BAB II

TINJAUAN PUSTAKA

2.1 Tinjauan Pustaka

2.1.1 Artificial intelligence for enhanced diagnostic precision of prostate cancer

Hamid, et al. telah membangun sebuah model *deep learning* yang ditujukan untuk mengklasifikasikan kelas jinak dan ganas pada prostat. Dalam praktiknya, penelitian ini memanfaatkan data berupa *whole slide image* (WSI) dengan pewarnaan *Hematoxylin dan Eosin* (H&E). Proses pengolahan data diawali dengan ekstraksi *region of interest* (ROI) dari WSI yang kemudian dipotong menjadi *patch* berukuran 512 x 512 piksel untuk mempermudah proses pelatihan model. Penelitian ini mengevaluasi beberapa arsitektur *deep learning* yang terdiri atas kelompok *Convolutional Neural Network* (CNN), *Transformer*, dan *Hybrid* dalam melakukan klasifikasi citra histopatologi prostat. Performa masing-masing model dievaluasi menggunakan beberapa metrik klasifikasi, seperti *accuracy*, *sensitivity*, *specificity*, *positive predictive value* (PPV), *negative predictive value* (NPV), *F1-score*, dan *area under the receiver operating characteristic curve* (AUC ROC). Hasil penelitian menunjukkan bahwa pendekatan *deep learning* mampu memberikan performa yang baik dalam membedakan jaringan prostat jinak dan ganas [11].

Terdapat sejumlah catatan teknis dari studi tersebut yang menjadi acuan penulis:

1. Data WSI dengan pewarnaan H&E terbukti menjadi format yang paling sesuai untuk tugas klasifikasi. Hal ini dikarenakan format tersebut menyajikan detail resolusi tinggi yang mempertahankan informasi arsitektur jaringan secara utuh.

2. Metode *patching* dengan ukuran 512 x 512 piksel dapat dipakai mengurangi kompleksitas komputasi pada citra histopatologi beresolusi tinggi.
3. Pendekatan *deep learning* menggunakan arsitektur CNN, *Transformer*, dan *Hybrid* terbukti mampu menghasilkan performa yang baik dalam proses klasifikasi citra histopatologi.
4. Metrik *accuracy*, *sensitivity*, *specificity*, PPV, NPV, *F1-score*, dan AUC dapat digunakan untuk mengevaluasi performa model klasifikasi.

2.1.2 CAMIL: channel attention-based multiple instance learning for whole slide image classification

Penelitian yang dilakukan oleh Mao, et al. mengembangkan *Channel Attention-based Multiple Instance Learning* (CAMIL), yaitu metode klasifikasi citra histopatologi berbasis *Multiple Instance Learning* (MIL) untuk menganalisis *whole slide image* (WSI). Pada pendekatan ini, setiap WSI direpresentasikan sebagai sekumpulan *patch* yang digabungkan menjadi sebuah *bag* dan diproses tanpa memerlukan anotasi pada tingkat *patch*. Model yang diusulkan memanfaatkan mekanisme *attention* untuk menentukan kontribusi setiap *patch* terhadap hasil klasifikasi sehingga informasi yang paling relevan dapat diberikan bobot yang lebih besar dalam proses pengambilan keputusan. Untuk mencegah terjadinya *data leakage*, pembagian data dilakukan pada level pasien sehingga data dari pasien yang sama tidak muncul pada *subset* yang berbeda. Selain itu, proses pelatihan dan evaluasi model dilakukan menggunakan metode *5-fold cross validation* untuk memperoleh estimasi performa yang lebih stabil dan representatif. Hasil penelitian menunjukkan bahwa metode CAMIL mampu memberikan performa yang baik pada tugas klasifikasi citra histopatologi dan menunjukkan peningkatan dibandingkan beberapa pendekatan *Multiple Instance Learning* yang telah ada sebelumnya [12].

Terdapat sejumlah catatan teknis dari studi tersebut yang menjadi acuan penulis:

1. Untuk menghindari *data leakage*, pembagian dataset akan menggunakan pendekatan *patient-level split*.
2. Metode *5-fold cross validation* dapat digunakan untuk menghasilkan proses pelatihan dan evaluasi model yang lebih stabil serta representatif.
3. Pendekatan *Attention-based Multiple Instance Learning* (MIL) dapat diterapkan untuk menggabungkan informasi dari sekumpulan *patch* menjadi satu representasi *bag* yang digunakan dalam proses klasifikasi.

