

Evaluasi Komparatif Swin Transformer dan MedViT untuk Klasifikasi Citra Kanker Kandung Kemih

Suryatna Sacadibrata¹, Sajarwo Anggai², Arya Adhyaksa Waskita³, Tukiayat⁴

^{1,2,3,4} Teknik Informatika S-2, Program Pascasarjana, Universitas Pamulang, Kota Tangerang Selatan, Banten
Email: ¹sacadibrata131003@gmail.com, ²dosen02832@unpam.ac.id, ³aawaskita@unpam.ac.id, ⁴dosen02711@unpam.ac.id

Abstrak—Kanker kandung kemih merupakan salah satu kanker urologi dengan tingkat kejadian dan kekambuhan yang tinggi, sehingga memerlukan metode deteksi dini yang akurat dan stabil untuk mendukung pengambilan keputusan klinis. Perkembangan model deep learning berbasis transformer membuka peluang untuk meningkatkan akurasi klasifikasi citra medis. Penelitian ini bertujuan membandingkan kinerja dua arsitektur vision transformer, yaitu MedViT dan Swin Transformer, dalam mengklasifikasikan citra kanker kandung kemih ke dalam empat kelas: High-Grade Carcinoma (HGC), Low-Grade Carcinoma (LGC), Non-Specific Tumor (NST), dan Non-Tumor Lesion (NTL). Dataset citra diperoleh dari platform Kaggle pada bulan Juli 2024 dan terdiri atas 1.754 gambar, masing-masing HGC=469, LGC=647, NST=504, dan NTL=134 gambar. Dataset dibagi menjadi data latih, validasi, dan uji dengan rasio 70:15:15 menggunakan pembagian acak terkontrol. Proses preprocessing meliputi resizing, normalisasi piksel, serta augmentasi untuk meningkatkan generalisasi model. Evaluasi model menggunakan akurasi, precision, recall, F1-score, confusion matrix, dan ROC-AUC. Hasil eksperimen menunjukkan bahwa MedViT mencapai akurasi data uji sebesar 95,49% dengan rata-rata ROC-AUC 0,97, sedangkan Swin Transformer memperoleh performa lebih tinggi dengan akurasi 98,87% dan ROC-AUC sempurna sebesar 1,00 pada seluruh kelas. Selain itu, Swin Transformer menunjukkan stabilitas training yang lebih baik dibandingkan MedViT, terutama pada kelas minoritas. Berdasarkan temuan ini dapat disimpulkan bahwa Swin Transformer memiliki potensi lebih unggul untuk diterapkan pada sistem klasifikasi kanker kandung kemih, meskipun MedViT tetap menawarkan efisiensi perhitungan dan kemampuan integrasi fitur lokal-global.

Kata Kunci—Kanker Kandung Kemih; Klasifikasi Citra Medis; MedViT; Swin Transformer; Vision Transformer

I. PENDAHULUAN

Kanker merupakan salah satu penyebab utama kematian di dunia dan menjadi tantangan besar dalam sistem kesehatan global [1]. Perkembangan teknologi kedokteran telah meningkatkan kemampuan deteksi dan penanganan kanker, namun akurasi diagnosis awal dan klasifikasi jenis kanker masih menjadi faktor penentu keberhasilan terapi. Oleh karena itu, pemanfaatan teknologi komputasi cerdas, khususnya dalam analisis citra medis, terus dikembangkan untuk mendukung proses diagnosis yang lebih cepat, objektif, dan konsisten [2].

Salah satu jenis kanker yang memiliki tingkat prevalensi dan kekambuhan tinggi adalah kanker kandung kemih (bladder cancer). Kanker kandung kemih (*bladder cancer*) merupakan salah satu keganasan yang paling sering terjadi pada saluran genitourinari [3]. Penyakit ini ditandai dengan pertumbuhan sel abnormal yang tidak terkendali pada jaringan kandung kemih, terutama pada lapisan urotelium yang melapisi bagian dalam kandung kemih. Lebih dari 90% kasus kanker kandung kemih diklasifikasikan sebagai urothelial bladder carcinoma (UBC), yang sebelumnya dikenal sebagai *transitional cell carcinoma* [4]. Penyakit ini umumnya diklasifikasikan ke dalam beberapa sub tipe histopatologis, seperti High-Grade Carcinoma (HGC), Low-Grade Carcinoma (LGC), Non-Specific Tumor (NST), dan Non-Tumor Lesion (NTL). Setiap sub tipe memiliki karakteristik morfologis dan implikasi klinis yang berbeda, sehingga klasifikasi yang akurat sangat penting dalam menentukan strategi pengobatan dan prognosis pasien.

Dalam praktik klinis, klasifikasi kanker kandung kemih masih mengandalkan pengamatan visual citra histopatologi oleh ahli patologi. Proses ini memerlukan pengalaman tinggi dan bersifat subjektif, sehingga berpotensi menimbulkan variasi interpretasi antar-pengamat, terutama pada kasus dengan perbedaan morfologi yang halus. Selain itu, meningkatnya jumlah sampel klinis juga menambah beban kerja tenaga medis, yang pada akhirnya dapat memengaruhi konsistensi dan efisiensi proses diagnosis.

Kemajuan pesat dalam bidang deep learning telah membuka peluang baru dalam analisis citra medis. Model berbasis Convolutional Neural Network (CNN) telah banyak digunakan dan terbukti efektif dalam mengekstraksi fitur visual dari citra medis. Namun, CNN memiliki keterbatasan dalam menangkap hubungan global antar fitur karena sifat operasinya yang berfokus pada area lokal. Keterbatasan ini menjadi tantangan tersendiri pada citra histopatologi kanker kandung kemih yang memiliki pola tekstur kompleks dan variasi struktur pada berbagai skala.

Untuk mengatasi keterbatasan tersebut, pendekatan berbasis Vision Transformer (ViT) mulai banyak diterapkan dalam tugas klasifikasi citra. Vision Transformer (ViT) merupakan arsitektur deep learning yang mengadaptasi mekanisme *Transformer*, yang

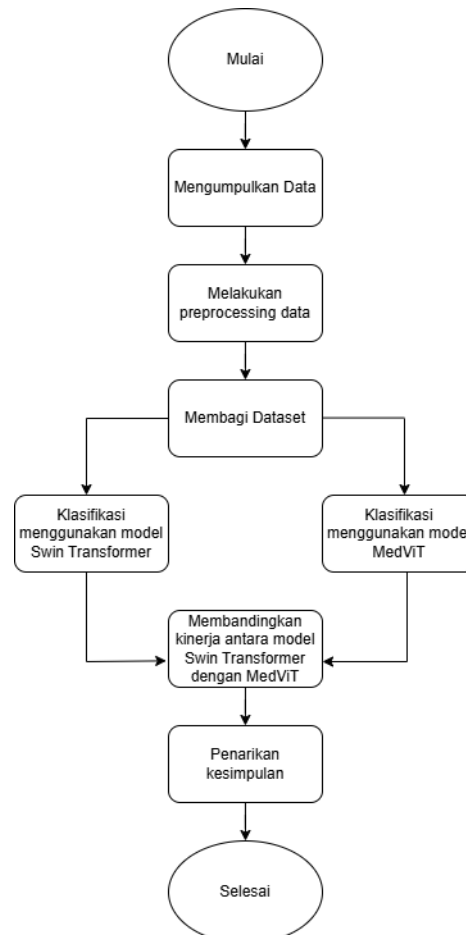
awalnya dikembangkan dalam bidang *Natural Language Processing* (NLP), untuk menyelesaikan permasalahan *computer vision*, khususnya klasifikasi citra. Berbeda dengan *Convolutional Neural Network* (CNN) yang mengandalkan operasi konvolusi untuk mengekstraksi fitur lokal, Vision Transformer memproses citra sebagai urutan (*sequence*) token visual dan memanfaatkan mekanisme *self-attention* untuk menangkap hubungan global antar bagian citra [5]. Mekanisme *self-attention* pada transformer memungkinkan model untuk mempelajari hubungan global antar bagian citra secara lebih efektif. Beberapa arsitektur transformer modern telah dikembangkan, seperti Swin Transformer yang menggunakan pendekatan window-based attention untuk efisiensi komputasi, serta MedViT yang dirancang khusus untuk karakteristik citra medis dengan mengombinasikan kekuatan CNN dan transformer.

Meskipun berbagai arsitektur Vision Transformer telah menunjukkan performa yang menjanjikan dalam domain medis, studi yang secara khusus membandingkan model transformer umum dan transformer yang dioptimalkan untuk citra medis pada klasifikasi kanker kandung kemih masih relatif terbatas. Perbandingan yang sistematis diperlukan untuk memahami keunggulan dan keterbatasan masing-masing arsitektur, terutama dalam konteks data medis yang memiliki ketidakseimbangan kelas dan variasi visual yang tinggi [6].

Berdasarkan latar belakang tersebut, penelitian ini bertujuan untuk menganalisis dan membandingkan performa MedViT dan Swin Transformer dalam klasifikasi citra kanker kandung kemih ke dalam empat kelas utama, yaitu HGC, LGC, NST, dan NTL. Evaluasi dilakukan menggunakan data latih, validasi, dan uji yang terpisah dengan skema pembagian 70/15/15 serta metrik evaluasi komprehensif, termasuk akurasi, precision, recall, F1-score, dan ROC-AUC. Diharapkan hasil penelitian ini dapat memberikan kontribusi dalam pengembangan sistem diagnosis kanker kandung kemih berbasis kecerdasan buatan yang lebih akurat dan andal.

II. METODE PENELITIAN

Bagian ini menjelaskan metodologi penelitian yang digunakan dalam pengembangan dan evaluasi sistem klasifikasi citra kanker kandung kemih berbasis *deep learning*. Metodologi penelitian disusun secara sistematis, mulai dari penjelasan dataset yang digunakan, tahapan eksplorasi dan praproses data citra, perancangan serta implementasi model MedViT dan Swin Transformer, hingga skema evaluasi performa model. Seluruh tahapan dirancang untuk memastikan proses eksperimen berjalan secara objektif, reproduktibel, dan mampu merefleksikan kemampuan generalisasi model dalam mengklasifikasikan subtype kanker kandung kemih secara akurat.



Gambar 1 Alur Penelitian

A. Tahap Pengumpulan Dataset

Dataset yang digunakan dalam penelitian ini merupakan kumpulan citra digital histopatologi kanker kandung kemih (bladder cancer) yang diperoleh dari platform Kaggle pada [link ini](#), sebuah repositori dataset publik yang banyak digunakan dalam penelitian ilmiah dan pengembangan model *machine learning*. Dataset ini dipilih karena menyediakan citra medis dengan kualitas yang memadai serta anotasi kelas yang jelas, sehingga sesuai untuk keperluan klasifikasi citra berbasis *deep learning*.

