. Mengonversi gambar RGB menjadi *Grayscale*

$$\begin{bmatrix} [51,51,61] & [51,51,61] & [50,50,60] & [50,50,60] & [50,50,60] \\ [53,51,62] & [53,51,62] & [53,51,62] & [53,51,62] & [52,51,59] \\ [55,53,64] & [54,52,63] & [53,51,62] & [50,48,59] & [49,48,56] \\ [55,53,64] & [55,53,64] & [54,52,63] & [52,50,61] & [49,48,56] \\ [53,51,62] & [55,53,64] & [57,55,66] & [56,54,65] & [53,52,60] \end{bmatrix} = \text{RGB}$$

Setiap elemen [R,G,B] menunjukkan intensitas warna merah, hijau dan biru (0-255). Untuk konversi ke *grayscale* maka matriks RGB dikalikan dengan rumus standar BT.601: $0.299R + 0.587G + 0.114B$. Berikut adalah matriks hasil konversi ke *grayscale*.

$$\begin{bmatrix} 52.14 & 52.14 & 51.14 & 51.14 & 51.14 \\ 52.85 & 52.85 & 52.85 & 52.85 & 52.21 \\ 54.85 & 53.85 & 52.85 & 49.85 & 49.21 \\ 54.85 & 54.85 & 53.85 & 51.85 & 49.21 \\ 52.85 & 54.85 & 56.85 & 55.85 & 53.21 \end{bmatrix} = \text{Grayscale}$$

2. Mengubah gambar ke bentuk biner untuk memudahkan deteksi area yang terang dan membentuk mask berdasarkan ambang batas (*threshold*) yang ditetapkan untuk menandai pantulan spekular. Ambang batas yang digunakan untuk menandai piksel intensitas tinggi yaitu 215-255. Nilai minimum 215 dipilih karena glare umumnya sangat terang (mendekati putih).

$$\text{mask}(x,y) = \begin{cases} 255, & \text{Igray}(x, y) > 215 \\ 0, & \text{lainnya} \end{cases}$$

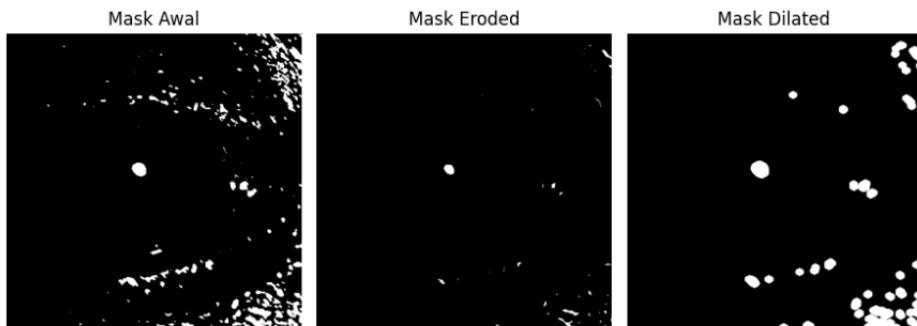
$$\begin{bmatrix} 40 & 45 & 220 \\ 42 & 230 & 240 \\ 50 & 55 & 60 \end{bmatrix} \rightarrow \begin{bmatrix} 0 & 0 & 255 \\ 0 & 255 & 255 \\ 0 & 0 & 0 \end{bmatrix}$$

3. Membentuk kernel yang digunakan untuk sebagai pola acuan pada operasi morfologi yang bergerak di seluruh area gambar, dan setiap kali jendela tersebut menutupi sekelompok piksel, sistem memeriksa hubungan antara bentuk kernel dan nilai piksel di dalamnya. Hasil dari pemeriksaan ini menentukan apakah piksel pusat tetap bernilai putih (255), menjadi hitam (0), atau sebaliknya. Ukuran kernel yang digunakan yaitu 7x7 berbentuk elips.

$$\begin{bmatrix} 0 & 0 & 1 & 1 & 1 & 0 & 0 \\ 0 & 1 & 1 & 1 & 1 & 1 & 0 \\ 1 & 1 & 1 & 1 & 1 & 1 & 1 \\ 1 & 1 & 1 & 1 & 1 & 1 & 1 \\ 1 & 1 & 1 & 1 & 1 & 1 & 1 \\ 0 & 1 & 1 & 1 & 1 & 1 & 0 \\ 0 & 0 & 1 & 1 & 1 & 0 & 0 \end{bmatrix}$$

4. Operasi morfologi erosi dan dilasi (*opening*) dilakukan untuk membersihkan *noise* kecil. Erosi adalah operasi morfologi yang tujuannya untuk menghapus bintik putih kecil yang tidak relevan dengan memanfaatkan kernel untuk

mengidentifikasi dan menghapus area putih yang tidak memiliki cukup piksel putih di sekitarnya. Jika semua piksel dalam area kernel adalah putih, maka piksel pusat tetap putih, dan sebaliknya jika ada satu saja piksel hitam di dalam jendela kernel, maka pusat berubah menjadi hitam. Sedangkan pada proses dilasi, kernel berperan sebaliknya dimana piksel pusat menjadi putih jika terdapat setidaknya satu piksel putih di dalam area kernel. Akibatnya area putih melebar, mengisi celah-celah kecil, dan memperluas batas objek. Bentuk dan ukuran kernel sangat menentukan hasil akhir dari operasi morfologi. Kernel berbentuk persegi cendrung menghasilkan perubahan yang bersudut tajam dan tidak alami, sedangkan kernel elips memberikan halus dan sesuai dengan bentuk pantulan cahaya pada kornea. Gambar 12 menunjukkan perbandingan *mask* sebelum dan setelah dilakukan operasi morfologi, yang menunjukkan kondisi *mask* setelah morfologi lebih bersih dari *mask* awal.