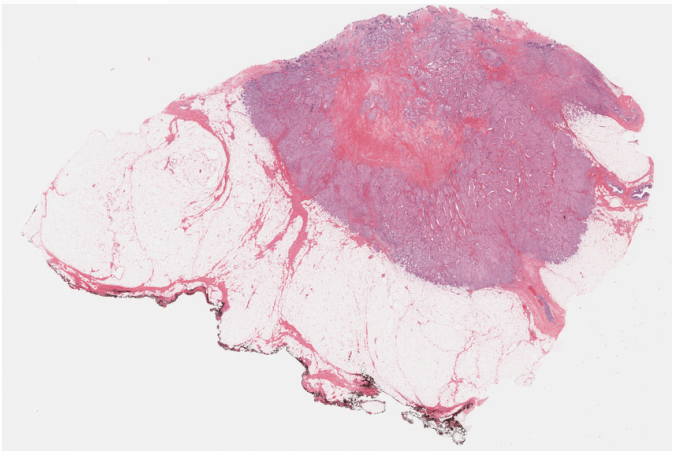
2.2 Tinjauan Teori

2.2.1 Histopatologi

Histopatologi adalah ilmu medis yang berfokus pada evaluasi penyakit lewat pengamatan mikroskopis pada spesimen jaringan tubuh. Dalam praktik klinis, prosedur ini diakui sebagai *gold standard* guna mendiagnosis tingkat keganasan sebuah kanker. Spesimen jaringan biasanya diambil dari tindakan biopsi ataupun pembedahan, yang selanjutnya melewati tahap preparasi seperti fiksasi dan pengirisan tipis. Tahapan ini diperlukan agar spesimen tersebut siap dianalisis menggunakan mikroskop oleh dokter spesialis patologi anatomi. Mengingat jaringan biologis aslinya tampak transparan dengan kontras visual yang minim, tahapan pewarnaan (*staining*) menjadi wajib dilakukan supaya arsitektur jaringan dan struktur sel di dalamnya terlihat jelas. Proses ini memungkinkan karakteristik morfologi jaringan diamati dengan lebih jelas sehingga dapat mendukung analisis dan penegakan diagnosis secara akurat [13].

2.2.2 Whole Slide Image

Whole Slide Image (WSI) adalah bentuk digitalisasi dari spesimen jaringan yang ditangkap melalui mikroskop digital beresolusi tinggi. Pemanfaatan teknologi ini memfasilitasi perekaman seluruh permukaan jaringan di atas kaca preparat menjadi sebuah gambar digital, yang selanjutnya siap dievaluasi menggunakan perangkat komputasi. Pada tahap pemindaiannya, tingkat pembesaran optik yang sering diterapkan berada di rentang 20× sampai 40×. Rentang pembesaran ini memastikan arsitektur jaringan dan morfologi selular tampak tajam guna menunjang keperluan riset maupun penegakan diagnosis. [14].



Gambar 2.1 Whole Slide Image [15]

2.2.3 Deep Learning

Deep learning adalah bagian dari *machine learning* yang digunakan untuk mempelajari representasi data secara otomatis. Dengan memanfaatkan interkoneksi antar neuron pada berbagai lapisannya,, *deep learning* mampu mengenali pola dan struktur kompleks dari data mentah, seperti citra, suara, maupun teks. Berdasarkan ketersediaan data berlabel, metode *deep learning* umumnya diklasifikasikan ke dalam tiga jenis:

