BAB II

TINJAUAN PUSTAKA

2.1 Justifikasi Solusi

Terdapat beberapa penelitian terkait dengan metode implementasi sistem yang mampu mensegmentasi tumor pada citra histopatologi.

2.1.1 A deep learning–based algorithm for tall cell detection in papillary thyroid carcinoma

Penelitian yang dilakukan oleh Stenman, et al membangun algoritma *deep learning* yang dapat mendeteksi *tall cell*. Algoritma yang dibangun terdiri atas dua tahap utama, yaitu segmentasi jaringan tumor dan klasifikasi area epitel menjadi *tall cell* dan *non tall cell*. Sistem ini dilatih menggunakan citra WSI dengan pewarnaan *Hematoxylin dan Eosin (H&E)* dan dievaluasi pada dataset pasien PTC yang berbeda. Hasil penelitian menunjukkan bahwa algoritma yang dikembangkan pada tahap pertama mampu mensegmentasi tumor dengan nilai *positive predictive value* (PPV; *precision*), sensitivitas, dan *F1-score* yang mencapai 99%. Algoritma pada tahap kedua dapat mengidentifikasi *tall cell* dengan akurasi tinggi, yaitu sensitivitas 93,7% dan spesifisitas 94,5%, serta menghasilkan skor persentase *tall cell* yang memiliki korelasi signifikan dengan *relapse free survival* pasien. Secara keseluruhan, metode berbasis *deep learning* ini terbukti dapat memberikan penilaian morfologi yang lebih objektif dan konsisten [11].

Beberapa poin penting yang dapat diambil oleh penulis adalah sebagai berikut:

- Segmentasi jaringan tumor pada citra WSI merupakan tahapan fundamental yang harus dilakukan terlebih dahulu untuk memisahkan area kanker dari jaringan sehat, sebelum dapat dilakukan analisis lanjutan seperti pengukuran morfologi sel

maupun perhitungan persentase penyebaran sel dalam tumor secara akurat.

- Pendekatan *deep learning* dapat diterapkan dalam proses pemisahan area tumor karena mampu mempelajari fitur jaringan tumor pada citra digital resolusi tinggi.
- Penggunaan citra WSI dengan pewarnaan H&E terkonfirmasi sebagai standar data yang ideal, karena mampu menyediakan resolusi tinggi yang mencakup konteks arsitektur jaringan secara menyeluruh untuk keperluan segmentasi.

2.1.2 GCSA-SegFormer: Transformer-Based Segmentation for Liver Tumor Pathological Images

Penelitian yang dilakukan oleh Li, et al. menerapkan pendekatan segmentasi berbasis arsitektur SegFormer untuk menangani analisis citra WSI dengan resolusi gigapiksel. Tujuan penelitian ini adalah melakukan segmentasi tumor hati secara presisi pada citra histopatologi. Data WSI diproses menggunakan pendekatan berbasis *patch*, di mana citra dipotong dengan teknik *sliding window* menjadi potongan berukuran 512x512 piksel sebelum digunakan sebagai input model. Hasil evaluasi model dibandingkan dengan beberapa arsitektur segmentasi lain, yaitu U-Net, SegNet, DeepLabV3, DeepLabV3+, PSPNet, HRNet, FCN, dan LR-ASPP. Berdasarkan hasil pengujian, model yang diusulkan menunjukkan nilai akurasi, mIoU, dan MPA yang lebih tinggi dibandingkan model lain [12].

Beberapa poin penting yang dapat diambil oleh penulis adalah sebagai berikut:

- Arsitektur SegFormer memiliki performa yang baik untuk tugas segmentasi tumor, sehingga dipilih sebagai model pada penelitian ini.

- Metode pemrosesan citra berbasis *patch* dengan ukuran 512x512 piksel dapat diterapkan dalam penelitian ini untuk memproses citra WSI secara efisien.