Gambar 12. Perbandingan mask sebelum dan setelah operasi morfologi

5. *Mask biner* yang telah dihasilkan pada tahap sebelumnya digunakan sebagai penanda lokasi piksel yang perlu diganti. Pada proses inpainting, area putih pada mask (bernilai piksel 255) digantikan dengan nilai piksel di sekitarnya. Metode Telea dipilih dalam penelitian ini karena pantulan cahaya pada gambar mata umumnya berukuran kecil, sehingga pendekatan ini mampu melakukan rekonstruksi dengan cepat dan menghasilkan hasil yang natural tanpa mengubah kontur anatomi mata. Berikut ditunjukkan skenario sederhana yang menggambarkan proses penggantian piksel berdasarkan jarak, tanpa mempertimbangkan faktor arah maupun gradien. Misalkan terdapat gambar RGB berukuran 5×5 dengan radius = 5, di mana mask menunjukkan area putih pada koordinat (2,2).

$$\text{Mask} = \begin{bmatrix} 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 255 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \end{bmatrix}$$

$$I_{RGB} = \begin{bmatrix} [46, 47, 45] & [47, 46, 48] & [45, 49, 47] & [46, 48, 49] & [47, 50, 48] \\ [48, 46, 47] & [49, 48, 46] & [47, 50, 48] & [48, 49, 51] & [50, 51, 49] \\ [46, 48, 49] & [48, 47, 50] & [240, 240, 240] & [49, 51, 48] & [51, 49, 50] \\ [47, 49, 48] & [48, 50, 47] & [49, 48, 51] & [50, 52, 49] & [52, 20, 51] \\ [48, 50, 49] & [50, 49, 52] & [51, 48, 50] & [52, 51, 49] & [53, 52, 51] \end{bmatrix}$$

Untuk setiap kanal R, G, dan B dihitung menggunakan rumus (5):

$$I(p) = \frac{\sum_q w_{pq} I(q)}{\sum_q w_{pq}} \text{ dengan } w_{pq} = \frac{1}{\|p-q\|^2}, q \neq p \quad (5)$$

Nilai d (jarak) dihitung berdasarkan jarak Euclidean yaitu dari pusat ke tetangga dengan menggunakan persamaan (6):

$$d = \sqrt{x^2 + y^2} \quad (6)$$

Selanjutnya piksel tetangga dikelompokkan berdasarkan jarak Euclidean dari pusat. Kelompok piksel tetangga dapat dilihat pada tabel 2 berikut.

Tabel 2. Kelompok piksel tetangga piksel pusat

Posisi relatif (x,y)	Jarak d	Jumlah	Bobot per titik $1/d^2$	Total Bobot
(0,±1), (±1,0)	1 (ortogonal)	4	1.000	4.000
(±1,±1)	$\sqrt{2}$ (diagonal)	4	0.500	2.000
(0,±2), (±2,0)	2	4	0.250	1.000
(±1,±2), (±2,±1)	$\sqrt{5}$	8	0.200	1.600
(±2,±2)	$\sqrt{8}$	4	0.125	0.500

Jumlah bobot = 4 + 2 + 1 + 1.6 + 0.5 = 9.1

Setelah jumlah bobot di dapatkan, berikutnya adalah mengalikan setiap nilai tetangga dengan bobotnya, lalu menjumlahkan per kelompok:

Kanal R:

- $d = 1$
(48+49+47+49) = 193
193 x 1 = 193
- $d = \sqrt{2}$
(49+48+48+50) = 195
195 x 0.5 = 97.5
- $d = 2$
(45+46+51+51) = 193
193 x 0.25 = 48.25
- $d = \sqrt{5}$

$$(47+48+50+46+47+50+52+52) = 392$$

$$392 \times 0.2 = 78.4$$

$$- d = \sqrt{8}$$

$$(46+47+48+53) = 194$$

$$194 \times 0.125 = 24.25$$

$$\text{Total R} = 193.00 + 97.50 + 48.25 + 78.40 + 24.25 = 441.40$$

$$\text{Kanal G: } 196.00 + 99.50 + 48.50 + 78.00 + 24.88 = 446.88$$

$$\text{Kanal B: } 197.00 + 96.50 + 49.00 + 78.60 + 24.13 = 445.23$$

Nilai piksel pengganti:

$$R = 441.40 / 9.1 = 49$$

$$G = 446.88 / 9.1 = 49$$

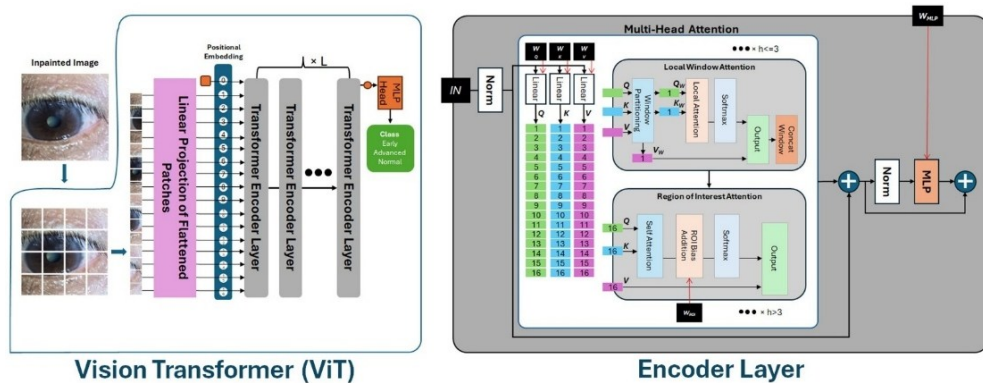
$$B = 445.23 / 9.1 = 49$$

2.4.3 Optimalisasi Arsitektur Vision Transformer

Optimalisasi arsitektur Vision Transformer (ViT) dilakukan dengan memodifikasi komponen utama pada *encoder*, yaitu mekanisme Multi-Head Self-Attention (MHSA). Dua pengembangan utama diterapkan pada penelitian ini, yaitu Local Window Attention (LWA) dan *Region of Interest* (ROI) Attention. Kedua pendekatan ini berfungsi untuk memperbaiki cara ViT menangkap hubungan antar-patch, baik pada konteks lokal maupun pada area yang secara klinis relevan terhadap klasifikasi stadium katarak.

Pada ViT standar, MHSA bekerja secara global sehingga setiap *patch* gambar dapat saling berinteraksi satu sama lain. Mekanisme ini efektif dalam memahami konteks global, tetapi cenderung kurang sensitif terhadap pola lokal yang berukuran kecil seperti area kekeruhan pada lensa mata. Untuk mengatasi hal ini, diterapkan *Local Window Attention* (LWA) sebagai varian dari MHSA yang membatasi perhitungan atensi hanya pada jendela (window) berukuran tetap, misalnya 7×7 patch. Dengan demikian, setiap head hanya memperhatikan token di dalam jendela tertentu tanpa menjangkau seluruh gambar. Pendekatan ini menjaga konsistensi spasial lokal, serta meningkatkan efisiensi dalam mengenali pola detail pada pupil.