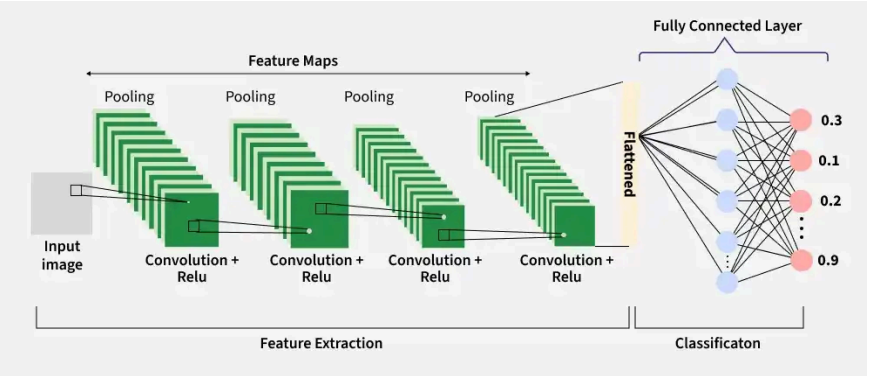
- *Supervised Learning*: Proses pelatihan yang sepenuhnya menggunakan data berlabel, sehingga setiap sampel input sudah memiliki target kelas yang pasti.

- *Unsupervised Learning*: Metode pembelajaran pada data yang tidak memiliki label klasifikasi. Pada kategori ini, model mencari pola kemiripan dalam data secara mandiri.
- *Semi-supervised Learning*: Teknik yang mengkombinasikan proporsi kecil data berlabel dengan dataset tanpa label dalam jumlah besar.

Secara konseptual, alur kerja *deep learning* dimulai dengan memberikan input data ke dalam arsitektur jaringan yang terdiri dari banyak lapisan. Setiap lapisannya secara berurutan mentransformasi data untuk melakukan ekstraksi fitur. Semakin dalam posisi sebuah lapisan, representasi fitur yang diekstrak akan menjadi semakin kompleks dibandingkan lapisan sebelumnya. Sepanjang fase pelatihan, model akan terus memperbarui parameter bobot pada koneksi antar neuronnya. Pembaruan ini bertujuan untuk meminimalkan tingkat *error* atau selisih antara tebakan prediksi dengan label aslinya, sehingga akurasi sistem dalam membaca pola data akan semakin optimal. [16].

2.2.4 Convolutional Neural Network

Convolutional Neural Network (CNN) merupakan arsitektur *deep learning* yang secara khusus dikembangkan untuk menganalisis data visual seperti citra digital. Keunggulan utama CNN terletak pada kemampuannya untuk mempelajari fitur-fitur penting dari gambar secara otomatis tanpa menghilangkan informasi spasial atau posisi antar pikselnya. CNN menjadi metode standar yang digunakan dalam berbagai tugas *computer vision*.



Gambar 2.2 Arsitektur CNN [17]

Arsitektur CNN umumnya tersusun dari tiga jenis lapisan utama yang bekerja secara berurutan untuk mentransformasi input menjadi output:

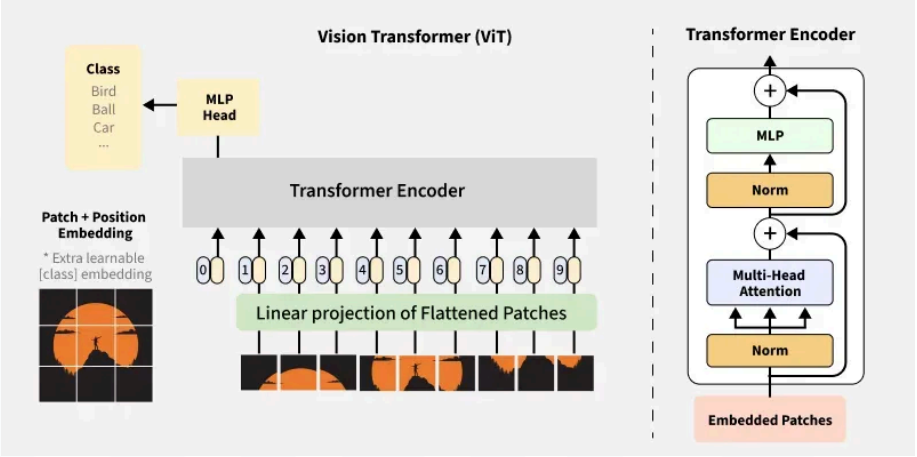
- *Convolutional Layer*, merupakan lapisan inti yang berfungsi untuk mengekstraksi fitur dari input gambar. Lapisan ini menggunakan filter (kernel) yang bergeser di sepanjang area gambar untuk mendeteksi pola-pola visual tertentu, seperti garis, tepi, atau bentuk geometris, dan menghasilkan luaran berupa *feature maps*.
- *Pooling Layer*, merupakan lapisan yang bertujuan untuk mengurangi dimensi dari *feature maps* yang dihasilkan *convolutional layer*. Proses ini dilakukan untuk mengurangi beban komputasi dan mencegah overfitting, namun tetap mempertahankan informasi fitur yang paling dominan.
- *Fully Connected Layer*, merupakan lapisan akhir di mana seluruh neuron terhubung penuh dengan neuron di lapisan sebelumnya. Lapisan ini menerima data fitur yang telah melewati proses *flattening*, lalu menghitung nilai probabilitas setiap kelas untuk menentukan hasil prediksi akhir.