2.1.3 *SegFormer: Simple and Efficient Design for Semantic Segmentation with Transformers*

Penelitian yang dilakukan oleh Xie, et al. memperkenalkan SegFormer sebagai arsitektur segmentasi semantik berbasis *Transformer*. SegFormer terdiri atas *encoder* berbasis *Mix-Transformer* (MiT) yang mengekstraksi fitur secara hierarkis dan *decoder* ringan berbasis *multilayer perceptron* (MLP) untuk menghasilkan peta segmentasi. Berbeda dengan *vision transformer* (ViT), arsitektur ini tidak menggunakan *positional encoding* eksplisit dan mampu memproses input dengan ukuran bervariasi. SegFormer dievaluasi pada beberapa dataset segmentasi dan menunjukkan performa yang kompetitif dibandingkan arsitektur segmentasi konvensional berbasis CNN [13].

Poin yang dapat diambil oleh penulis adalah sebagai berikut:

- Penggunaan *backbone* MiT pada arsitektur SegFormer dapat diterapkan dalam penelitian ini, mengingat *backbone* tersebut merupakan bagian integral dari desain SegFormer sebagai *encoder* untuk segmentasi.

2.1.4 *Spatially Varying Label Smoothing: Capturing Uncertainty from Expert Annotations*

Penelitian yang dilakukan oleh Müller, et al. membahas teknik *label smoothing* dalam konteks pembelajaran model *deep learning* dengan mempertimbangkan ketidakpastian yang muncul dari proses anotasi oleh ahli. Penelitian ini menyoroti bahwa pada label *ground truth* yang digunakan dalam proses pelatihan model tidak selalu sepenuhnya pasti, terutama ketika anotasi dilakukan secara manual.

Label smoothing diperkenalkan sebagai teknik regularisasi untuk mengurangi *overconfidence* model selama proses pelatihan, di mana model cenderung menghasilkan prediksi dengan tingkat keyakinan yang sangat tinggi terhadap satu kelas sehingga kurang mampu melakukan generalisasi pada data baru. Pada proses *training*, label *ground truth* umumnya direpresentasikan menggunakan *one-hot encoding*, yaitu representasi label dengan satu kelas bernilai 1 dan kelas lainnya bernilai 0, yang mendorong model untuk mempelajari prediksi dengan sangat tegas. *Label smoothing* memodifikasi label *one-hot* agar model tidak terlalu memaksimalkan keyakinan pada satu kelas selama proses optimasi [14].

Poin yang dapat diambil oleh penulis sebagai berikut:

- *Label smoothing* dapat diterapkan dalam proses pelatihan model segmentasi untuk mengurangi *overconfidence* model terhadap label *ground truth* yang tidak *pixel-perfect*.