Sementara itu, *Region of Interest* (ROI) Attention diterapkan untuk memastikan bahwa perhatian model terfokus pada area penting gambar mata, yaitu lensa bola mata yang meliputi iris dan pupil. Tidak semua bagian gambar mata relevan untuk diagnosis katarak seperti misalnya kulit, bulu mata, maupun noise dari kamera. Oleh karena itu, mask ROI digunakan untuk meningkatkan fokus mekanisme self-attention hanya pada area bola mata. Dengan menambahkan mask ROI sebagai bias dalam perhitungan perhatian, patch yang berada di dalam area ROI memperoleh bobot atensi yang lebih tinggi dibandingkan patch di luar ROI. Hal ini memastikan bahwa model memberikan perhatian yang lebih besar pada area klinis yang relevan untuk klasifikasi stadium katarak. Optimalisasi arsitektur Vision Transformer (ViT) dapat dilihat lebih detail pada gambar 13.



Gambar 13. Arsitektur Vision Transformer (ViT) yang di optimalisasi

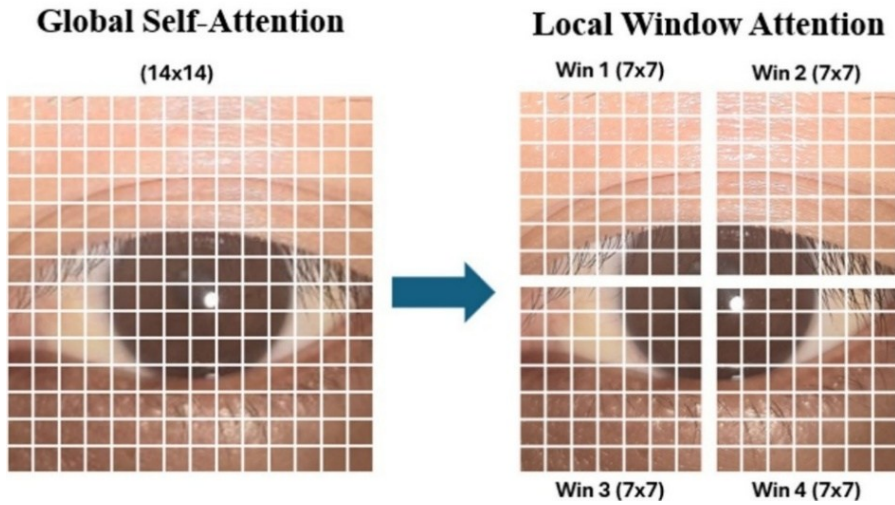
Integrasi antara LWA dan ROI Attention di dalam struktur MHSA menjadikan ViT mampu memahami hubungan lokal dan global secara seimbang. LWA memperkuat representasi fitur spasial lokal, sedangkan ROI Attention memastikan fokus pembelajaran berada pada area diagnostik utama (iris dan pupil). Kombinasi keduanya menghasilkan representasi fitur yang lebih diskriminatif dan stabil, serta meningkatkan performa klasifikasi stadium katarak.

1. *Local window attention* (LWA)

Pada model Vision Transformer (ViT) standar, setiap patch gambar dapat saling berinteraksi satu sama lain melalui mekanisme global self-attention. Mekanisme ini memberikan fleksibilitas penuh bagi setiap patch untuk berkomunikasi dengan seluruh patch lainnya, sehingga mampu menangkap fitur global secara komprehensif (Hu et al., 2023). Namun demikian, pendekatan ini kurang optimal untuk menangkap fitur lokal yang sangat penting dalam gambar medis.

Pada kasus gambar mata penderita katarak, fitur-fitur lokal seperti titik-titik opasitas (*opacity spots*) sering kali terabaikan karena perhatian model menyebar terlalu luas. Akibatnya, patch kritis yang memuat detail kekeruhan dapat tertutupi oleh patch dominan lainnya. Untuk mengatasi hal tersebut, penelitian ini menerapkan mekanisme LWA.

Untuk membatasi area perhatian agar lebih fokus pada konteks lokal, seluruh patch hasil embedding dibagi ke dalam beberapa jendela (*window*) yang berukuran 7x7 seperti yang diperlihatkan pada gambar 14. Karena total patch adalah 14x14, maka terbentuk empat jendela (*window*), yang masing berisi 49 patch. Setiap patch di dalam satu jendela hanya berinteraksi melalui mekanisme self-attention lokal, sehingga memperkuat sensitivitas model terhadap pola tekstur atau opasitas yang berada di area berdekatan.



Gambar 14. Mekanisme *local window attention*

Pada mekanisme Local Window Attention (LWA), setiap token hasil embedding dalam satu window direpresentasikan sebagai tensor input (7):

$$X_w \in \mathbb{R}^{B \times N_w \times D} \quad (7)$$

dengan:

X = Input, B = Ukuran batch,

N_w = Jumlah patch dalam satu window,

D = dimensi embedding.

Untuk setiap jendela yang terdiri dari N_w *patch*, dilakukan proses proyeksi linier terhadap embedding menjadi tiga komponen, yaitu *query* (Q_w), *key* (K_w) dan *value* (V_w) menggunakan matriks bobot $W_Q, W_K, W_V \in \mathbb{R}^{D \times d_h}$ dimana d_h adalah dimensi head dari mekanisme *multi-head attention*. Selanjutnya dihitung skor *local-attention* masing-masing jendela dengan persamaan (8):

$$A_w (\text{Hasil Softmax}) = \text{softmax} \left(\frac{Q_w K_w^T}{\sqrt{d_h}} + B_{LWA} \right) \quad (8)$$

Contoh langkah-langkah kerja LWA dijelaskan sebagai berikut:

- Image input RGB ukuran 4 x 4 dengan *patch size* = 2

$$X = \begin{bmatrix} [10, 8, 6] & [12, 9, 7] & [14, 10, 8] & [16, 11, 9] \\ [11, 9, 7] & [13, 10, 8] & [15, 11, 9] & [17, 12, 10] \\ [12, 10, 8] & [14, 11, 9] & [16, 12, 10] & [18, 13, 11] \\ [13, 11, 9] & [15, 12, 10] & [17, 13, 11] & [19, 14, 12] \end{bmatrix}$$