Secara prinsip, CNN bekerja dengan membangun hierarki fitur yang semakin kompleks di setiap lapisannya. Pada tahap awal, jaringan akan mengenali fitur sederhana seperti goresan atau tepi. Lapisan-lapisan berikutnya kemudian menggabungkan fitur-fitur dasar tersebut untuk

mengenali pola yang lebih rumit, seperti tekstur atau bagian objek, hingga akhirnya *fully connected layer* menerjemahkan pola abstrak tersebut menjadi sebuah prediksi akhir [18].

2.2.5 Vision Transformer

Vision Transformer (ViT) awalnya dibangun untuk kebutuhan *Natural Language Processing* (NLP), sebelum akhirnya diadaptasi secara luas untuk menyelesaikan berbagai tugas di bidang *computer vision*. Jika *Convolutional Neural Network* (CNN) mengandalkan operasi konvolusi untuk mengolah informasi, ViT memiliki pendekatan berbeda dengan memecah citra menjadi sekumpulan *patches*. Model ini kemudian menggunakan mekanisme *self-attention* guna mengenali keterkaitan global antar bagian pada citra tersebut. Lewat cara ini, ViT mampu menangkap *long-range dependencies* tanpa harus terkekang oleh batasan ukuran *receptive field* lokal yang biasa ditemui pada CNN.



Gambar 2.3 Arsitektur Vision Transformer [19]

Secara garis besar, arsitektur ViT beroperasi melalui tahapan-tahapan berikut:

I. Image Patching & Flattening

Proses diawali dengan membagi citra masukan berdimensi H x W x C ke dalam sejumlah *patch* dengan ukuran seragam P x P.

Masing-masing *patch* selanjutnya diubah menjadi representasi vektor satu dimensi lewat proses *flattening*, sehingga siap diterima oleh arsitektur Transformer.

II. Linear Projection of Flattened Patches

Vektor-vektor yang telah melalui tahap *flattening* lalu diproyeksikan secara linear ke dalam suatu ruang dimensi spesifik. Luaran dari proyeksi yang dapat dilatih ini dikenal dengan istilah *patch embeddings*, yang berfungsi sebagai representasi informasi dasar dari setiap potongan gambar.

III. Positional Embedding

Karena sifat asli Transformer tidak mampu mengenali urutan tata letak spasial, sistem perlu menyisipkan *positional embedding* ke dalam setiap *patch embedding*. Penambahan ini krusial agar model dapat memahami lokasi relatif setiap *patch* pada gambar aslinya.

IV. Transformer Encoder

Sekumpulan *patch embeddings* yang telah dibekali informasi posisi kemudian diteruskan ke dalam *Transformer Encoder*. Komponen ini tersusun dari sejumlah blok repetitif yang masing-masingnya memuat *Multi-Head Self-Attention* (MSA) dan *Feed-Forward Network* (FFN). Untuk menjaga kestabilan selama proses pelatihan berlangsung, tahap ini turut mengimplementasikan *Layer Normalization* serta *skip connections*.

Keberadaan mekanisme *self-attention* menjadi pembeda paling mencolok antara ViT dengan model arsitektur lainnya. Melalui mekanisme ini, sistem dapat mengevaluasi tingkat hubungan antara satu *patch* dengan semua *patch* lainnya di waktu yang sama, sehingga konteks gambar secara keseluruhan (*global context*) bisa tertangkap. Pada tahapan perhitungannya, tiap vektor input (X) akan ditransformasikan ke dalam

tiga bentuk representasi, yakni *Query* (Q), *Key* (K), dan *Value* (V). Komponen *Query* bertugas mencari informasi yang relevan, *Key* bertindak sebagai acuan pencariannya, sementara *Value* membawa intisari informasi yang nanti akan dipertimbangkan pada proses pembentukan fitur akhir.