Seluruh citra dalam dataset disimpan dalam format gambar digital standar, seperti *.jpg* dan *.png*, guna memastikan kompatibilitas dengan berbagai pustaka pemrosesan citra dan kerangka kerja pembelajaran mendalam. Dataset disusun dalam struktur folder berdasarkan kelas, sehingga memudahkan proses pemuatan data secara otomatis menggunakan pustaka PyTorch.

Dataset ini mencakup empat kelas utama, yaitu High-Grade Carcinoma (HGC), Low-Grade Carcinoma (LGC), Non-Specific Tumor (NST), dan Non-Tumor Lesion (NTL). Keempat kelas tersebut merepresentasikan variasi sub tipe kanker kandung kemih yang umum dijumpai dalam praktik klinis dan memiliki karakteristik morfologis yang berbeda pada citra histopatologi.

Secara keseluruhan, dataset yang digunakan dalam penelitian ini berjumlah 1.754 citra, yang kemudian dibagi ke dalam data latih, data validasi, dan data uji untuk mendukung proses pelatihan dan evaluasi model. Distribusi jumlah citra antar kelas tidak sepenuhnya seimbang, dengan kelas LGC dan NST memiliki jumlah data yang lebih banyak dibandingkan kelas NTL. Kondisi ini mencerminkan distribusi data medis nyata dan menjadi pertimbangan penting dalam pemilihan metrik evaluasi yang digunakan pada penelitian ini.

B. Tahap Preprocessing Data

Tahap *preprocessing* data merupakan langkah penting dalam penelitian ini untuk memastikan bahwa citra masukan memiliki kualitas dan format yang sesuai sebelum diproses oleh model *deep learning*. Preprocessing bertujuan untuk menyeragamkan ukuran citra, menormalkan nilai piksel, serta meningkatkan stabilitas dan konvergensi model selama proses pelatihan.

Pada penelitian ini, seluruh citra histopatologi kanker kandung kemih terlebih dahulu dilakukan penyaringan format file, sehingga hanya file citra dengan ekstensi valid seperti *.jpg*, *.jpeg*, *.png*, *.bmp*, dan *.tif* yang digunakan. Langkah ini dilakukan untuk menghindari kesalahan pemuatan data akibat adanya file non-gambar di dalam folder dataset.

Selanjutnya, seluruh citra diubah ukurannya (*resizing*) menjadi 224×224 piksel. Ukuran ini dipilih karena merupakan ukuran input standar yang digunakan oleh arsitektur MedViT dan Swin Transformer, sehingga memungkinkan pemanfaatan bobot pra-latih (*pretrained weights*) serta menjaga konsistensi input antar model. Proses *resizing* juga membantu menurunkan beban komputasi dan mempercepat proses pelatihan.

Setelah proses *resizing*, citra dikonversi ke dalam bentuk tensor dan dilakukan normalisasi nilai piksel ke rentang $[-1, 1]$. Normalisasi ini bertujuan untuk menstabilkan distribusi data input, sehingga proses optimisasi menggunakan *gradient descent* dapat berjalan lebih efektif dan mengurangi risiko ketidakstabilan numerik selama pelatihan model.

Proses preprocessing citra dalam penelitian ini diimplementasikan menggunakan pustaka `torchvision.transforms`, sebagaimana ditunjukkan pada potongan kode berikut:

```
transform = transforms.Compose([
    transforms.Resize((224, 224)),
    transforms.ToTensor(),
    transforms.Normalize(mean=[0.5, 0.5, 0.5],
                        std=[0.5, 0.5, 0.5])
])
```

C. Tahap Pembagian Dataset

Pembagian dataset merupakan tahapan penting dalam penelitian ini untuk memastikan bahwa proses pelatihan, validasi, dan evaluasi model dilakukan secara objektif serta mampu menggambarkan kemampuan generalisasi model terhadap data yang belum pernah dilihat sebelumnya. Oleh karena itu, dataset citra kanker kandung kemih dibagi ke dalam tiga subset utama, yaitu data latih (*training set*), data validasi (*validation set*), dan data uji (*test set*).

Dalam penelitian ini, dataset dibagi menjadi tiga subset dengan proporsi 70% data latih, 15% data validasi, dan 15% data uji. Pemilihan rasio ini tidak ditetapkan secara arbitrer, melainkan mengikuti praktik terbaik dalam penelitian klasifikasi citra medis, di mana alokasi terbesar diberikan pada data pelatihan agar model mampu mempelajari distribusi fitur secara optimal. Proporsi validasi dan uji yang relatif seimbang bertujuan memastikan proses tuning hyperparameter berlangsung objektif serta menghindari overfitting, sekaligus menyediakan subset independen untuk mengevaluasi generalisasi model setelah pelatihan selesai.

Pembagian dataset dilakukan secara acak namun terkontrol menggunakan fungsi `train_test_split` dari pustaka *scikit-learn* dengan parameter `random_state = 42`. Penggunaan nilai *random state* yang tetap bertujuan untuk menjaga reproduktibilitas eksperimen, sehingga hasil penelitian dapat diulang dan diverifikasi oleh peneliti lain. Selain itu, pembagian dilakukan secara terpisah untuk setiap kelas agar distribusi data antar kelas tetap terjaga dan tidak menimbulkan bias kelas.

Proses pembagian dataset dilakukan dalam dua tahap. Tahap pertama membagi dataset menjadi data latih (70%) dan data sementara (30%). Tahap kedua membagi data sementara tersebut menjadi data validasi (15%) dan data uji (15%). Potongan kode berikut menunjukkan implementasi pembagian dataset yang digunakan dalam penelitian ini:

```
# 70% data latih, 30% data sementara
train_images, temp_images = train_test_split(
    images, test_size=0.30, random_state=42, shuffle=True
)

# 30% data sementara dibagi menjadi 15% validasi dan 15% data uji
val_images, test_images = train_test_split(
    temp_images, test_size=0.50, random_state=42, shuffle=True
)
```

D. Klasifikasi Dengan Model Swin Transformer

1) Pengertian

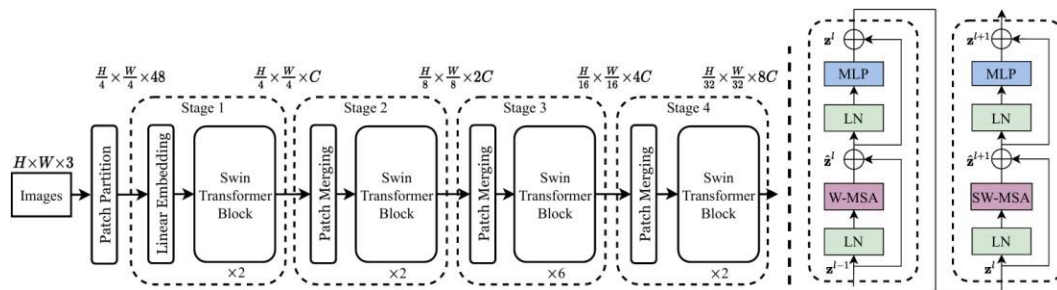
Swin Transformer merupakan arsitektur *vision transformer* hierarkis yang diperkenalkan oleh Liu et al. (2021) untuk mengatasi keterbatasan Transformer konvensional dalam pemrosesan citra beresolusi tinggi. Berbeda dengan Vision Transformer (ViT) yang menerapkan *global self-attention* dengan kompleksitas kuadratik, Swin Transformer menggunakan mekanisme *self-attention* berbasis jendela lokal (*window-based self-attention*) yang memiliki kompleksitas linear terhadap ukuran citra. Pendekatan ini memungkinkan Swin Transformer berfungsi sebagai *general-purpose backbone* untuk berbagai tugas visi komputer, termasuk klasifikasi citra, deteksi objek, dan segmentasi semantik [7].

Swin Transformer (Shifted Window Transformer) merupakan arsitektur *vision transformer* hierarkis yang diperkenalkan oleh Liu et al. (2021) untuk mengatasi keterbatasan Vision Transformer (ViT) dalam memproses citra beresolusi tinggi. Berbeda dengan ViT konvensional yang menggunakan *global self-attention* dan memiliki kompleksitas komputasi kuadratik terhadap ukuran citra, Swin Transformer menerapkan mekanisme *self-attention* berbasis jendela (*window-based self-attention*), sehingga lebih efisien dan skalabel [8].

Model ini dirancang khusus untuk tugas visi komputer seperti klasifikasi citra, deteksi objek, dan segmentasi citra medis, dengan kemampuan menangkap informasi lokal dan global secara bertahap melalui struktur hierarkis. Oleh karena itu, Swin Transformer sangat sesuai digunakan dalam klasifikasi citra kanker kandung kemih yang memiliki variasi tekstur dan pola jaringan yang kompleks [9].

2) Arsitektur

Arsitektur Swin Transformer terdiri dari beberapa *stage* yang tersusun secara hierarkis, sebagaimana ditunjukkan pada Gambar arsitektur Swin Transformer. Setiap *stage* terdiri dari blok Swin Transformer yang mengombinasikan *Window Multi-Head Self-Attention* (W-MSA) dan *Shifted Window Multi-Head Self-Attention* (SW-MSA).



Sumber : Huang J et al, 2022 [9]

Gambar 2 Arsitektur Swin Transformer

Secara umum, alur arsitektur Swin Transformer adalah sebagai berikut:

- **Patch Partition**
Citra input berukuran $H \times W \times 3$ dibagi menjadi patch kecil berukuran tetap (misalnya 4×4 piksel). Setiap patch kemudian diratakan menjadi vektor.
- **Linear Embedding**
Patch yang telah dibagi diproyeksikan ke dimensi fitur tertentu menggunakan lapisan *linear embedding*.
- **Swin Transformer Block**
Setiap blok Swin Transformer terdiri dari:
 - Layer Normalization (LN)
 - Window-based Multi-Head Self-Attention (W-MSA) atau Shifted Window MSA (SW-MSA)
 - Multilayer Perceptron (MLP)
 - Residual Connection
- **Patch Merging**
Pada setiap transisi antar *stage*, dilakukan proses *patch merging* untuk mengurangi resolusi spasial dan meningkatkan dimensi fitur, sehingga membentuk representasi hierarkis seperti pada Convolutional Neural Network (CNN).

3) Cara Kerja

Cara kerja Swin Transformer diawali dengan pemrosesan citra input melalui *patch partition* dan *linear embedding*. Selanjutnya, fitur diekstraksi menggunakan mekanisme *window-based self-attention* (W-MSA), di mana perhatian hanya dihitung dalam jendela lokal tertentu untuk menekan kompleksitas komputasi.