- Flatten dan *linear projection* (Embedding) dengan dimensi embedding = 8

$$X1 = [10, 8, 6, 12, 9, 7, 11, 9, 7, 13, 10, 8]$$

$$X2 = [14, 10, 8, 16, 11, 9, 15, 11, 9, 17, 12, 10]$$

$$X3 = [12, 10, 8, 14, 11, 9, 13, 11, 9, 15, 12, 10]$$

$$X4 = [16, 12, 10, 18, 13, 11, 17, 13, 11, 19, 14, 12]$$

$$W = \begin{bmatrix} 1 & 0 & 0 & 1 & 0 & 1 & 1 & 1 \\ 0 & 1 & 0 & -1 & 1 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 & -1 & 1 & 0 & 0 \\ 1 & 0 & 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & -1 & 1 & 0 & 1 & 0 \\ 0 & 0 & 1 & 0 & -1 & 0 & 0 & 1 \\ 1 & 0 & 0 & 1 & 0 & -1 & 1 & 0 \\ 0 & 1 & 0 & -1 & 1 & -1 & 0 & 0 \\ 0 & 0 & 1 & 0 & -1 & -1 & 0 & 0 \\ 1 & 0 & 0 & 1 & 0 & 0 & 0 & 1 \\ 0 & 1 & 0 & -1 & 1 & 0 & 1 & 0 \\ 0 & 0 & 1 & 0 & -1 & 0 & 0 & 1 \end{bmatrix}, b = [1 \ -1 \ 0 \ 2 \ 0 \ -2 \ 1 \ 0]$$

$$\begin{aligned} X1 &= (10 \times 1) + (8 \times 0) + (6 \times 0) + (12 \times 1) + (9 \times 0) + (7 \times 0) + (11 \times 1) + (9 \times 0) \\ &\quad + (7 \times 0) + (13 \times 1) + (10 \times 0) + (8 \times 0) + b1 \\ &= (10 + 12 + 11 + 13) + 1 \\ &= 46 + 1 \\ &= 47 \end{aligned}$$

$$x1 = [47 \ 35 \ 28 \ 12 \ 8 \ -5 \ 41 \ 47]$$

$$x2 = [63 \ 43 \ 36 \ 20 \ 8 \ -5 \ 53 \ 61]$$

$$x3 = [55 \ 43 \ 36 \ 12 \ 8 \ -5 \ 49 \ 57]$$

$$x4 = [71 \ 51 \ 44 \ 20 \ 8 \ -5 \ 61 \ 71]$$

- Tambahkan token CLS (*Classification Token*) dan *positional embedding*
Token CLS = [0.5, 0.1, -0.3, 0.2, 0.0, 0.4, -0.2, 0.3]

$$\text{Positional embedding } E = \begin{bmatrix} 0.10, 0.20, 0.30, 0.40, 0.50, 0.60, 0.70, 0.80 \\ 0.15, 0.25, 0.35, 0.45, 0.55, 0.65, 0.75, 0.85 \\ 0.20, 0.30, 0.40, 0.50, 0.60, 0.70, 0.80, 0.90 \\ 0.25, 0.35, 0.45, 0.55, 0.65, 0.75, 0.85, 0.95 \\ 0.30, 0.35, 0.45, 0.55, 0.65, 0.75, 0.85, 0.95 \end{bmatrix}$$

$$Z_0 = \begin{bmatrix} X_{CLS} \\ x1 \\ x2 \\ x3 \\ x4 \end{bmatrix} + E$$

Hasil (Z):

$$Z_0 = \begin{bmatrix} 0.60 & 0.30 & 0.00 & 0.60 & 0.50 & 1.00 & 0.50 & 1.10 \\ 47.15 & 35.25 & 28.35 & 12.45 & 8.55 & -4.35 & 41.75 & 47.85 \\ 63.20 & 43.30 & 36.40 & 20.50 & 8.60 & -4.30 & 53.80 & 61.90 \\ 55.25 & 43.35 & 36.45 & 12.55 & 8.65 & -4.25 & 49.85 & 57.95 \\ 71.30 & 51.40 & 44.50 & 20.60 & 8.70 & -4.20 & 61.90 & 72.00 \end{bmatrix}$$

- Pembentukan Query (Q), Key (K), dan Value (V)

$$h \text{ (head)} = 2, D_h = \frac{D \text{ (depth)}}{h} = \frac{8}{2} = 4$$

Bobot (W)s:

$$W_Q = \begin{bmatrix} 0.5 & 0.3 & 0.2 & 0.1 & 0.1 & 0.2 & 0.3 & 0.4 \\ 0.2 & 0.5 & 0.1 & 0.3 & 0.3 & 0.1 & 0.2 & 0.4 \\ 0.3 & 0.2 & 0.5 & 0.3 & 0.1 & 0.3 & 0.4 & 0.2 \\ 0.1 & 0.4 & 0.3 & 0.5 & 0.2 & 0.2 & 0.1 & 0.3 \\ 0.2 & 0.1 & 0.2 & 0.3 & 0.4 & 0.3 & 0.2 & 0.1 \\ 0.3 & 0.2 & 0.1 & 0.1 & 0.2 & 0.4 & 0.3 & 0.2 \\ 0.4 & 0.3 & 0.2 & 0.1 & 0.1 & 0.3 & 0.5 & 0.2 \\ 0.1 & 0.2 & 0.3 & 0.4 & 0.5 & 0.2 & 0.1 & 0.3 \end{bmatrix} \quad W_K = \begin{bmatrix} 0.2 & 0.2 & 0.1 & 0.0 & 0.0 & 0.1 & 0.2 & 0.2 \\ 0.1 & 0.2 & 0.0 & 0.2 & 0.2 & 0.0 & 0.1 & 0.2 \\ 0.2 & 0.1 & 0.2 & 0.1 & 0.0 & 0.2 & 0.2 & 0.1 \\ 0.0 & 0.2 & 0.2 & 0.2 & 0.1 & 0.1 & 0.0 & 0.2 \\ 0.1 & 0.0 & 0.1 & 0.2 & 0.2 & 0.2 & 0.1 & 0.0 \\ 0.2 & 0.1 & 0.0 & 0.0 & 0.1 & 0.2 & 0.2 & 0.1 \\ 0.2 & 0.2 & 0.1 & 0.0 & 0.0 & 0.2 & 0.2 & 0.1 \\ 0.0 & 0.1 & 0.2 & 0.2 & 0.2 & 0.1 & 0.0 & 0.2 \end{bmatrix}$$

$$W_V = \begin{bmatrix} 0.1 & 0.1 & 0.0 & 0.0 & 0.0 & 0.0 & 0.1 & 0.1 \\ 0.0 & 0.1 & 0.0 & 0.1 & 0.1 & 0.0 & 0.0 & 0.1 \\ 0.1 & 0.0 & 0.1 & 0.0 & 0.0 & 0.1 & 0.1 & 0.0 \\ 0.0 & 0.1 & 0.1 & 0.1 & 0.1 & 0.1 & 0.1 & 0.1 \\ 0.0 & 0.0 & 0.0 & 0.1 & 0.1 & 0.1 & 0.0 & 0.0 \\ 0.1 & 0.0 & 0.0 & 0.0 & 0.0 & 0.1 & 0.1 & 0.0 \\ 0.1 & 0.1 & 0.0 & 0.0 & 0.0 & 0.1 & 0.1 & 0.0 \\ 0.0 & 0.0 & 0.1 & 0.1 & 0.1 & 0.0 & 0.0 & 0.1 \end{bmatrix}$$

