BAB III
METODOLOGI PENELITIAN

3.1 Metode Penelitian

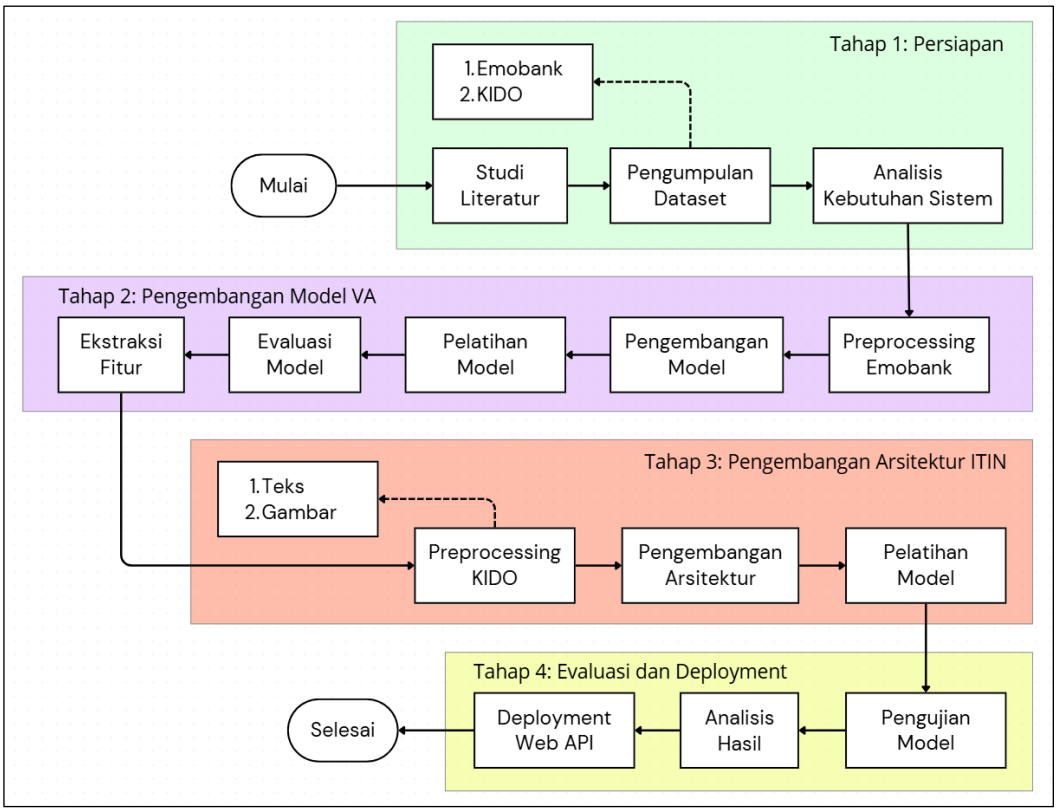
Penelitian ini menggunakan metode eksperimental dengan pendekatan *deep learning* untuk mengembangkan sistem klasifikasi emosi berbasis gambar coretan anak secara multimodal. Metode eksperimental merupakan metode penelitian yang bertujuan untuk mencari pengaruh perlakuan tertentu terhadap yang lain dalam kondisi yang terkendalikan (Sugiyono, 2019). Metode eksperimental dipilih karena penelitian ini bertujuan untuk menguji hipotesis bahwa integrasi fitur *Valence-Arousal* (VA) dapat meningkatkan akurasi klasifikasi emosi dibandingkan dengan pendekatan konvensional yang hanya mengandalkan label diskret.

3.2 Alur Penelitian

Alur penelitian diadaptasi dari kerangka eksperimental Zhu dkk. (2022) yang mencakup tahap persiapan data, pengembangan model, dan evaluasi. Terdapat dua modifikasi utama yaitu penambahan tahap khusus pengembangan VA Regressor dan tahap evaluasi yang diperluas dengan subproses *deployment* antarmuka web sebagai keluaran sistem yang dapat digunakan secara interaktif oleh praktisi atau pengguna.