Untuk memungkinkan pertukaran informasi antar jendela, Swin Transformer menggunakan *shifted window mechanism* (SW-MSA), yaitu dengan menggeser posisi jendela pada layer berikutnya. Mekanisme ini memungkinkan model menangkap hubungan global antar patch tanpa harus menghitung *self-attention* secara penuh pada seluruh citra.

Setelah melalui beberapa *stage* Swin Transformer dan proses *patch merging*, fitur akhir diratakan dan diteruskan ke lapisan klasifikasi (*fully connected layer*) untuk menghasilkan probabilitas kelas kanker kandung kemih, yaitu HGC, LGC, NST, dan NTL.

4) Kelebihan

Swin Transformer memiliki beberapa keunggulan utama, antara lain [10]:

- Efisiensi komputasi karena *self-attention* dihitung secara lokal pada jendela tertentu.
- Struktur hierarkis yang menyerupai CNN, sehingga cocok untuk citra medis beresolusi tinggi.
- Kemampuan menangkap konteks lokal dan global melalui mekanisme *shifted window*.
- Performa tinggi pada berbagai tugas visi komputer, termasuk klasifikasi citra medis.
- Mudah diintegrasikan dengan model pralatih (pretrained) melalui pustaka seperti timm.

5) Kekurangan

Di samping keunggulannya, Swin Transformer juga memiliki beberapa keterbatasan, antara lain:

- Kompleksitas arsitektur lebih tinggi dibandingkan CNN konvensional.
- Kebutuhan sumber daya komputasi yang besar, terutama pada proses pelatihan.
- Sensitif terhadap ukuran patch dan jendela, sehingga memerlukan penyesuaian hiperparameter yang cermat.
- Kurang optimal pada dataset berukuran sangat kecil tanpa teknik regularisasi atau augmentasi yang memadai.

E. Klasifikasi Dengan Model MedViT

1) Pengertian

Medical Vision Transformer (MedViT) merupakan arsitektur *deep learning* berbasis Transformer yang dirancang khusus untuk menangani karakteristik citra medis. Berbeda dengan Vision Transformer standar yang hanya mengandalkan mekanisme *self-attention* global, MedViT mengombinasikan kekuatan Convolutional Neural Network (CNN) dan Transformer dalam satu arsitektur terpadu. Pendekatan ini bertujuan untuk menangkap fitur lokal beresolusi tinggi yang penting dalam citra medis, sekaligus memodelkan hubungan global antar fitur secara efektif [11].

MedViT dikembangkan untuk mengatasi keterbatasan CNN konvensional yang cenderung fokus pada fitur lokal, serta kekurangan Transformer murni yang memerlukan data besar dan kurang optimal dalam menangkap detail spasial halus. Oleh karena itu, MedViT sangat sesuai untuk tugas klasifikasi citra medis, termasuk klasifikasi kanker kandung kemih berbasis citra histopatologi atau mikroskopik.

2) Arsitektur

Arsitektur MedViT (Medical Vision Transformer) dirancang untuk menggabungkan keunggulan Convolutional Neural Network (CNN) dalam menangkap fitur lokal dan Vision Transformer (ViT) dalam memodelkan hubungan global. Berbeda dengan ViT konvensional, MedViT menggunakan pendekatan hierarchical multi-stage serta dua blok utama, yaitu Local Feature Perception (LFP) dan Global Feature Perception (GFP), yang dioptimalkan untuk karakteristik citra medis.

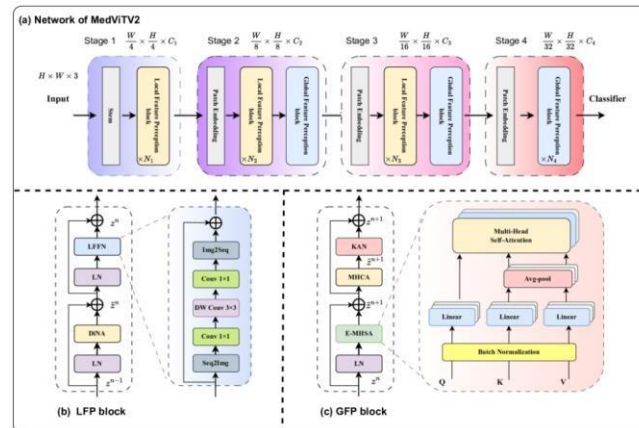
Pengembangan MedViT didorong oleh temuan bahwa MedViT versi awal (MedViTV1), meskipun menunjukkan performa yang baik, mengalami penurunan akurasi ketika jumlah parameter meningkat. Hal ini menunjukkan keterbatasan dalam skalabilitas model. Oleh karena itu, MedViT V2 diperkenalkan dengan memasukkan komponen baru yang lebih efisien dan ekspresif, yaitu Dilated Neighborhood Attention (DiNA) dan Kolmogorov–Arnold Networks (KAN), untuk meningkatkan performa pada dataset medis berskala kecil hingga menengah.

Selain mengatasi masalah skalabilitas, arsitektur MedViT dirancang dengan struktur hierarkis empat tahap (*stage*) yang memungkinkan ekstraksi fitur secara bertingkat. Pada setiap *stage*, resolusi spasial fitur dikurangi sementara jumlah kanal meningkat, sehingga model mampu merepresentasikan pola visual dari tingkat rendah (tekstur dan tepi) hingga pola semantik tingkat tinggi yang lebih kompleks. Setiap *stage* terdiri dari kombinasi LFP block dan GFP block, yang bekerja secara berurutan untuk mengekstraksi dan menggabungkan fitur lokal maupun global secara efektif.

LFP block menggunakan operasi konvolusi (Conv 1×1 dan Depthwise Conv 3×3) untuk mengekstraksi fitur lokal dan mempertahankan bias spasial yang penting dalam citra medis, khususnya untuk mendeteksi pola morfologis kanker. Sementara itu, GFP block memanfaatkan mekanisme attention berbasis Multi-Head Channel Attention (MHCA), Efficient Multi-Head Self-Attention (E-MHSA), serta KAN layer untuk memodelkan dependensi global antar patch secara lebih efisien dibanding *self-attention* standar. Integrasi antara LFP dan GFP menghasilkan representasi fitur yang kaya dan mampu menangkap variasi struktur kanker kandung kemih secara lebih akurat.

Selanjutnya, proses *embedding patch* dan *attention* dalam MedViT juga dikombinasikan dengan *batch normalization* dan *residual connection* untuk memastikan stabilitas training, mencegah gradien menghilang, serta mempercepat konvergensi.

Output akhir dari tahap hierarkis diteruskan ke classification head guna menghasilkan prediksi kelas kanker kandung kemih. Dengan desain modular yang efisien, MedViT mampu menghasilkan performa tinggi meskipun menggunakan dataset medis berskala relatif kecil, sehingga menjadikan model ini lebih sesuai diterapkan pada permasalahan klinis nyata dibanding arsitektur transformer konvensional.



Sumber : Karim R et al, 2023 [12]

Gambar 3 Arsitektur MedViT

- Struktur Umum MedViT**
 Seperti ditunjukkan pada Gambar arsitektur, MedViT terdiri dari empat stage utama, yaitu Stage 1 hingga Stage 4, dengan resolusi fitur yang semakin mengecil dan jumlah kanal yang meningkat secara bertahap:
 - Stage 1: Resolusi tinggi, fokus pada ekstraksi fitur lokal awal
 - Stage 2–3: Kombinasi fitur lokal dan global
 - Stage 4: Representasi fitur global tingkat tinggi sebelum klasifikasi
 Setiap stage diawali dengan patch embedding, yang berfungsi mengubah citra atau feature map menjadi representasi token untuk diproses oleh blok transformer.
- Local Feature Perception (LFP) Block**
 LFP block bertujuan untuk menangkap detail lokal, seperti tekstur, tepi sel, dan pola mikroskopis yang sangat penting pada citra medis.
 Komponen utama LFP block:
 - Depthwise Convolution (DW Conv 3×3): Menangkap informasi spasial lokal secara efisien
 - Conv 1×1: Menyesuaikan dimensi kanal
 - Img2Seq dan Seq2Img: Mengonversi antara representasi spasial dan sekuensial
 - DINA (Dynamic Inverted Attention): Memperkuat fitur penting lokal
 - Layer Normalization (LN) dan Local Feed Forward Network (LFN)
 LFP block memungkinkan MedViT mempertahankan keunggulan CNN dalam mengekstraksi fitur lokal yang kaya, yang sering kali hilang pada Vision Transformer murni.
- Global Feature Perception (GFP) Block**
 GFP block berfungsi untuk memodelkan hubungan global antar region citra, yang penting untuk memahami struktur makroskopis jaringan.
 Komponen utama GFP block:
 - E-MHSA (Efficient Multi-Head Self-Attention): Versi efisien dari self-attention untuk mengurangi kompleksitas komputasi
 - MHCA (Multi-Head Convolutional Attention): Menggabungkan perhatian berbasis konvolusi
 - KAN (Kernel Attention Network): Memperkuat representasi global
 - Batch Normalization dan Linear Projection pada proses pembentukan Query, Key, dan Value
 GFP block memungkinkan MedViT menangkap konteks global tanpa biaya komputasi sebesar ViT standar.
- Hierarchical Patch Embedding**
 Setiap stage menggunakan patch embedding bertingkat, di mana resolusi spasial berkurang ($H/4 \rightarrow H/8 \rightarrow H/16 \rightarrow H/32$) dan jumlah kanal meningkat ($C_1 \rightarrow C_4$). Pendekatan ini menyerupai desain CNN modern seperti ResNet, namun tetap mempertahankan keunggulan transformer.
- Classifier**
 Pada akhir Stage 4, fitur global hasil ekstraksi MedViT diteruskan ke classifier, yang terdiri dari lapisan linear untuk menghasilkan prediksi kelas. Pada penelitian ini, classifier digunakan untuk mengklasifikasikan citra ke dalam empat kelas kanker kandung kemih: HGC, LGC, NST, dan NTL.