2.2 Tinjauan Teori

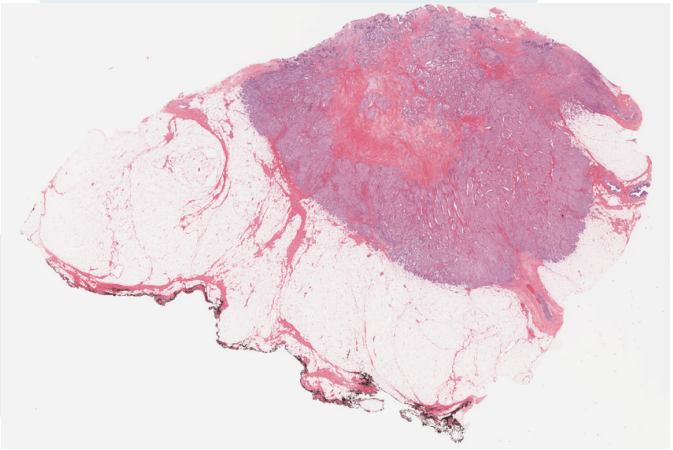
2.2.1 Histopatologi

Histopatologi merupakan cabang ilmu kedokteran yang mempelajari penyakit melalui pemeriksaan mikroskopis terhadap sampel jaringan tubuh. Dalam dunia medis, pemeriksaan ini dianggap sebagai *gold standard* untuk penegakan diagnosis tumor dan penentuan sifat keganasan suatu kanker. Sampel jaringan biasanya diperoleh melalui prosedur biopsi atau operasi bedah, kemudian diproses melalui serangkaian tahapan fiksasi dan pemotongan tipis agar dapat diamati di bawah mikroskop oleh dokter spesialis patologi anatomi. Jaringan biologis asli cenderung transparan dan sulit dibedakan strukturnya, oleh karena itu diperlukan proses pewarnaan (*staining*) untuk memberikan kontras visual pada komponen sel agar struktur

morfologi sel dan arsitektur jaringan dapat teramati dengan jelas untuk keperluan analisis diagnostik [15].

2.2.2 Whole Slide Image

Whole Slide Image (*WSI*) merupakan representasi digital dari sampel jaringan yang diperoleh melalui proses pemindaian menggunakan mikroskop digital beresolusi tinggi. Teknologi ini memungkinkan seluruh area jaringan pada kaca objek direkam secara menyeluruh menjadi citra digital yang dapat diamati melalui komputer. Dalam proses pemindaian, pembesaran optik yang umum digunakan adalah 20× hingga 40×, yang mampu menampilkan detail morfologi sel dengan jelas untuk keperluan diagnostik maupun penelitian. Setiap file WSI memiliki dimensi yang sangat besar hingga mencapai resolusi gigapiksel dan biasanya menyimpan metadata penting seperti *microns per pixel*, informasi pembesaran, serta data pewarnaan jaringan [16].



Gambar 2.1 Whole Slide Image
Sumber: Abdulsadig [23]

2.2.3 Deep Learning

Deep learning merupakan salah satu cabang dari *machine learning* yang menggunakan arsitektur *artificial neural networks* dengan *deep neural networks* untuk mempelajari representasi data secara otomatis. Melalui lapisan-lapisan *neuron* yang saling terhubung, *deep learning*

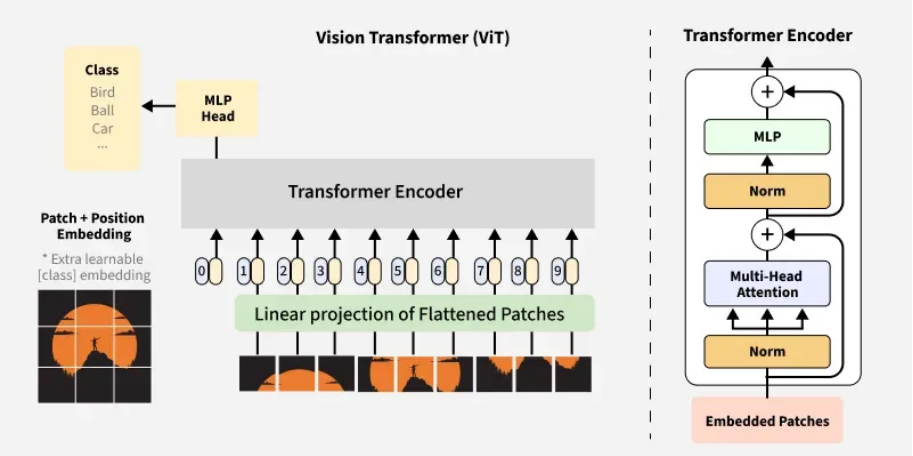
mampu mengenali struktur kompleks dari data mentah, misalnya gambar, suara, atau teks. *Deep learning* dapat dibedakan menjadi tiga pendekatan utama:

- *Supervised Learning*, yaitu pembelajaran menggunakan data berlabel, di mana setiap citra atau sampel telah memiliki kelas target yang diketahui.
- *Unsupervised Learning*, yaitu pembelajaran dengan data tanpa label, di mana model berusaha menemukan pola kemiripan dalam data, contohnya melalui teknik *clustering* atau *autoencoder*.
- *Semi-supervised Learning*, yaitu pendekatan gabungan yang memanfaatkan sebagian kecil data berlabel dan sebagian besar data tidak berlabel.