Hasil:

$$Q = \begin{bmatrix} 1.13 & 1.19 & 0.96 & 1.19 & 1.08 & 1.30 & 1.09 & 1.47 \\ 62.28 & 63.52 & 54.86 & 52.66 & 47.00 & 54.87 & 62.28 & 64.20 \\ 85.10 & 86.78 & 74.48 & 70.34 & 62.60 & 73.17 & 85.10 & 87.60 \\ 74.49 & 76.27 & 65.41 & 61.85 & 55.07 & 63.77 & 74.49 & 76.82 \\ 96.11 & 97.83 & 84.33 & 79.75 & 71.21 & 82.54 & 96.11 & 98.91 \end{bmatrix}$$

$$K = \begin{bmatrix} 0.50 & 0.61 & 0.50 & 0.50 & 0.54 & 0.63 & 0.50 & 0.67 \\ 26.96 & 34.51 & 27.48 & 23.66 & 19.14 & 25.61 & 26.96 & 35.12 \\ 35.01 & 45.56 & 36.32 & 30.50 & 24.38 & 33.46 & 35.01 & 46.37 \\ 32.66 & 41.22 & 32.77 & 28.15 & 22.82 & 30.72 & 32.66 & 42.03 \\ 40.71 & 52.27 & 41.61 & 34.99 & 28.06 & 38.57 & 40.71 & 53.28 \end{bmatrix}$$

$$V = \begin{bmatrix} 0.21 & 0.20 & 0.17 & 0.25 & 0.19 & 0.20 & 0.21 & 0.26 \\ 11.29 & 13.66 & 8.87 & 10.41 & 9.17 & 7.43 & 11.29 & 14.27 \\ 14.91 & 18.08 & 11.88 & 13.43 & 11.38 & 9.45 & 14.91 & 18.89 \\ 13.73 & 16.10 & 10.70 & 12.25 & 11.00 & 9.07 & 13.73 & 16.91 \\ 17.35 & 20.52 & 13.71 & 15.27 & 13.21 & 11.09 & 17.35 & 21.53 \end{bmatrix}$$

- Proses *local window attention*

Window 1 (kolom 1-4):

$$Q_1 = \begin{bmatrix} 62.28 & 63.52 & 54.86 & 52.66 \\ 85.10 & 86.78 & 74.48 & 70.34 \\ 74.49 & 76.27 & 65.41 & 61.85 \\ 96.11 & 97.83 & 84.33 & 79.75 \end{bmatrix} \quad K_1 = \begin{bmatrix} 26.96 & 34.51 & 27.48 & 23.66 \\ 35.01 & 45.56 & 36.32 & 30.50 \\ 32.66 & 41.22 & 32.77 & 28.15 \\ 40.71 & 52.27 & 41.61 & 34.99 \end{bmatrix}$$

$$V_1 = \begin{bmatrix} 11.29 & 13.66 & 8.87 & 10.41 \\ 14.91 & 18.08 & 11.88 & 13.43 \\ 13.73 & 16.10 & 10.70 & 12.25 \\ 17.35 & 20.52 & 13.71 & 15.27 \end{bmatrix}$$

Matriks bias LWA (B_{LWA}):

$$B_{LWA} = \begin{bmatrix} 0.4 & 0.3 & 0.3 & 0.2 \\ 0.3 & 0.4 & 0.2 & 0.3 \\ 0.3 & 0.2 & 0.4 & 0.3 \\ 0.2 & 0.3 & 0.3 & 0.4 \end{bmatrix}$$

Hitung skor logit (S) per-window ditambahkan bias LWA:

$$S_1 = \frac{Q_1 K_1}{\sqrt{d_h}} = \frac{Q_1 K_1}{2} + B_{LWA}$$

$$\begin{aligned} S_1 &= \begin{bmatrix} 3312.32 & 4336.52 & 3966.25 & 4990.45 \\ 4500.01 & 5891.77 & 5388.61 & 6780.36 \\ 3950.58 & 5172.44 & 4730.63 & 5952.48 \\ 5085.76 & 6658.59 & 6099.98 & 7662.82 \end{bmatrix} + B_{LWA} = \begin{bmatrix} 0.4 & 0.3 & 0.3 & 0.2 \\ 0.3 & 0.4 & 0.2 & 0.3 \\ 0.3 & 0.2 & 0.4 & 0.3 \\ 0.2 & 0.3 & 0.3 & 0.4 \end{bmatrix} \\ &= \begin{bmatrix} 3312.72 & 4336.82 & 3966.55 & 4990.65 \\ 4500.31 & 5892.17 & 5388.81 & 6780.66 \\ 3950.88 & 5172.64 & 4731.03 & 5952.78 \\ 5085.96 & 6658.89 & 6100.28 & 7663.22 \end{bmatrix} \end{aligned}$$

Normalisasi dengan Softmax:

$$A_1 \text{ (Hasil Softmax)} = \text{Softmax}(S_1) = \begin{bmatrix} 0.08 & 0.22 & 0.15 & 0.55 \\ 0.07 & 0.18 & 0.14 & 0.61 \\ 0.09 & 0.20 & 0.15 & 0.56 \\ 0.06 & 0.16 & 0.12 & 0.66 \end{bmatrix}$$

Hitung Z_1 (Output) dari jendela pertama:

$$\begin{aligned} Z_1 &= A_1 \times V_1 \\ &= \begin{bmatrix} 15.79 & 18.77 & 12.47 & 14.02 \\ 15.98 & 18.98 & 12.62 & 14.18 \\ 15.77 & 18.75 & 12.46 & 14.01 \\ 16.16 & 19.19 & 12.77 & 14.32 \end{bmatrix} \end{aligned}$$