3) Cara Kerja

Secara umum, alur kerja MedViT dapat dijelaskan sebagai berikut:

- **Patch Embedding**
Citra input berukuran $H \times W \times 3$ terlebih dahulu diproses melalui tahap *patch embedding*, di mana citra dibagi menjadi beberapa patch dan diproyeksikan ke dalam ruang vektor berdimensi tertentu. Proses ini mengubah representasi citra dua dimensi menjadi bentuk sekuensial yang sesuai untuk diproses oleh arsitektur transformer.
- **Local Feature Processing (LFP Block)**
Setelah proses embedding, fitur diproses menggunakan *Local Feature Processing (LFP) block*. LFP memanfaatkan operasi konvolusi seperti *depthwise convolution* dan *pointwise convolution* untuk mengekstraksi informasi lokal, termasuk tepi, tekstur, dan pola morfologis halus yang sangat penting dalam citra medis. Tahap ini bertujuan mempertahankan bias lokal yang umumnya hilang pada transformer konvensional.
- **Global Feature Processing (GFP Block)**
Fitur hasil LFP selanjutnya diteruskan ke *Global Feature Processing (GFP) block*, yang menggunakan mekanisme *self-attention* untuk memodelkan hubungan global antar patch. Pada tahap ini, MedViT mampu menangkap konteks global citra, seperti hubungan antar area jaringan dan pola abnormal yang tersebar luas. GFP juga mengintegrasikan komponen seperti *efficient multi-head self-attention* dan *Kolmogorov–Arnold Networks (KAN)* untuk meningkatkan efisiensi dan daya representasi model.
- **Hierarchical Multi-Stage Processing**
Proses LFP dan GFP dilakukan secara bertingkat pada beberapa stage. Pada setiap stage, resolusi spasial fitur secara bertahap diperkecil melalui patch merging, sementara dimensi kanal ditingkatkan. Pendekatan hierarkis ini memungkinkan MedViT mengekstraksi representasi fitur dari tingkat rendah hingga tingkat tinggi secara progresif.
- **Klasifikasi Akhir**
Fitur keluaran dari stage terakhir dirangkum dan diteruskan ke *classification head* untuk menghasilkan prediksi kelas. Dalam penelitian ini, tahap ini menghasilkan klasifikasi subtype kanker kandung kemih, yaitu HGC, LGC, NST, dan NTL.

4) Kelebihan

Model MedViT memiliki sejumlah keunggulan yang menjadikannya sangat sesuai untuk aplikasi klasifikasi citra medis, khususnya pada dataset dengan ukuran terbatas dan struktur visual yang kompleks.

- MedViT mampu mengintegrasikan fitur lokal dan global secara efektif. Kombinasi LFP block dan GFP block memungkinkan model menangkap detail tekstur halus sekaligus hubungan spasial global, yang sangat penting dalam membedakan subtype penyakit pada citra medis.
- MedViT dirancang dengan efisiensi komputasi yang tinggi. Penggunaan mekanisme perhatian lokal seperti DiNA serta arsitektur hierarkis membuat kompleksitas waktu dan memori lebih rendah dibandingkan Vision Transformer konvensional, sehingga model lebih mudah dilatih pada perangkat dengan sumber daya terbatas.
- MedViT menunjukkan ketahanan terhadap ukuran dataset yang kecil, yang merupakan karakteristik umum pada data medis. Integrasi komponen seperti *Kolmogorov–Arnold Networks (KAN)* sebagai alternatif MLP membantu mengurangi jumlah parameter dan meningkatkan kemampuan generalisasi model.

5) Kekurangan

Meskipun memiliki berbagai keunggulan, MedViT juga memiliki beberapa keterbatasan sebagai berikut:

- Kombinasi LFP dan GFP dalam beberapa stage membuat arsitektur MedViT lebih kompleks dibanding CNN konvensional, sehingga membutuhkan pemahaman arsitektur yang lebih mendalam.
- Meskipun lebih efisien dibanding ViT murni, MedViT tetap memerlukan sumber daya komputasi yang relatif besar dibanding model CNN ringan, terutama saat digunakan pada resolusi citra tinggi.
- Performa MedViT sangat dipengaruhi oleh jumlah stage, ukuran patch, dan dimensi embedding, sehingga diperlukan proses *hyperparameter tuning* yang cermat.
- MedViT dirancang khusus untuk citra medis, sehingga performanya belum tentu optimal jika diterapkan langsung pada dataset citra umum tanpa penyesuaian arsitektur.

F. Evaluasi Model

Evaluasi model pada penelitian ini dilakukan untuk menilai kemampuan MedViT dan Swin Transformer dalam mengklasifikasikan citra kanker kandung kemih ke dalam empat kelas, yaitu High-Grade Carcinoma (HGC), Low-Grade Carcinoma (LGC), Non-Specific Tumor (NST), dan Non-Tumor Lesion (NTL). Proses evaluasi dirancang secara komprehensif dengan memanfaatkan beberapa metrik evaluasi yang umum digunakan dalam klasifikasi citra medis, sehingga hasil yang diperoleh mampu merefleksikan performa model secara menyeluruh dan objektif.

- **Loss of function**

Loss function bertugas membandingkan nilai prediksi model dengan nilai sebenarnya (*ground truth*) secara kontinu selama proses pelatihan. Hasil perbandingan tersebut digunakan untuk memperbarui bobot model guna meminimalkan kesalahan yang dihasilkan. Dalam penelitian ini, digunakan *cross-entropy loss* sebagai fungsi *loss*. Secara formal, fungsi ini dapat

didefinisikan sebagai berikut:

$$L = \frac{1}{N} \sum_i L_i = -\frac{1}{N} \sum_i \sum_{c=1}^m y_{ic} \log(p_{ic}) \quad (1)$$

dengan N menyatakan jumlah sampel, m jumlah kelas, y_{ic} merupakan label biner dari kelas sebenarnya (0 atau 1), dan p_{ic} adalah probabilitas prediksi bahwa sampel ke- i termasuk ke dalam kelas ke- c . Pemantauan nilai *loss* pada data latih dan data validasi juga digunakan untuk mendeteksi potensi *overfitting* selama proses pelatihan.

- Akurasi

Akurasi digunakan untuk mengukur persentase prediksi yang benar dibandingkan dengan total jumlah data yang diuji. Metrik ini memberikan gambaran umum mengenai kemampuan model dalam mengklasifikasikan citra kanker kandung kemih secara keseluruhan. Meskipun akurasi bersifat intuitif, pada dataset medis dengan distribusi kelas yang tidak sepenuhnya seimbang, akurasi perlu dikombinasikan dengan metrik lain agar evaluasi lebih representatif.

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \quad (2)$$

di mana:

TP (True Positive) adalah jumlah data yang diprediksi benar sebagai kelas positif,

TN (True Negative) adalah jumlah data yang diprediksi benar sebagai kelas negatif,

FP (False Positive) adalah jumlah data yang diprediksi positif tetapi sebenarnya negatif,

FN (False Negative) adalah jumlah data yang diprediksi negatif tetapi sebenarnya positif.

- Confusion Matrix

Confusion matrix merupakan salah satu metode evaluasi yang digunakan untuk menganalisis performa model klasifikasi secara lebih rinci. Berbeda dengan akurasi yang hanya memberikan gambaran umum, confusion matrix menyajikan informasi detail mengenai jumlah prediksi yang benar dan salah pada setiap kelas. Dengan demikian, confusion matrix sangat penting untuk memahami pola kesalahan model, terutama pada permasalahan klasifikasi citra medis yang melibatkan beberapa kelas.

Dalam penelitian ini, confusion matrix digunakan untuk mengevaluasi kemampuan model MedViT dan Swin Transformer dalam mengklasifikasikan citra kanker kandung kemih ke dalam empat kelas, yaitu High-Grade Carcinoma (HGC), Low-Grade Carcinoma (LGC), Non-Specific Tumor (NST), dan Non-Tumor Lesion (NTL). Matriks ini disusun dalam bentuk tabel dua dimensi, di mana baris merepresentasikan kelas sebenarnya (*actual class*), sedangkan kolom merepresentasikan kelas hasil prediksi (*predicted class*).

Setiap elemen pada confusion matrix menggambarkan jumlah sampel yang termasuk dalam kategori tertentu, yaitu:

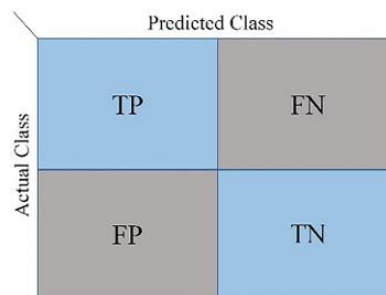
TP (True Positive) adalah jumlah data yang diprediksi benar sebagai kelas positif,

TN (True Negative) adalah jumlah data yang diprediksi benar sebagai kelas negatif,

FP (False Positive) adalah jumlah data yang diprediksi positif tetapi sebenarnya negatif,

FN (False Negative) adalah jumlah data yang diprediksi negatif tetapi sebenarnya positif.

Melalui confusion matrix, dapat dianalisis kelas mana yang paling sering mengalami kesalahan klasifikasi serta sejauh mana model mampu membedakan subtype kanker kandung kemih yang memiliki karakteristik morfologis yang mirip. Analisis ini menjadi sangat penting dalam konteks medis, karena kesalahan klasifikasi pada kelas tertentu dapat berdampak langsung pada pengambilan keputusan klinis.



Sumber : Liyanage V, 2023 [13]

Gambar 4 Confusion Matrix

- Precision, Recall, dan F1-Score

Untuk mengevaluasi performa model secara lebih mendalam, terutama pada dataset dengan distribusi kelas yang tidak seimbang, digunakan metrik precision, recall, dan F1-score.

- Precision mengukur tingkat ketepatan prediksi positif yang dilakukan oleh model dan didefinisikan sebagai:

$$Precision = \frac{TP}{TP + FP} \quad (3)$$

- Recall mengukur kemampuan model dalam mendeteksi seluruh data positif yang sebenarnya dan didefinisikan sebagai:

$$Recall = \frac{TP}{TP + FN} \quad (4)$$

- F1-score merupakan rata-rata harmonik antara precision dan recall, yang memberikan keseimbangan antara keduanya:

$$F1 - Score = 2 \times \frac{Precision \times Recall}{Precision + Recall} \quad (5)$$

III. HASIL DAN PEMBAHASAN

Bab ini menyajikan hasil eksperimen dan pembahasan terkait kinerja model MedViT dan Swin Transformer dalam melakukan klasifikasi citra kanker kandung kemih. Evaluasi dilakukan menggunakan dataset yang telah dibagi ke dalam data latih, validasi, dan uji dengan skema 70/15/15. Hasil yang diperoleh dianalisis secara komprehensif menggunakan berbagai metrik evaluasi, meliputi nilai *loss*, akurasi, *confusion matrix*, *precision*, *recall*, *F1-score*, serta kurva ROC-AUC untuk menilai kemampuan klasifikasi dan generalisasi masing-masing model.

Pembahasan pada bab ini difokuskan pada analisis performa model selama proses pelatihan dan validasi, perbandingan hasil evaluasi pada data uji, serta interpretasi kesalahan klasifikasi yang terjadi pada setiap kelas kanker kandung kemih. Selain itu, dilakukan pula perbandingan langsung antara MedViT dan Swin Transformer untuk mengidentifikasi keunggulan dan keterbatasan masing-masing arsitektur dalam menangani karakteristik citra medis yang kompleks. Analisis ini diharapkan dapat memberikan pemahaman yang lebih mendalam mengenai efektivitas penerapan Vision Transformer pada klasifikasi kanker kandung kemih.