Secara prinsip, *deep learning* bekerja dengan memasukkan data ke dalam jaringan saraf yang terdiri dari beberapa lapisan, dan setiap lapisan melakukan proses transformasi matematis untuk mengekstraksi fitur yang semakin kompleks dari lapisan sebelumnya. Selama proses pelatihan, model memperbarui bobot koneksi antar *neuron* secara bertahap untuk meminimalkan kesalahan antara hasil prediksi dan target sebenarnya. [17].

2.2.4 Vision Transformer

Vision Transformer (ViT) merupakan adaptasi dari arsitektur *Transformer* yang awalnya dirancang untuk *Natural Language Processing* (NLP), yang kemudian digunakan dalam tugas *computer vision*. ViT memperlakukan input sebagai *patches* dan menggunakan mekanisme *self attention* untuk menangkap hubungan global antar piksel secara langsung. Pendekatan ini memungkinkan model untuk memahami *long range dependencies* di seluruh area gambar tanpa terbatasi oleh ukuran *receptive field lokal* seperti pada CNN.



Gambar 2.2 Arsitektur Vision Transformer
Sumber: GeeksforGeeks [24]

Arsitektur ViT umumnya bekerja sebagai berikut:

I. Image Patching & Flattening

Langkah pertama dalam ViT adalah membagi input 2D berukuran H x W x C menjadi serangkaian *patch* berukuran tetap (P x P). Setiap *patch* kemudian di *flattened* menjadi vektor 1D berdimensi $P^2 x C$. Proses ini bertujuan untuk mengubah data spasial 2D menjadi urutan data linear yang dapat diproses oleh blok *Transformer*.

II. Linear Projection of Flattened Patches

Vektor *patch* yang telah diratakan kemudian dipetakan ke dalam ruang dimensi melalui lapisan proyeksi linear yang dapat dilatih. Hasil proyeksi ini disebut sebagai *patch embeddings*. Pada tahap ini, model mulai mempelajari representasi fitur tingkat rendah dari setiap potongan gambar.

III. Positional Encoding

ViT menambahkan vektor posisi ke dalam setiap *patch embedding* agar model dapat memahami struktur spasial input gambar.

IV. Transformer Encoder

Patch embedding yang telah memuat informasi posisi dimasukkan ke dalam *Transformer Encoder*. *Encoder* ini terdiri dari tumpukan blok yang masing-masing berisi dua lapisan utama yaitu *Multi-Head Self-Attention* (MSA) dan *Feed-Forward Network* (FFN). Mekanisme *Layer Normalization* diterapkan sebelum setiap blok, dan *skip connections* diterapkan setelah setiap blok untuk menjaga aliran gradien tetap stabil selama pelatihan.

Komponen inti yang membedakan ViT dari arsitektur lain adalah mekanisme *self-attention*. Mekanisme ini memungkinkan model untuk mengevaluasi tingkat kepentingan atau relevansi antara satu *patch* dengan seluruh *patch* lainnya dalam citra secara bersamaan.

Dalam proses komputasinya, setiap vektor input X ditransformasikan menjadi tiga representasi vektor yang berbeda:

- *Query* (Q): Representasi fitur yang mencari informasi relevan.
- *Key* (K): Representasi fitur yang menjadi indeks atau target pencarian.
- *Value* (V): Representasi informasi yang sebenarnya.