Window 2 (kolom 5-8):

$$Q = \begin{bmatrix} 47.00 & 54.87 & 62.28 & 64.20 \\ 62.60 & 73.17 & 85.10 & 87.60 \\ 55.07 & 63.77 & 74.49 & 76.82 \\ 71.21 & 82.54 & 96.11 & 98.91 \end{bmatrix} \quad K = \begin{bmatrix} 19.14 & 25.61 & 26.96 & 35.12 \\ 24.38 & 33.46 & 35.01 & 46.37 \\ 22.82 & 30.72 & 32.66 & 42.03 \\ 28.06 & 38.57 & 40.71 & 53.28 \end{bmatrix}$$

$$V = \begin{bmatrix} 9.17 & 7.43 & 11.29 & 14.27 \\ 11.38 & 9.45 & 14.91 & 18.89 \\ 11.00 & 9.07 & 13.73 & 16.91 \\ 13.21 & 11.09 & 17.35 & 21.53 \end{bmatrix}$$

$$S_2 = \begin{bmatrix} 3119.69 & 4069.89 & 3745.57 & 4695.78 \\ 4221.73 & 5508.31 & 5068.95 & 6355.54 \\ 3696.98 & 4823.39 & 4439.05 & 5565.47 \\ 4771.03 & 6224.88 & 5728.69 & 7182.54 \end{bmatrix}$$

Normalisasi dengan Softmax (A_2):

$$A_2 = \text{Softmax}(S_2) = \begin{bmatrix} 0.10 & 0.25 & 0.18 & 0.47 \\ 0.06 & 0.24 & 0.15 & 0.55 \\ 0.08 & 0.24 & 0.17 & 0.51 \\ 0.05 & 0.22 & 0.14 & 0.59 \end{bmatrix}$$

Hitung Z_2 (Output) dari jendela kedua:

$$Z_2 = A_2 \times V_2$$

$$Z_2 = \begin{bmatrix} 11.96 & 9.96 & 15.49 & 19.32 \\ 12.18 & 10.16 & 15.83 & 19.74 \\ 12.08 & 10.07 & 15.68 & 19.55 \\ 12.28 & 10.25 & 15.99 & 19.92 \end{bmatrix}$$

Gabungkan hasil output jendela 1 (Z_1) dan jendela 2 (Z_2) dan tambahkan kembali token CLS:

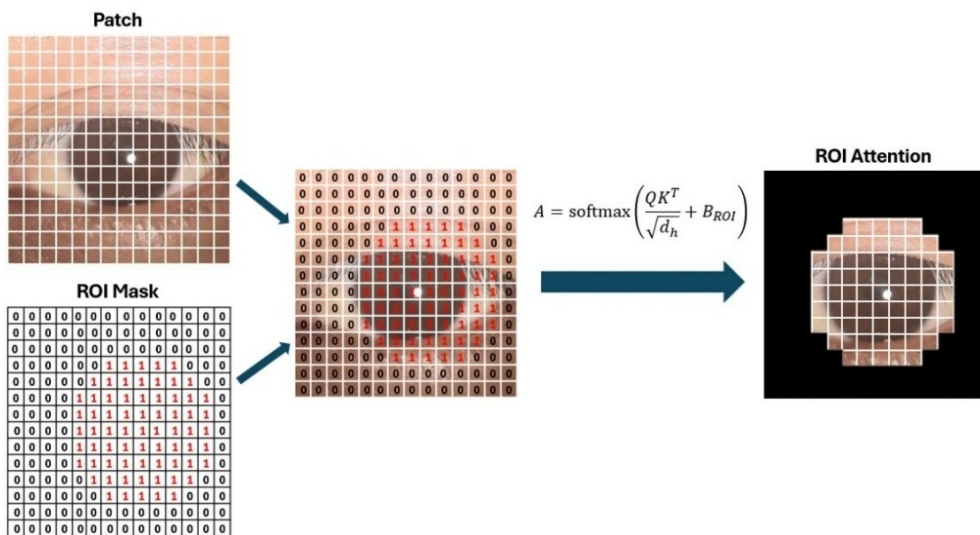
$$Z = \begin{bmatrix} 0.60 & 0.30 & 0.00 & 0.60 & 0.50 & 1.00 & 0.50 & 1.10 \\ 15.79 & 18.77 & 12.47 & 14.02 & 11.96 & 9.96 & 15.49 & 19.32 \\ 15.98 & 18.98 & 12.62 & 14.18 & 12.18 & 10.16 & 15.83 & 19.74 \\ 15.77 & 18.75 & 12.46 & 14.01 & 12.08 & 10.07 & 15.68 & 19.55 \\ 16.16 & 19.19 & 12.77 & 14.32 & 12.28 & 10.25 & 15.99 & 19.92 \end{bmatrix}$$

2. Region of interest (ROI) attention

Pada klasifikasi stadium katarak, tidak semua area gambar mata memiliki relevansi diagnostik. Area di luar bola mata, seperti kulit, bulu mata, atau pantulan cahaya dari kamera, tidak berkontribusi terhadap identifikasi tingkat keparahan katarak. Fokus utama dalam diagnosis justru terletak pada area lensa mata, khususnya iris dan

pupil. Untuk mengatasi hal tersebut, digunakan mekanisme *Region of Interest (ROI) Attention* yang bertujuan meningkatkan fokus perhatian model pada area penting gambar, yaitu bagian bola mata. Mekanisme ini memanfaatkan ROI mask untuk menandai area relevan yang harus mendapatkan perhatian lebih besar (Faizal et al., 2023).

Untuk memperkuat perhatian pada area iris dan pupil, sebuah bias matriks ROI ditambahkan ke dalam perhitungan *attention*. Matriks ini memberikan nilai 2.0 untuk patch yang berada di dalam area ROI dan 0.0 untuk patch di luar ROI. Penambahan bias ini menyebabkan bobot perhatian pada area bola mata meningkat secara signifikan, sehingga model lebih sensitif terhadap perubahan intensitas, tekstur, dan kekeruhan pada lensa mata yang merupakan fitur-fitur penting dalam penentuan stadium katarak. Gambar 15 menunjukkan mekanisme *region of interest attention (ROI)*.