A. Hasil Pelatihan Model MedViT

Pada tahap ini, model MedViT dilatih untuk mengklasifikasikan citra kanker kandung kemih ke dalam empat kelas (HGC, LGC, NST, NTL) menggunakan pembagian data train/validation/test = 70/15/15. Berdasarkan hasil split dataset, jumlah data yang digunakan adalah: Train = 1225, Validation = 263, dan Test = 266 citra, dengan rincian per kelas: HGC (328/70/71), LGC (452/97/98), NST (352/76/76), dan NTL (93/20/21) untuk Train/Val/Test.

Epoch 1/15	Train Loss: 0.8020	Train Acc: 0.72	Val Loss: 0.4929	Val Acc: 0.84	Time: 25.9s
Epoch 2/15	Train Loss: 0.3336	Train Acc: 0.88	Val Loss: 0.3559	Val Acc: 0.89	Time: 25.2s
Epoch 3/15	Train Loss: 0.1743	Train Acc: 0.95	Val Loss: 0.3528	Val Acc: 0.90	Time: 24.5s
Epoch 4/15	Train Loss: 0.1136	Train Acc: 0.97	Val Loss: 0.2366	Val Acc: 0.95	Time: 24.3s
Epoch 5/15	Train Loss: 0.0679	Train Acc: 0.98	Val Loss: 0.2842	Val Acc: 0.95	Time: 24.8s
Epoch 6/15	Train Loss: 0.0577	Train Acc: 0.99	Val Loss: 0.2626	Val Acc: 0.95	Time: 24.3s
Epoch 7/15	Train Loss: 0.0343	Train Acc: 0.99	Val Loss: 0.2466	Val Acc: 0.96	Time: 23.7s
Epoch 8/15	Train Loss: 0.0272	Train Acc: 1.00	Val Loss: 0.2765	Val Acc: 0.96	Time: 23.6s
Epoch 9/15	Train Loss: 0.0263	Train Acc: 0.99	Val Loss: 0.2596	Val Acc: 0.96	Time: 23.6s
Epoch 10/15	Train Loss: 0.0117	Train Acc: 1.00	Val Loss: 0.2445	Val Acc: 0.97	Time: 23.8s
Epoch 11/15	Train Loss: 0.0265	Train Acc: 0.99	Val Loss: 0.3370	Val Acc: 0.94	Time: 23.3s
Epoch 12/15	Train Loss: 0.0134	Train Acc: 1.00	Val Loss: 0.2817	Val Acc: 0.96	Time: 23.9s
Epoch 13/15	Train Loss: 0.0122	Train Acc: 1.00	Val Loss: 0.2280	Val Acc: 0.97	Time: 24.0s
Epoch 14/15	Train Loss: 0.0060	Train Acc: 1.00	Val Loss: 0.2616	Val Acc: 0.97	Time: 23.8s
Epoch 15/15	Train Loss: 0.0051	Train Acc: 1.00	Val Loss: 0.3308	Val Acc: 0.97	Time: 24.1s

Total training time: 6.05 minutes

Gambar 5 Training Model MedViT

Berdasarkan Gambar 5, dapat diamati bahwa proses pelatihan model MedViT menunjukkan konvergensi yang cepat dan stabil selama 15 epoch. Pada epoch awal, nilai *training loss* masih relatif tinggi sebesar 0,8020 dengan akurasi pelatihan 0,72, sementara akurasi validasi berada pada 0,84. Hal ini menunjukkan bahwa model mulai mempelajari pola dasar dari data citra kanker kandung kemih.

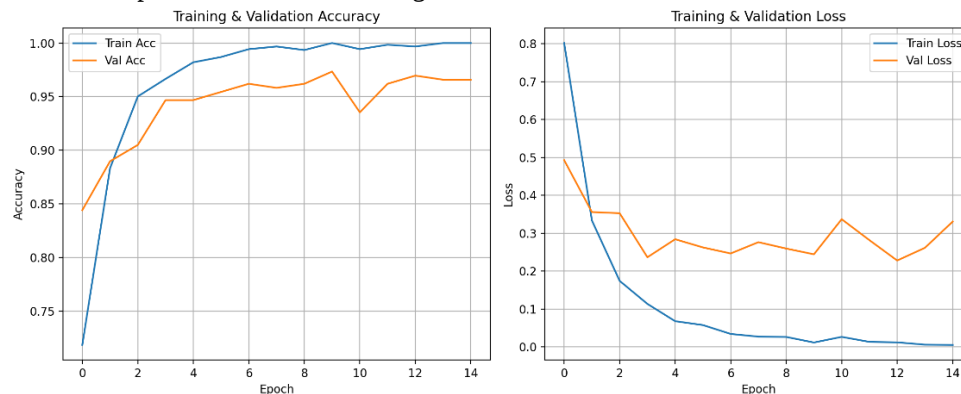
Seiring bertambahnya jumlah epoch, nilai *training loss* mengalami penurunan yang signifikan dan konsisten, hingga mencapai 0,0051 pada epoch ke-15. Penurunan loss ini diikuti dengan peningkatan *training accuracy* yang mencapai 100% mulai sekitar epoch ke-8 dan tetap stabil hingga akhir pelatihan. Kondisi ini mengindikasikan bahwa model MedViT mampu mempelajari representasi fitur yang sangat baik pada data latih.

Pada sisi validasi, *validation accuracy* juga menunjukkan peningkatan yang stabil dari 0,84 pada epoch pertama menjadi 0,97 pada epoch ke-15. Meskipun terdapat fluktuasi kecil pada nilai *validation loss* di beberapa epoch, nilai akurasi validasi tetap berada pada kisaran tinggi (94%–97%), yang menandakan bahwa model memiliki kemampuan generalisasi yang baik terhadap data yang tidak digunakan dalam pelatihan.

Perbedaan antara *training accuracy* dan *validation accuracy* relatif kecil, serta tidak terdapat peningkatan *validation loss* yang signifikan secara berkelanjutan. Hal ini menunjukkan bahwa model MedViT tidak mengalami overfitting yang serius, meskipun akurasi pelatihan mencapai nilai maksimum. Dengan kata lain, model tetap mampu mempertahankan performa yang konsisten pada data validasi. Selain itu, waktu pelatihan per epoch berada pada kisaran 23–26 detik, dengan total waktu pelatihan sekitar 6,05 menit. Hal ini menunjukkan bahwa arsitektur MedViT memiliki efisiensi komputasi yang baik dan dapat dilatih secara efektif pada dataset medis berskala menengah.

B. Evaluasi Model MedViT pada Data Uji

Evaluasi model MedViT pada data uji dilakukan untuk mengukur kemampuan generalisasi model dalam mengklasifikasikan citra kanker kandung kemih yang tidak pernah digunakan selama proses pelatihan maupun validasi. Evaluasi ini menjadi tahap krusial untuk memastikan bahwa performa tinggi yang diperoleh pada data latih dan validasi tidak disebabkan oleh *overfitting*, melainkan mencerminkan kemampuan model dalam menangani data baru.

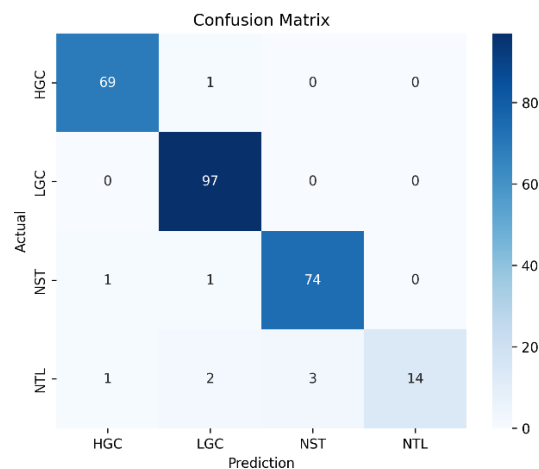


Gambar 6 kurva training dan validation accuracy serta training dan validation loss

Berdasarkan Gambar 6 yang menampilkan kurva *training & validation accuracy* serta *training & validation loss*, dapat dianalisis bahwa model MedViT menunjukkan kinerja pelatihan yang sangat baik dan stabil. Pada grafik akurasi, terlihat bahwa akurasi data latih meningkat secara signifikan sejak epoch awal, dari sekitar 72% pada epoch pertama hingga mencapai 100% pada epoch ke-8 dan seterusnya. Hal ini menunjukkan bahwa model mampu mempelajari pola fitur citra kanker kandung kemih secara efektif.

Akurasi validasi juga menunjukkan tren peningkatan yang konsisten, dimulai dari sekitar 84% pada epoch pertama dan mencapai kisaran 95–97% pada epoch akhir. Meskipun terdapat sedikit fluktuasi pada beberapa epoch (misalnya penurunan pada epoch ke-10), akurasi validasi kembali meningkat dan stabil. Kondisi ini mengindikasikan bahwa model memiliki kemampuan generalisasi yang baik terhadap data yang tidak digunakan selama proses pelatihan.

Dari sisi *loss function*, grafik menunjukkan bahwa training loss menurun secara drastis dari sekitar 0,80 pada epoch pertama hingga mendekati 0,00 pada epoch terakhir, yang menandakan proses optimasi berjalan dengan sangat efektif. Sementara itu, *validation loss* juga mengalami penurunan secara umum dan berada pada rentang 0,23–0,33 di akhir pelatihan. Walaupun nilai *validation loss* tidak menurun sedrastis *training loss*, pola yang relatif stabil menunjukkan bahwa model tidak mengalami *underfitting*.



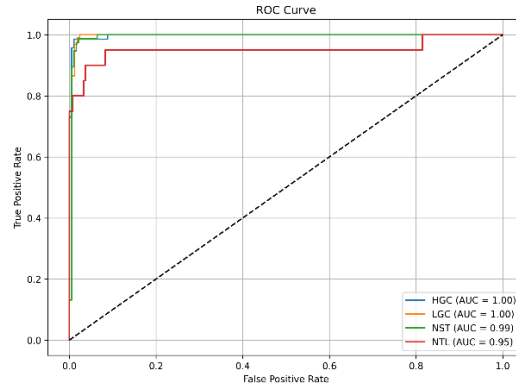
Gambar 7 Confusion Matrix Model MedViT

Berdasarkan confusion matrix yang ditunjukkan pada Gambar 7, dapat dianalisis bahwa model MedViT menunjukkan kinerja klasifikasi yang sangat baik pada sebagian besar kelas kanker kandung kemih. Pada kelas High-Grade Carcinoma (HGC), dari total 70 data uji, sebanyak 69 citra berhasil diklasifikasikan dengan benar, sementara 1 citra salah diklasifikasikan sebagai Low-Grade Carcinoma (LGC). Hal ini menunjukkan bahwa model memiliki kemampuan yang sangat baik dalam mengenali karakteristik HGC, dengan tingkat kesalahan yang sangat rendah.