Attention weights dihitung menggunakan fungsi *Scaled Dot-Product Attention*. Rumus matematisnya dinyatakan sebagai berikut:

$$A = softmax\left(\frac{QK^T}{\sqrt{d_k}}\right)V$$

Di mana QK^T mengukur kemiripan antara *Query* dan *Key*. Nilai ini kemudian dibagi dengan akar dimensi *Key* ($\sqrt{d_k}$) untuk penskalaan dan dinormalisasi menggunakan fungsi *Softmax* menjadi probabilitas. Probabilitas ini kemudian digunakan untuk memberikan bobot pada vektor *Value* (V). Untuk meningkatkan kemampuan representasi, ViT

menggunakan MSA dimana proses *attention* dijalankan beberapa kali secara paralel. Setiap *head* dapat fokus mempelajari aspek hubungan yang berbeda. Output dari seluruh *head* kemudian digabungkan dan diproyeksikan kembali untuk menghasilkan representasi fitur akhir [18].

2.2.5 Segmentasi

Segmentasi merupakan proses pemisahan sebuah gambar menjadi beberapa bagian atau wilayah yang memiliki karakteristik serupa, seperti warna, tekstur, atau intensitas piksel. Tujuan utama dari segmentasi adalah untuk menyederhanakan representasi suatu gambar agar lebih bermakna dan mudah dianalisis. Melalui proses ini, bagian-bagian penting dari gambar dapat diisolasi dari latar belakang, sehingga memungkinkan komputer untuk memahami struktur atau objek yang terkandung di dalamnya. Hasil segmentasi sering kali menjadi tahap awal dari berbagai tugas dalam bidang *computer vision*, seperti *object detection*, *image classification*, *object tracking*, hingga analisis bentuk dan ukuran. Secara umum, segmentasi dibedakan menjadi tiga jenis:

- *Semantic Segmentation*, yaitu proses pengelompokan setiap piksel ke dalam kelas tertentu tanpa membedakan objek individual.
- *Instance Segmentation*, yaitu segmentasi yang tidak hanya mengenali kelas objek, tetapi juga membedakan antarobjek yang termasuk dalam kelas yang sama.
- *Panoptic Segmentation*, yaitu pendekatan yang menggabungkan kedua metode sebelumnya, dengan memberikan label kelas pada setiap piksel sekaligus mengidentifikasi batas setiap objek individual.

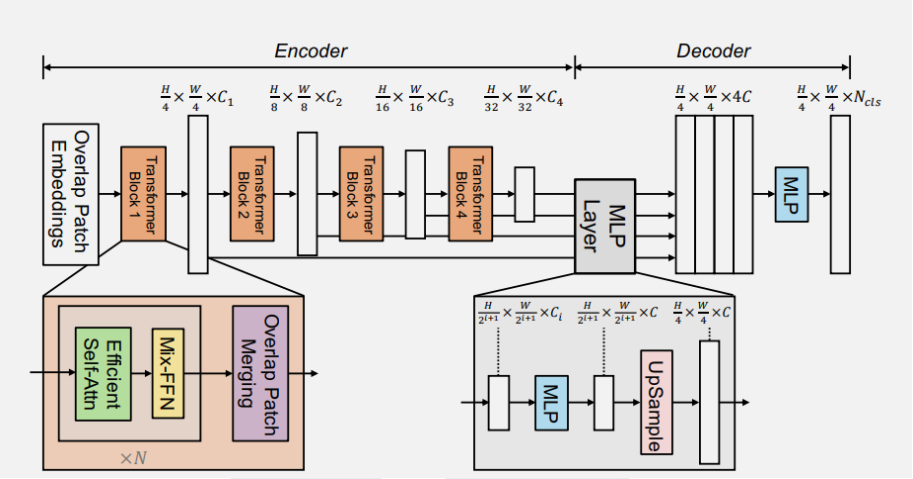
Secara prinsip, segmentasi bekerja dengan menganalisis nilai piksel dan hubungan spasial antar wilayah untuk menentukan batas

objek secara otomatis. Algoritma modern umumnya menggunakan pendekatan *Deep Learning*, baik berbasis CNN maupun *Transformer*, yang mampu mempelajari pola visual kompleks dari data pelatihan. Setiap lapisan dalam jaringan memproses citra untuk mengekstraksi fitur-fitur penting, seperti *edge*, bentuk, atau pola tekstur, kemudian menggabungkannya untuk menghasilkan *segmentation map* yang menandai wilayah objek secara akurat. Hasil akhir dari proses ini berupa gambar dengan area yang telah diberi label sesuai kelas atau identitas objek. [19].