Rumus *attention-weights* adalah sebagai berikut:

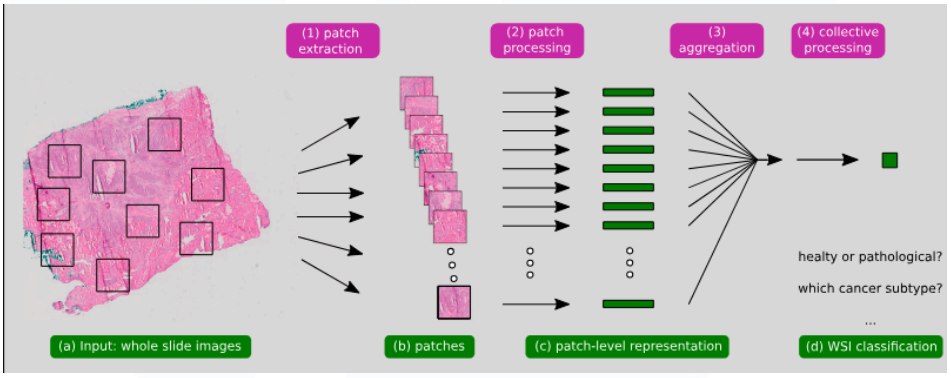
$$A = softmax\left(\frac{QK^T}{\sqrt{d_k}}\right)V$$

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 [20].

2.2.6 Multiple Instance Learning

Multiple Instance Learning (MIL) merupakan salah satu pendekatan *weakly supervised learning* yang mempelajari hubungan antara sekumpulan data (*instances*) dengan satu label pada tingkat kelompok (*bag*). Berbeda dengan pembelajaran konvensional yang memberikan label pada setiap sampel, MIL hanya memerlukan label pada tingkat *bag*, sedangkan setiap *instance* di dalamnya tidak memiliki label individual. Pendekatan ini banyak digunakan pada bidang *digital pathology* karena anotasi pada tingkat patch atau piksel memerlukan waktu dan tenaga yang besar, sedangkan label pada tingkat *Whole Slide Image* (WSI) umumnya telah tersedia sebagai bagian dari proses diagnosis rutin.

Dalam analisis citra histopatologi, satu WSI atau ROI dapat dipandang sebagai sebuah *bag* yang terdiri atas sejumlah patch atau *instances*. Setiap patch terlebih dahulu diubah menjadi representasi fitur dalam bentuk vektor, kemudian seluruh vektor tersebut dikumpulkan untuk membentuk satu *bag*. Informasi dari kumpulan vektor tersebut digunakan untuk menghasilkan prediksi kelas citra secara keseluruhan



Gambar 2.4 Alur Multiple Instance Learning [21]

Secara umum, MIL melibatkan beberapa tahapan utama, yaitu ekstraksi *patch* dari citra asli, pembentukan representasi fitur untuk setiap *patch*, penggabungan informasi dari seluruh *instance* dalam satu *bag*, serta klasifikasi untuk menghasilkan prediksi akhir. Berdasarkan cara pengolahannya, MIL dapat dibedakan menjadi *instance-based MIL* dan *embedding-based MIL*. Pada *instance-based MIL*, setiap *instance* memberikan kontribusi langsung terhadap keputusan klasifikasi. Sementara *embedding-based MIL* terlebih dahulu mengubah setiap *instance* menjadi vektor fitur sebelum dilakukan proses agregasi. Pendekatan *embedding-based* dilaporkan memiliki kemampuan yang lebih baik untuk tugas klasifikasi pada tingkat WSI dibandingkan pendekatan *instance-based*.

Seiring perkembangan *deep learning*, berbagai arsitektur MIL modern mulai mengintegrasikan mekanisme *attention* pada tahap agregasi fitur.

Mekanisme ini memungkinkan model memberikan bobot yang berbeda pada setiap *instance* sesuai tingkat relevansinya terhadap proses klasifikasi. Dengan demikian, area jaringan yang memiliki karakteristik penting dapat memberikan kontribusi yang lebih besar terhadap keputusan akhir model, sehingga meningkatkan kemampuan klasifikasi.[21]