Untuk kelas Low-Grade Carcinoma (LGC), seluruh 97 citra uji berhasil diprediksi dengan benar tanpa adanya kesalahan klasifikasi ke kelas lain. Hasil ini mengindikasikan bahwa fitur-fitur yang dipelajari oleh MedViT sangat representatif dalam membedakan LGC dari kelas lainnya, baik tumor tingkat tinggi maupun non-tumor.

Pada kelas Non-Specific Tumor (NST), dari total 76 citra uji, sebanyak 74 citra diklasifikasikan dengan benar, sementara 2 citra mengalami kesalahan klasifikasi, masing-masing diprediksi sebagai HGC dan LGC. Kesalahan ini menunjukkan adanya kemiripan visual atau pola morfologis tertentu antara NST dan kelas tumor lainnya, yang masih menjadi tantangan bagi model.

Sementara itu, pada kelas Non-Tumor Lesion (NTL), model berhasil mengklasifikasikan 14 dari 20 citra dengan benar, sedangkan 6 citra salah diklasifikasikan ke dalam kelas HGC, LGC, dan NST. Hal ini menunjukkan bahwa kelas NTL merupakan kelas yang paling menantang bagi model, kemungkinan disebabkan oleh variasi visual yang tinggi atau kemiripan fitur dengan jaringan tumor, sehingga memerlukan perhatian khusus dalam proses pelatihan atau penambahan data.



Gambar 8 ROC AUC Model MedViT

Berdasarkan Gambar 8, kurva Receiver Operating Characteristic (ROC) digunakan untuk mengevaluasi kemampuan model MedViT dalam membedakan masing-masing kelas kanker kandung kemih terhadap kelas lainnya. Kurva ROC memvisualisasikan hubungan antara True Positive Rate (TPR) dan False Positive Rate (FPR) pada berbagai nilai ambang keputusan (threshold). Semakin dekat kurva ROC ke sudut kiri atas, semakin baik performa model dalam melakukan klasifikasi.

Hasil yang diperoleh menunjukkan bahwa model MedViT memiliki performa klasifikasi yang sangat baik untuk seluruh kelas. Kelas High-Grade Carcinoma (HGC) dan Low-Grade Carcinoma (LGC) masing-masing mencapai nilai AUC = 1.00, yang menandakan kemampuan pemisahan kelas yang sempurna. Hal ini menunjukkan bahwa model hampir tidak mengalami kesalahan dalam membedakan kedua kelas tersebut dari kelas lainnya.

Kelas Non-Specific Tumor (NST) juga menunjukkan performa yang sangat tinggi dengan nilai AUC = 0.99, yang mengindikasikan bahwa model mampu mengenali pola karakteristik NST secara konsisten dengan tingkat kesalahan yang sangat rendah. Sementara itu, kelas Non-Tumor Lesion (NTL) memiliki nilai AUC = 0.95, yang meskipun sedikit lebih rendah dibandingkan kelas lainnya, tetap menunjukkan kemampuan diskriminasi yang sangat baik.

Tabel 1 Classification Report Model MedViT

Class	Precision	Recall	F1-Score	Support
HGC	0.9848	0.9155	0.9489	71
LGC	0.9231	0.9796	0.9505	98
NST	0.9615	0.9868	0.9740	76
NTL	1.0000	0.8571	0.9231	21
Accuracy			0.9549	266
Macro avg	0.9674	0.9348	0.9491	266
Weighted avg	0.9566	0.9549	0.9546	266

Berdasarkan classification report pada Gambar 9, dapat dianalisis bahwa model MedViT menunjukkan kinerja yang sangat baik dalam mengklasifikasikan citra kanker kandung kemih ke dalam empat kelas, yaitu HGC, LGC, NST, dan NTL. Evaluasi dilakukan menggunakan metrik precision, recall, dan F1-score, yang memberikan gambaran performa model secara lebih rinci dibandingkan akurasi saja.

Untuk kelas High-Grade Carcinoma (HGC), model memperoleh precision sebesar 0,9848, recall 0,9155, dan F1-score 0,9489. Nilai precision yang tinggi menunjukkan bahwa sebagian besar prediksi HGC oleh model adalah benar, meskipun recall yang sedikit lebih rendah mengindikasikan masih terdapat sejumlah kecil data HGC yang salah diklasifikasikan ke kelas lain.

Pada kelas Low-Grade Carcinoma (LGC), model menunjukkan performa yang sangat baik dengan precision 0,9231, recall 0,9796, dan F1-score 0,9505. Nilai recall yang sangat tinggi menandakan bahwa hampir seluruh data LGC berhasil terdeteksi dengan benar oleh model, meskipun precision sedikit lebih rendah akibat adanya prediksi positif palsu dalam jumlah kecil.

Kelas Non-Specific Tumor (NST) memperoleh precision 0,9615, recall 0,9868, dan F1-score 0,9740, yang merupakan salah satu performa terbaik di antara seluruh kelas. Hal ini menunjukkan bahwa MedViT sangat efektif dalam mengenali pola citra NST, baik dari sisi ketepatan prediksi maupun kelengkapan deteksi.

Sementara itu, pada kelas Non-Tumor Lesion (NTL), model mencapai precision 1,0000, recall 0,8571, dan F1-score 0,9231. Precision sempurna menandakan bahwa semua prediksi NTL yang dihasilkan model benar, namun nilai recall yang lebih rendah menunjukkan bahwa masih terdapat beberapa citra NTL yang salah diklasifikasikan ke kelas lain, kemungkinan disebabkan oleh kemiripan visual antara lesi non-tumor dan tumor tertentu.

C. Hasil Pelatihan Model Swin Transformer

Pada bagian ini disajikan hasil pelatihan model Swin Transformer dalam melakukan klasifikasi citra kanker kandung kemih ke dalam empat kelas, yaitu High-Grade Carcinoma (HGC), Low-Grade Carcinoma (LGC), Non-Specific Tumor (NST), dan Non-Tumor Lesion (NTL). Model Swin Transformer dilatih menggunakan dataset yang sama dengan model MedViT, sehingga perbandingan kinerja kedua model dapat dilakukan secara adil dan objektif. Proses pelatihan dilakukan selama beberapa epoch dengan pemantauan nilai loss dan akurasi pada data latih serta data validasi untuk mengevaluasi proses konvergensi model. Hasil pelatihan ini mencakup analisis kurva akurasi dan loss, performa klasifikasi pada data uji, serta evaluasi menggunakan confusion matrix, ROC curve, dan metrik evaluasi lainnya guna menilai kemampuan generalisasi Swin Transformer dalam konteks klasifikasi citra medis.

Epoch 1/15	Train Loss: 0.6781	Train Acc: 0.76	Val Loss: 0.3096	Val Acc: 0.89	Time: 31.9s
Epoch 2/15	Train Loss: 0.1355	Train Acc: 0.96	Val Loss: 0.0929	Val Acc: 0.98	Time: 28.3s
Epoch 3/15	Train Loss: 0.0533	Train Acc: 0.99	Val Loss: 0.0867	Val Acc: 0.97	Time: 28.0s
Epoch 4/15	Train Loss: 0.0312	Train Acc: 1.00	Val Loss: 0.0399	Val Acc: 0.99	Time: 27.9s
Epoch 5/15	Train Loss: 0.0310	Train Acc: 0.99	Val Loss: 0.0629	Val Acc: 0.98	Time: 27.7s
Epoch 6/15	Train Loss: 0.0084	Train Acc: 1.00	Val Loss: 0.0630	Val Acc: 0.98	Time: 27.5s
Epoch 7/15	Train Loss: 0.0059	Train Acc: 1.00	Val Loss: 0.0515	Val Acc: 0.98	Time: 27.5s
Epoch 8/15	Train Loss: 0.0053	Train Acc: 1.00	Val Loss: 0.0608	Val Acc: 0.98	Time: 27.6s
Epoch 9/15	Train Loss: 0.0296	Train Acc: 0.99	Val Loss: 0.0961	Val Acc: 0.97	Time: 27.7s
Epoch 10/15	Train Loss: 0.0214	Train Acc: 0.99	Val Loss: 0.0802	Val Acc: 0.97	Time: 27.7s
Epoch 11/15	Train Loss: 0.0092	Train Acc: 1.00	Val Loss: 0.0710	Val Acc: 0.99	Time: 27.9s
Epoch 12/15	Train Loss: 0.0059	Train Acc: 1.00	Val Loss: 0.0832	Val Acc: 0.99	Time: 27.6s
Epoch 13/15	Train Loss: 0.0472	Train Acc: 0.99	Val Loss: 0.0551	Val Acc: 0.98	Time: 27.7s
Epoch 14/15	Train Loss: 0.0196	Train Acc: 1.00	Val Loss: 0.0191	Val Acc: 0.99	Time: 27.7s
Epoch 15/15	Train Loss: 0.0022	Train Acc: 1.00	Val Loss: 0.0226	Val Acc: 0.99	Time: 28.0s

Total training time: 7.01 minutes

Gambar 9 Training Model Swin Transformer

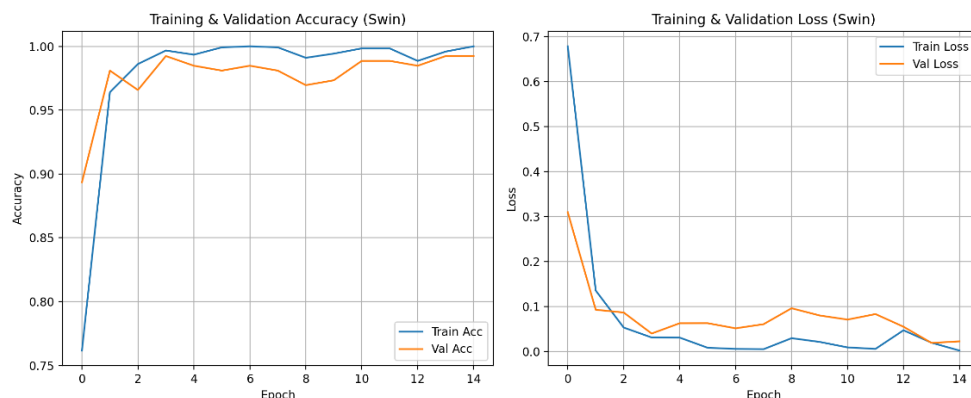
Berdasarkan Gambar 10, proses pelatihan model Swin Transformer menunjukkan performa yang sangat baik dan stabil sepanjang 15 epoch. Pada epoch pertama, model masih berada pada fase awal pembelajaran dengan nilai train accuracy sebesar 0,76 dan validation accuracy sebesar 0,89, serta nilai training loss 0,6781 dan validation loss 0,3096. Kondisi ini menunjukkan bahwa model mulai mampu menangkap pola dasar dari citra kanker kandung kemih.