2.2.6 Arsitektur Segformer

SegFormer merupakan arsitektur *deep learning* yang digunakan untuk tugas segmentasi semantik dengan pendekatan *transformer*. SegFormer dirancang untuk memanfaatkan kemampuan *transformer* dalam menangkap informasi konteks dari area gambar yang luas, dengan tetap mempertahankan efisiensi komputasi yang diperlukan pada proses segmentasi. Namun, penggunaan *transformer* konvensional seringkali terkendala oleh kompleksitas komputasi dan ketergantungan pada ukuran input tertentu [13].

Untuk mengatasi keterbatasan tersebut, SegFormer menggunakan arsitektur *encoder–decoder* yang efisien dan fleksibel. SegFormer tidak menggunakan *positional encoding* eksplisit dan mengekstraksi fitur gambar secara bertahap dari resolusi tinggi ke resolusi yang lebih rendah. Pendekatan ini memungkinkan model untuk mempelajari informasi visual pada berbagai tingkat detail. Selain itu, SegFormer menggunakan *decoder* yang ringan untuk menyatukan informasi dari beberapa tingkat resolusi tersebut dan menghasilkan peta segmentasi tanpa menambah kompleksitas model secara signifikan.



Gambar 2.3 Framework Arsitektur Segformer
Sumber: Xie [13]

Secara keseluruhan, SegFormer terdiri atas dua modul utama, yaitu *encoder* berbasis MiT dan *decoder* berbasis MLP.

- Encoder Mix-Transformer (MiT)

  Encoder pada arsitektur SegFormer menggunakan Mix Transformer (MiT) sebagai *backbone* untuk ekstraksi fitur. Secara konseptual, MiT masih berlandaskan mekanisme *self-attention* sebagaimana pada *Vision Transformer,* namun dirancang khusus untuk kebutuhan segmentasi semantik yang memerlukan representasi fitur pada berbagai tingkat resolusi. Berbeda dengan *Vision Transformer* konvensional yang menggunakan *patch non-overlapping*, MiT menerapkan *overlapping patch embedding* pada tahap awal untuk menjaga hubungan informasi antar area gambar yang berdekatan. Fitur yang dihasilkan kemudian diproses melalui beberapa *stage transformer* yang tersusun secara hierarkis, di mana peta fitur diproses dari resolusi tinggi ke resolusi yang lebih rendah, dengan jumlah informasi fitur yang direpresentasikan meningkat pada setiap tahap pemrosesan. Selain itu, MiT mengadopsi *Mix Feed-Forward Network* (Mix-FFN) agar fitur

yang dihasilkan mencakup informasi global dan lokal secara seimbang.

- Decoder Multilayer Perceptron (MLP)

  Decoder pada arsitektur SegFormer menggunakan *multilayer perceptron* (MLP) untuk mengubah fitur hasil ekstraksi *encoder* menjadi peta segmentasi akhir. *Decoder* menerima keluaran fitur dari setiap *stage encoder* MiT yang memiliki ukuran resolusi berbeda, kemudian memproyeksikannya ke dimensi yang sama menggunakan lapisan MLP agar seluruh fitur memiliki format yang seragam. Selanjutnya, fitur-fitur tersebut diubah ke resolusi spasial yang sama melalui operasi *upsampling* dan digabungkan menjadi satu peta fitur yang mengandung informasi dari seluruh tingkat resolusi. Tahap akhir *decoder* memetakan peta fitur tersebut ke dalam peta segmentasi, di mana setiap piksel diberikan prediksi kelas.