Gambar 15. Mekanisme *region of interest (ROI) attention*

Berikut adalah langkah-langkah Region of Interest Attention dengan menggunakan output (Z) dari mekanisme Local Window Attention sebelumnya sebagai input:

- Membentuk mask ROI berdasarkan dimensi input (5x8)

$$\text{ROI Mask} = \begin{bmatrix} 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 1 & 1 & 1 & 0 & 0 \\ 0 & 0 & 1 & 1 & 1 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 & 1 & 0 & 0 & 0 \end{bmatrix}$$

Flatten ROI Mask: [0, 1, 1, 1, 1]

CLS : [0 0 0 0 0 0 0 0] = 0

X1 : [0 0 0 1 1 0 0 0] = 1

$$X2 : [0 \ 0 \ 1 \ 1 \ 1 \ 1 \ 0 \ 0] = 1$$

$$X3 : [0 \ 0 \ 1 \ 1 \ 1 \ 1 \ 0 \ 0] = 1$$

$$X4 : [0 \ 0 \ 0 \ 1 \ 1 \ 0 \ 0 \ 0] = 1$$

- Membentuk Query (Q), Key (K), dan Value (V)

Bobot:

$$W_Q = \begin{bmatrix} 0.3 & 0.1 & 0.2 & 0.1 & 0.0 & 0.2 & 0.1 & 0.3 \\ 0.1 & 0.2 & 0.1 & 0.0 & 0.1 & 0.3 & 0.1 & 0.2 \\ 0.2 & 0.1 & 0.3 & 0.1 & 0.1 & 0.2 & 0.0 & 0.1 \\ 0.1 & 0.1 & 0.1 & 0.2 & 0.0 & 0.1 & 0.2 & 0.1 \\ 0.2 & 0.0 & 0.2 & 0.1 & 0.1 & 0.3 & 0.1 & 0.0 \\ 0.0 & 0.1 & 0.2 & 0.1 & 0.1 & 0.2 & 0.0 & 0.2 \\ 0.1 & 0.2 & 0.1 & 0.1 & 0.0 & 0.1 & 0.3 & 0.1 \\ 0.2 & 0.1 & 0.0 & 0.2 & 0.1 & 0.2 & 0.1 & 0.2 \end{bmatrix} \quad W_K = \begin{bmatrix} 0.2 & 0.1 & 0.1 & 0.1 & 0.0 & 0.1 & 0.1 & 0.2 \\ 0.1 & 0.1 & 0.1 & 0.0 & 0.1 & 0.2 & 0.1 & 0.1 \\ 0.1 & 0.1 & 0.2 & 0.1 & 0.1 & 0.1 & 0.0 & 0.1 \\ 0.1 & 0.1 & 0.1 & 0.1 & 0.0 & 0.1 & 0.1 & 0.1 \\ 0.1 & 0.0 & 0.1 & 0.1 & 0.1 & 0.2 & 0.1 & 0.0 \\ 0.0 & 0.1 & 0.1 & 0.1 & 0.1 & 0.1 & 0.0 & 0.1 \\ 0.1 & 0.1 & 0.1 & 0.1 & 0.0 & 0.1 & 0.2 & 0.1 \\ 0.1 & 0.1 & 0.0 & 0.1 & 0.1 & 0.1 & 0.1 & 0.1 \end{bmatrix}$$

$$W_V = \begin{bmatrix} 0.1 & 0.1 & 0.1 & 0.0 & 0.1 & 0.0 & 0.1 & 0.1 \\ 0.0 & 0.1 & 0.1 & 0.1 & 0.0 & 0.1 & 0.1 & 0.0 \\ 0.1 & 0.0 & 0.1 & 0.1 & 0.1 & 0.0 & 0.1 & 0.0 \\ 0.0 & 0.1 & 0.0 & 0.1 & 0.1 & 0.1 & 0.0 & 0.1 \\ 0.1 & 0.0 & 0.1 & 0.0 & 0.1 & 0.1 & 0.0 & 0.1 \\ 0.1 & 0.1 & 0.0 & 0.1 & 0.0 & 0.1 & 0.1 & 0.0 \\ 0.0 & 0.1 & 0.1 & 0.0 & 0.1 & 0.0 & 0.1 & 0.1 \\ 0.1 & 0.0 & 0.1 & 0.1 & 0.0 & 0.1 & 0.0 & 0.1 \end{bmatrix}$$

Hasil:

$$Q = \begin{bmatrix} 0.64 & 0.49 & 0.56 & 0.60 & 0.29 & 0.89 & 0.52 & 0.77 \\ 18.31 & 14.01 & 16.11 & 13.24 & 7.25 & 23.68 & 14.04 & 18.54 \\ 18.60 & 14.23 & 16.35 & 13.46 & 7.37 & 24.05 & 14.27 & 18.83 \\ 18.39 & 14.07 & 16.17 & 13.32 & 7.29 & 23.79 & 14.12 & 18.62 \\ 18.79 & 14.38 & 16.52 & 13.59 & 7.44 & 24.29 & 14.42 & 19.03 \end{bmatrix}$$

$$K = \begin{bmatrix} 0.42 & 0.41 & 0.35 & 0.43 & 0.29 & 0.54 & 0.41 & 0.47 \\ 12.36 & 10.58 & 11.09 & 9.90 & 7.25 & 14.85 & 11.08 & 12.16 \\ 12.55 & 10.75 & 11.26 & 10.07 & 7.37 & 15.08 & 11.27 & 12.35 \\ 12.41 & 10.63 & 11.13 & 9.96 & 7.29 & 14.92 & 11.15 & 12.21 \\ 12.68 & 10.86 & 11.37 & 10.17 & 7.44 & 15.24 & 11.39 & 12.48 \end{bmatrix}$$

$$V = \begin{bmatrix} 0.32 & 0.30 & 0.30 & 0.30 & 0.22 & 0.35 & 0.24 & 0.33 \\ 6.95 & 7.40 & 9.38 & 7.45 & 6.97 & 7.40 & 7.25 & 7.66 \\ 7.07 & 7.51 & 9.53 & 7.57 & 7.08 & 7.52 & 7.36 & 7.79 \\ 6.99 & 7.43 & 9.43 & 7.48 & 7.00 & 7.45 & 7.27 & 7.71 \\ 7.14 & 7.59 & 9.63 & 7.64 & 7.15 & 7.60 & 7.44 & 7.87 \end{bmatrix}$$

Hitung skor logit ditambahkan bias ROI:

$$S = \frac{QK}{\sqrt{d_h}} = \frac{QK^T}{2} + B_{ROI}$$

$$S = \begin{bmatrix} 1.03 & 27.84 & 28.29 & 27.98 & 28.58 \\ 27.06 & 734.75 & 746.43 & 738.20 & 754.13 \\ 27.49 & 746.34 & 758.20 & 749.85 & 766.03 \\ 27.19 & 738.18 & 749.91 & 741.65 & 757.65 \\ 27.77 & 753.98 & 765.96 & 757.53 & 773.87 \end{bmatrix} + [0, 2, 2, 2, 2]$$

$$S = \begin{bmatrix} 1.03 & 29.84 & 30.29 & 29.98 & 30.58 \\ 27.06 & 736.75 & 748.43 & 740.20 & 756.13 \\ 27.49 & 748.34 & 760.20 & 751.85 & 768.03 \\ 27.19 & 740.18 & 751.91 & 743.65 & 759.65 \\ 27.77 & 755.98 & 767.96 & 759.53 & 775.87 \end{bmatrix}$$