Seiring bertambahnya jumlah epoch, terjadi peningkatan akurasi yang signifikan dan penurunan nilai loss yang konsisten. Pada epoch ke-4 hingga epoch ke-5, training accuracy telah mencapai 1,00, sementara validation accuracy berada pada kisaran 0,98–0,99, yang menandakan bahwa model telah mempelajari representasi fitur yang kuat dan relevan. Nilai training loss juga menurun drastis hingga mendekati nol, mencerminkan konvergensi model yang cepat dan efektif.

Selama seluruh proses pelatihan, validation accuracy tetap tinggi dan stabil, berkisar antara 0,97 hingga 0,99, tanpa adanya penurunan signifikan meskipun training accuracy telah mencapai nilai maksimum. Hal ini mengindikasikan bahwa model tidak mengalami overfitting yang berat, karena performa pada data validasi tetap sejalan dengan performa pada data latih.

Waktu pelatihan per epoch relatif konsisten, yaitu sekitar 27–28 detik, dengan total waktu pelatihan mencapai 7,01 menit. Konsistensi waktu ini menunjukkan efisiensi komputasi Swin Transformer meskipun menggunakan mekanisme self-attention hierarkis.

D. Evaluasi Model Swin Transformer pada Data Uji



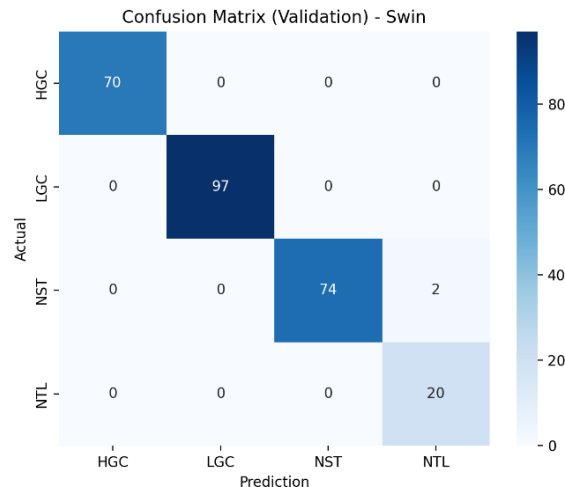
Gambar 10 kurva training dan validation accuracy serta training dan validation loss

Berdasarkan Gambar 11, ditunjukkan kurva akurasi dan loss pelatihan serta validasi pada model Swin Transformer selama 15 epoch. Pada grafik akurasi, terlihat bahwa akurasi pelatihan meningkat sangat cepat dari sekitar 0,76 pada epoch pertama hingga

mendekati 1,00 pada epoch ke-3, dan kemudian stabil hingga akhir pelatihan. Hal ini menunjukkan bahwa model Swin Transformer mampu mempelajari pola data pelatihan dengan sangat efektif dalam jumlah epoch yang relatif sedikit.

Akurasi validasi juga menunjukkan tren peningkatan yang konsisten, dimulai dari sekitar 0,89 pada epoch awal dan mencapai nilai di atas 0,98 pada sebagian besar epoch selanjutnya. Perbedaan antara kurva akurasi pelatihan dan validasi terlihat sangat kecil, yang mengindikasikan bahwa model memiliki kemampuan generalisasi yang baik dan tidak hanya menghafal data latih.

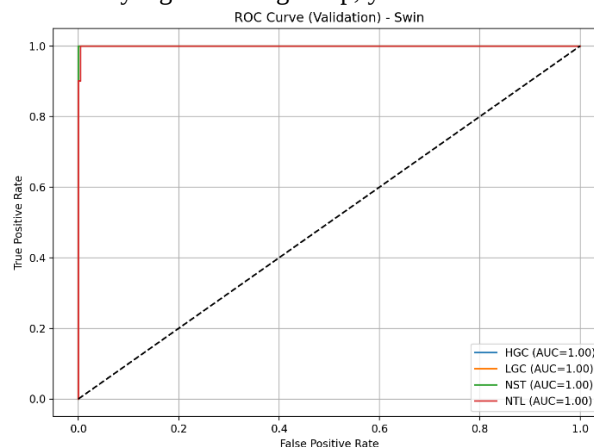
Pada grafik loss, terlihat bahwa training loss menurun secara signifikan dari nilai awal yang cukup tinggi hingga mendekati nol pada epoch-epoch akhir. Sementara itu, validation loss juga menunjukkan penurunan yang stabil, meskipun terdapat sedikit fluktuasi pada beberapa epoch. Fluktuasi kecil ini masih berada dalam batas wajar dan tidak menunjukkan peningkatan loss yang signifikan secara terus-menerus.



Gambar 11 Confusion Matrix Model Swin Transformer

Hasil **confusion matrix** pada **Gambar 12** menunjukkan bahwa model Swin Transformer mampu mengklasifikasikan sebagian besar citra kanker kantung kemih dengan **tingkat ketepatan yang sangat tinggi**. Seluruh data pada kelas **HGC** dan **LGC** berhasil diklasifikasikan secara sempurna tanpa kesalahan prediksi.

Pada kelas **NST**, terdapat **dua data** yang salah diklasifikasikan sebagai **NTL**, sementara mayoritas data NST tetap terklasifikasi dengan benar. Pada kelas **NTL**, seluruh data berhasil dikenali dengan tepat. Kesalahan yang terjadi relatif sangat kecil dan hanya muncul pada kelas dengan karakteristik visual yang cenderung mirip, yaitu antara **NST dan NTL**.



Gambar 12 ROC AUC Model Swin Transformer

Berdasarkan Gambar 13, kurva ROC (Receiver Operating Characteristic) menunjukkan bahwa model Swin Transformer memiliki kemampuan diskriminasi yang sangat tinggi dalam membedakan keempat kelas kanker kantung kemih, yaitu HGC, LGC, NST, dan NTL. Hal ini ditunjukkan oleh nilai Area Under Curve (AUC) sebesar 1.00 untuk seluruh kelas, yang menandakan performa klasifikasi yang mendekati sempurna.

Kurva ROC untuk masing-masing kelas berada jauh di atas garis diagonal (garis acak), yang berarti model memiliki True Positive Rate (TPR) yang sangat tinggi dengan False Positive Rate (FPR) yang sangat rendah. Kondisi ini menunjukkan bahwa Swin Transformer mampu mengidentifikasi citra dari setiap kelas dengan tingkat kesalahan yang minimal.

Nilai AUC sebesar 1.00 mengindikasikan bahwa model hampir tidak mengalami tumpang tindih prediksi antar kelas, sehingga sangat andal dalam konteks klasifikasi citra medis. Hal ini penting dalam aplikasi medis, karena kesalahan klasifikasi dapat berdampak langsung terhadap keputusan klinis.

Tabel 2 Classification Report Model Swin Transformer

Class	Precision	Recall	F1-Score	Support
HGC	1.0000	1.0000	1.0000	71
LGC	1.0000	0.9898	0.9949	98
NST	0.9967	0.9737	0.9801	76
NTL	0.9130	1.0000	0.9545	21
Accuracy			0.9887	266
Macro avg	0.9749	0.9909	0.9824	266
Weighted avg	0.9893	0.9887	0.9888	266

Berdasarkan Gambar 14, ditunjukkan hasil evaluasi kinerja model Swin Transformer pada data uji menggunakan metrik *precision*, *recall*, dan *F1-score* untuk masing-masing kelas kanker kandung kemih, yaitu HGC, LGC, NST, dan NTL.

Secara keseluruhan, model Swin Transformer menunjukkan kinerja yang sangat tinggi, dengan akurasi pengujian sebesar 98,87%. Nilai ini mengindikasikan bahwa sebagian besar citra uji berhasil diklasifikasikan dengan benar ke dalam kelas yang sesuai.

Untuk kelas High-Grade Carcinoma (HGC), model mencapai *precision*, *recall*, dan *F1-score* sebesar 1.000, yang berarti seluruh citra HGC pada data uji berhasil diklasifikasikan secara sempurna tanpa kesalahan. Hal ini menunjukkan kemampuan model yang sangat baik dalam mengenali karakteristik citra kanker derajat tinggi.

Pada kelas Low-Grade Carcinoma (LGC), model juga menunjukkan performa yang sangat baik dengan *precision* sebesar 1.000, *recall* sebesar 0.9898, dan *F1-score* sebesar 0.9949. Nilai *recall* yang sedikit lebih rendah mengindikasikan adanya sejumlah kecil citra LGC yang salah diklasifikasikan, namun secara keseluruhan performa tetap sangat optimal.

Untuk kelas Non-Specific Tumor (NST), diperoleh *precision* sebesar 0.9867, *recall* sebesar 0.9737, dan *F1-score* sebesar 0.9801. Hasil ini menunjukkan bahwa meskipun sebagian kecil citra NST mengalami kesalahan klasifikasi, model tetap mampu mengenali pola kelas NST dengan tingkat akurasi yang sangat tinggi.

Sementara itu, kelas Non-Tumor Lesion (NTL) memperoleh *precision* sebesar 0.9130, *recall* sebesar 1.000, dan *F1-score* sebesar 0.9545. Nilai *recall* yang sempurna menunjukkan bahwa seluruh citra NTL berhasil terdeteksi dengan benar, meskipun terdapat beberapa kesalahan prediksi dari kelas lain yang diklasifikasikan sebagai NTL, yang tercermin dari nilai *precision* yang sedikit lebih rendah dibandingkan kelas lainnya.

Selain itu, nilai *macro average F1-score* sebesar 0.9824 dan *weighted average F1-score* sebesar 0.9888 menunjukkan bahwa performa model stabil dan konsisten di seluruh kelas, serta tidak terlalu terpengaruh oleh ketidakseimbangan jumlah data antar kelas.

E. Perbandingan Kinerja MedViT dan Swin Transformer

Perbandingan kinerja model MedViT dan Swin Transformer dilakukan untuk mengetahui efektivitas masing-masing arsitektur dalam mengklasifikasikan citra kanker kandung kemih ke dalam empat kelas, yaitu High-Grade Carcinoma (HGC), Low-Grade Carcinoma (LGC), Non-Specific Tumor (NST), dan Non-Tumor Lesion (NTL). Kedua model dievaluasi menggunakan dataset, skema pembagian data, serta metrik evaluasi yang sama, sehingga perbandingan yang dilakukan bersifat adil dan objektif. Analisis kinerja difokuskan pada hasil pelatihan dan pengujian, meliputi nilai akurasi, *loss*, *confusion matrix*, ROC-AUC, serta metrik *precision*, *recall*, dan *F1-score*. Dengan membandingkan hasil evaluasi tersebut, bagian ini bertujuan untuk mengidentifikasi keunggulan dan keterbatasan masing-masing model, serta menentukan model yang lebih optimal dan andal untuk tugas klasifikasi citra medis, khususnya pada kasus kanker kandung kemih.

Tabel 3 Tabel Perbandingan Kinerja MedViT dan Swin Transformer

Aspek Evaluasi	MedViT	Swin Transformer
Arsitektur Dasar	Hybrid CNN-Transformer (LFP + GFP)	Hierarchical Vision Transformer
Jumlah Kelas	4 kelas (HGC, LGC, NST, NTL)	4 kelas (HGC, LGC, NST, NTL)
Akurasi Data Uji	95.49%	98.87%
Precision (Weighted Avg)	95.66%	98.93%
Recall (Weighted Avg)	95.49%	98.87%
F1-Score (Weighted Avg)	95.46%	98.88%
ROC-AUC (Rata-rata Kelas)	0.97	1.00
Kinerja Kelas Minor (NTL)	Precision tinggi, recall lebih rendah	Recall & precision sangat tinggi
Stabilitas Training	Stabil, val loss fluktuatif ringan	Sangat stabil
Waktu Training	± 6.05 menit	± 7.01 menit

Berdasarkan Tabel 1, dapat dilihat bahwa baik MedViT maupun Swin Transformer sama-sama mampu melakukan klasifikasi citra kanker kandung kemih ke dalam empat kelas, yaitu HGC, LGC, NST, dan NTL, dengan performa yang sangat baik. Namun demikian, terdapat perbedaan karakteristik dan tingkat kinerja antara kedua model tersebut.

Dari sisi arsitektur, MedViT mengadopsi pendekatan hybrid CNN–Transformer yang menggabungkan *Local Feature Processing (LFP)* dan *Global Feature Processing (GFP)*. Pendekatan ini memungkinkan MedViT menangkap fitur lokal dan global secara seimbang, yang sangat penting pada citra medis. Sementara itu, Swin Transformer menggunakan Hierarchical Vision Transformer dengan mekanisme *shifted window attention* yang dirancang untuk menangkap konteks global secara efisien pada berbagai skala resolusi.

Ditinjau dari akurasi data uji, Swin Transformer menunjukkan kinerja yang lebih unggul dengan akurasi sebesar 98,87%, dibandingkan MedViT yang mencapai 95,49%. Perbedaan ini menunjukkan bahwa Swin Transformer memiliki kemampuan generalisasi yang lebih kuat pada dataset pengujian.

Pada metrik evaluasi lanjutan, yaitu precision, recall, dan F1-score (weighted average), Swin Transformer secara konsisten memperoleh nilai yang lebih tinggi dibandingkan MedViT. Hal ini mengindikasikan bahwa Swin Transformer tidak hanya lebih akurat secara keseluruhan, tetapi juga lebih seimbang dalam meminimalkan kesalahan prediksi pada seluruh kelas.

Dari aspek ROC-AUC, Swin Transformer mencapai nilai 1.00, yang menunjukkan kemampuan pemisahan kelas yang sangat sempurna. Sebaliknya, MedViT memperoleh nilai 0.97, yang tetap tergolong sangat tinggi, namun masih berada sedikit di bawah Swin Transformer.

Kinerja pada kelas minor (NTL) menunjukkan perbedaan yang cukup signifikan. MedViT memiliki precision tinggi namun recall lebih rendah, yang mengindikasikan bahwa model cenderung lebih berhati-hati dalam memprediksi kelas NTL, tetapi masih melewatkan sebagian sampel yang seharusnya terdeteksi. Sebaliknya, Swin Transformer menunjukkan recall dan precision yang sangat tinggi, menandakan kemampuan yang lebih baik dalam mendeteksi kelas minor secara konsisten.

Dari sisi stabilitas training, kedua model menunjukkan proses pelatihan yang stabil. Namun, Swin Transformer terlihat lebih konsisten dengan fluktuasi loss yang minimal dibandingkan MedViT yang masih mengalami fluktuasi ringan pada validation loss. Meskipun demikian, MedViT memiliki waktu pelatihan yang lebih singkat ($\pm 6,05$ menit) dibandingkan Swin Transformer ($\pm 7,01$ menit), sehingga MedViT lebih efisien dari segi komputasi.

Secara keseluruhan, hasil perbandingan ini menunjukkan bahwa Swin Transformer unggul dalam hal performa klasifikasi dan stabilitas, sedangkan MedViT menawarkan keunggulan dari sisi efisiensi dan desain arsitektur yang lebih ringan, sehingga keduanya memiliki potensi penggunaan yang berbeda tergantung pada kebutuhan klinis dan keterbatasan sumber daya komputasi.

IV. KESIMPULAN

Penelitian ini berfokus pada penerapan dan perbandingan dua arsitektur deep learning berbasis Vision Transformer, yaitu MedViT dan Swin Transformer, dalam tugas klasifikasi citra kanker kandung kemih. Berdasarkan seluruh rangkaian eksperimen yang telah dilakukan, mulai dari pelatihan model hingga evaluasi pada data uji, diperoleh beberapa temuan utama yang dirangkum sebagai berikut.

- 1) MedViT dan Swin Transformer sama-sama mampu mencapai kinerja klasifikasi yang sangat baik pada dataset kanker kandung kemih dengan empat kelas (HGC, LGC, NST, dan NTL), yang ditunjukkan oleh nilai akurasi, precision, recall, dan F1-score yang tinggi pada data uji.
- 2) Model Swin Transformer menunjukkan performa yang lebih unggul secara konsisten dibandingkan MedViT, dengan akurasi data uji sebesar 98,87%, weighted F1-score 98,88%, serta nilai ROC-AUC sempurna (1.00) pada seluruh kelas, menandakan kemampuan diskriminasi kelas yang sangat kuat.
- 3) MedViT tetap menunjukkan keunggulan dalam efisiensi dan stabilitas pelatihan, dengan waktu training yang lebih singkat dan performa yang kompetitif (akurasi 95,49%), sehingga model ini berpotensi digunakan pada skenario dengan keterbatasan sumber daya komputasi atau ukuran dataset yang relatif kecil.
- 4) Analisis confusion matrix dan classification report menunjukkan bahwa kelas minoritas (NTL) merupakan kelas yang paling menantang bagi kedua model. Namun, Swin Transformer mampu mempertahankan recall yang sangat tinggi pada kelas ini, sementara MedViT masih menunjukkan kecenderungan kesalahan klasifikasi antar kelas yang memiliki karakteristik visual serupa.
- 5) Secara keseluruhan, Swin Transformer lebih direkomendasikan untuk aplikasi klinis yang menuntut akurasi dan generalisasi tinggi, sedangkan MedViT dapat menjadi alternatif yang efektif dan ringan untuk sistem pendukung keputusan medis dengan keterbatasan data atau komputasi.

DAFTAR PUSTAKA

- [1] R. Shadab, R. B. Nerli, S. R. Bidi, and S. C. Ghagane, "Risk Factors for Bladder Cancer: Results of a Survey of Hospital Patients," *J Cancer Allied Spec*, vol. 8, no. 2, p. 4, 2022, doi: 10.37029/jcas.
- [2] L. Lutviana, Rian Ardianto, and Purwono, "CNN-based Classification of Bladder Tissue Lesions from Endoscopy Images," *IT Journal Research and Development*, vol. 9, no. 2, pp. 95–107, Mar. 2025, doi: 10.25299/itjrd.2025.17867.
- [3] A. Lopez-Beltran, M. S. Cookson, B. J. Guercio, and L. Cheng, "Advances in diagnosis and treatment of bladder cancer," *BMJ*, Feb. 2024, doi: 10.1136/bmj-2023-076743.
- [4] S. Alouiini, "Risk Factors Associated with Urothelial Bladder Cancer," *Int J Environ Res Public Health*, vol. 21, no. 7, pp. 1–12, Jul. 2024, doi: 10.3390/ijerph21070954.
- [5] S. H. Lee, S. Lee, and B. C. Song, "Vision Transformer for Small-Size Datasets," *Computer Vision and Pattern Recognition*, Dec. 2021, doi: <https://doi.org/10.48550/arXiv.2112.13492>.
- [6] E. U. Henry, O. Emebo, and C. A. Omonhinmin, "Vision Transformers in Medical Imaging: A Review," *Computer Vision and Pattern Recognition*, Nov. 2022, doi: <https://doi.org/10.48550/arXiv.2211.10043>.
- [7] Z. Liu *et al.*, "Swin Transformer: Hierarchical Vision Transformer using Shifted Windows," *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 10012–10022, 2021, Accessed: Dec. 14, 2025. [Online]. Available: https://openaccess.thecvf.com/content/ICCV2021/papers/Liu_Swin_Transformer_Hierarchical_Vision_Transformer_Using_Shifted_Windows_ICCV_2021_paper.pdf
- [8] Z. Liu *et al.*, "Swin Transformer V2: Scaling Up Capacity and Resolution," *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 12009–12019, 2022, Accessed: Dec. 14, 2025. [Online]. Available: https://openaccess.thecvf.com/content/CVPR2022/papers/Liu_Swin_Transformer_V2_Scaling_Up_Capacity_and_Resolution_CVPR_2022_paper.pdf
- [9] J. Huang *et al.*, "Swin transformer for fast MRI," *Neurocomputing*, vol. 493, pp. 281–304, Jul. 2022, doi: 10.1016/j.neucom.2022.04.051.
- [10] Z. Xie *et al.*, "Self-Supervised Learning with Swin Transformers," *Computer Vision and Pattern Recognition*, May 2021, doi: <https://doi.org/10.48550/arXiv.2105.04553>.
- [11] Y. Liu, "Medical Image Classification Based on Transformer Model and Ordinal Loss," in *Proceedings of the 1st International Conference on Engineering Management, Information Technology and Intelligence (EMITI 2024, INSTICC, Sep. 2024*, pp. 708–713. doi: 10.5220/0012969200004508.
- [12] R. Karim, H. Zhao, R. P. Wildes, and M. Siam, "MED-VT: Multiscale Encoder-Decoder Video Transformer with Application to Object Segmentation," *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 6323–6333, 2023.
- [13] V. Liyanage, M. Tao, J. S. Park, K. N. Wang, and S. Azimi, "Malignant and non-malignant oral lesions classification and diagnosis with deep neural networks," *J Dent*, vol. 137, Oct. 2023, doi: 10.1016/j.jdent.2023.104657.