Normalisasi dengan Softmax:

$$A = \text{Softmax}(S) = \begin{bmatrix} 0.01 & 0.20 & 0.26 & 0.22 & 0.31 \\ 0.01 & 0.01 & 0.02 & 0.01 & 0.95 \\ 0.01 & 0.01 & 0.02 & 0.01 & 0.95 \\ 0.01 & 0.01 & 0.02 & 0.01 & 0.95 \\ 0.01 & 0.01 & 0.02 & 0.01 & 0.95 \end{bmatrix}$$

Hitung Z (Output):

$$Z = A \times V$$

$$Z = \begin{bmatrix} 6.98 & 7.42 & 9.42 & 7.48 & 6.99 & 7.43 & 7.27 & 7.70 \\ 7.07 & 7.51 & 9.53 & 7.56 & 7.08 & 7.52 & 7.36 & 7.79 \\ 7.07 & 7.51 & 9.53 & 7.56 & 7.08 & 7.52 & 7.36 & 7.79 \\ 7.07 & 7.51 & 9.53 & 7.56 & 7.08 & 7.52 & 7.36 & 7.79 \\ 7.07 & 7.51 & 9.53 & 7.56 & 7.08 & 7.52 & 7.36 & 7.79 \end{bmatrix}$$

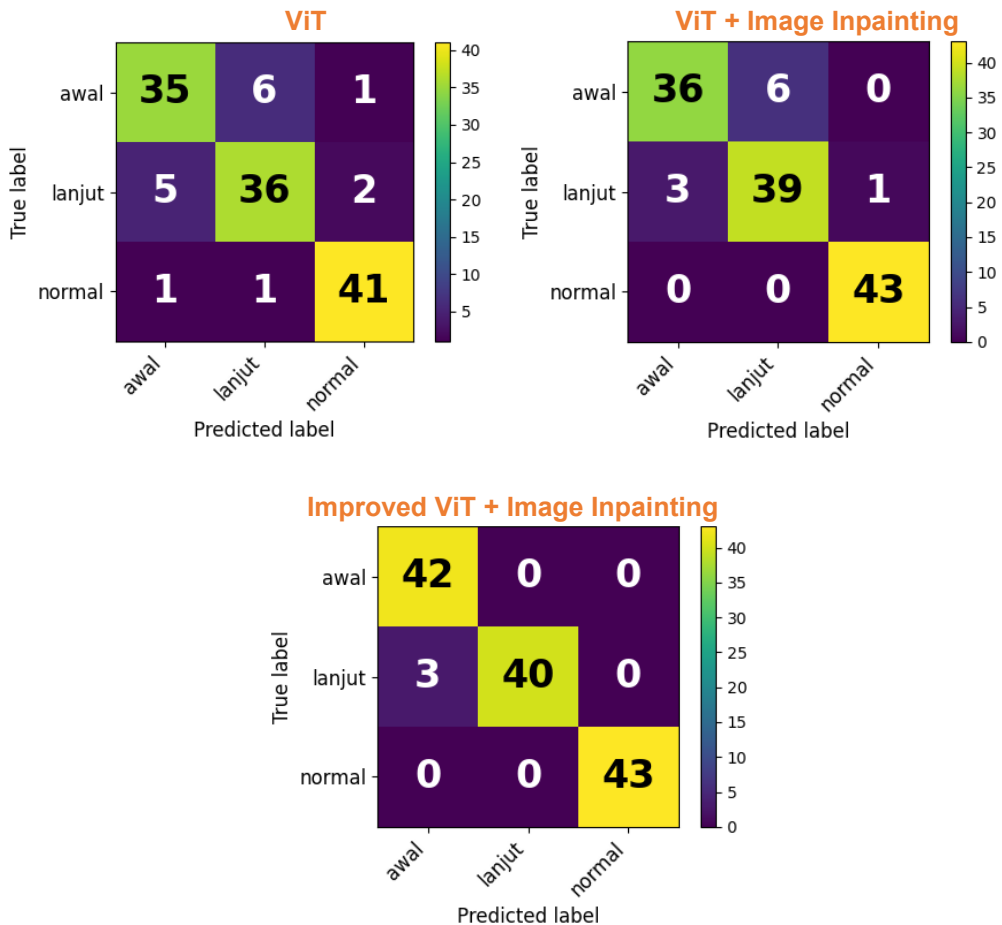
2.5 Evaluasi

Tahapan evaluasi dilakukan untuk menilai kinerja model klasifikasi stadium katarak yang telah dilatih menggunakan parameter yang ditunjukkan pada Tabel 3.

Tabel 3. Hyperparameter yang digunakan dalam pelatihan model

Hyperparameter	Value
Width	224
Height	224
Learning Rate	0.0001
Number Heads	12
Depth	6
Embed Dimension	768
Batch Size	8

Evaluasi performa dilakukan menggunakan *confusion matrix* yang terdiri atas tiga komponen utama, yaitu *True Positive* (TP), *False Positive* (FP), dan *False Negative* (FN) yang dapat dilihat pada gambar 16 berikut.



Gambar 16. *Confusion matrix* antar model

Dari *confusion matrix* tersebut dihitung sejumlah metrik evaluasi, yaitu akurasi, *precision*, *recall*, *F1-score*, dan *macro-F1* (Eyo & Abbey, 2022). Nilai akurasi merepresentasikan proporsi prediksi yang benar terhadap keseluruhan data uji, *precision* menunjukkan ketepatan model dalam memprediksi setiap kelas, sedangkan *recall* mengukur kemampuan model dalam mendeteksi seluruh sampel dari kelas sebenarnya. *F1-score* menggambarkan keseimbangan antara *precision* dan *recall*, sementara *macro-F1* merupakan rata-rata *F1-score* dari seluruh kelas yang mencerminkan performa keseluruhan model secara seimbang. Skenario dengan nilai *macro-F1* tertinggi dipilih sebagai model terbaik karena menunjukkan keseimbangan kinerja antar kelas dan kemampuan generalisasi yang baik dalam mendeteksi stadium katarak.