Secara keseluruhan, penelitian ini dilaksanakan melalui empat tahapan utama yang sistematis dan hierarkis sebagaimana diilustrasikan pada Gambar 3.1. Alur penelitian ini disusun dengan mengadopsi dan mengadaptasi sejumlah pendekatan dari penelitian terdahulu, meliputi penggunaan dataset KIDO sebagai sumber data utama (Çiftçi dkk., 2025), arsitektur *Image-Text Interaction Network* (ITIN) sebagai kerangka dasar model yang dimodifikasi (Zhu dkk., 2022), integrasi dimensi *Valence-Arousal* berdasarkan *Russell's Circumplex Model of Affect*

(Boateng dkk., 2022) yang direalisasikan melalui pendekatan regresi berbasis RoBERTa (Liu dkk., 2019; Mendes & Martins, 2023) yang dilatih pada dataset EmoBank (Buechel & Hahn, 2017), serta sejumlah modifikasi tambahan yang terinspirasi dari *weighted logit fusion* (Ye, 2025), *objectness score* (Choi dkk., 2019), strategi pembobotan modalitas visual dominan pada konteks seni gambar (Bose dkk., 2021), dan kerapatan tinta sebagai fitur visual informatif (Nath dkk., 2025). Keluaran dari setiap tahap menjadi masukan langsung bagi tahap berikutnya sehingga seluruh alur membentuk satu kerangka kerja penelitian yang terstruktur.



Gambar 3.1 Diagram Alur Penelitian

Tahap pertama berupa persiapan yang difokuskan pada persiapan aspek konseptual, menyiapkan data, dan merumuskan kebutuhan sistem. Tahap kedua yaitu menempatkan pengembangan VA Regressor sebagai prasyarat sebelum memasuki tahap pengembangan model utama, karena fitur VA yang dihasilkan akan

digunakan sebagai masukan pada tahap ketiga. Tahap ketiga merupakan inti penelitian yang mengembangkan arsitektur ITIN dengan mengintegrasikan seluruh modalitas. Terakhir, tahap keempat meliputi evaluasi menyeluruh dan deployment sistem dalam bentuk antarmuka web yang dapat digunakan secara interaktif.

3.3 Persiapan

3.3.1 Studi Literatur

Tahapan pertama yang dilakukan adalah studi literatur yang bertujuan untuk menggali landasan teoritis, mengidentifikasi kesenjangan penelitian, serta menetapkan arah pengembangan sistem. Studi literatur diarahkan pada tiga kelompok utama berikut:

a. Kelompok 1 mengkaji karakteristik visual yang berkorelasi dengan kondisi emosional subjek serta penggunaan dataset KIDO (Çiftçi dkk., 2025) sebagai satu-satunya korpus publik yang menyediakan pasangan gambar coretan anak dan teks refleksi dengan anotasi emosi biner.
b. Kelompok 2 mengkaji pendekatan fusi modalitas gambar dan teks, dengan fokus pada ITIN (Zhu dkk., 2022) sebagai arsitektur referensi karena mekanisme *cross-modal alignment* dan *adaptive gating* terbukti efektif pada tugas analisis sentimen multimodal.
c. Kelompok 3 menelusuri konsep *Valence-Arousal* melalui *Russell's Circumplex Model of Affect* (Boateng dkk., 2022), dataset EmoBank (Buechel & Hahn, 2017) sebagai korpus anotasi VA standar, dan pendekatan regresi berbasis RoBERTa (Mendes & Martins, 2023).

Luaran dari tahapan ini disampaikan pada Bab 1 dalam bentuk latar belakang dan rumusan masalah, serta pada Bab 2 dalam bentuk dasar teori dan tinjauan pustaka.

3.3.2 Pengumpulan Dataset

Tahapan pengumpulan dataset dilakukan untuk menyiapkan seluruh data yang dibutuhkan oleh sistem, baik untuk pelatihan komponen VA Regressor maupun untuk pelatihan model ITIN utama. Terdapat dua dataset yang digunakan dalam penelitian ini yaitu dataset utama KIDO dan dataset sekunder EmoBank.

3.3.2.1 KIDO

Dataset utama yang digunakan dalam penelitian ini adalah KIDO (*Children's Drawings Dataset*) yang dikembangkan oleh Çiftçi dkk. (2025) dan dapat diakses melalui repositori GitHub (https://github.com/serdarciftci/KIDO). Dataset tersebut dikumpulkan secara sistematis lebih dari enam bulan dari lingkungan pendidikan di kota Şanlıurfa negara Turki yang melibatkan 5.430 siswa sekolah dasar dan menengah berusia 6 hingga 14 tahun dari 27 sekolah perkotaan.

Proses pengumpulan data dilakukan menggunakan media kertas fisik untuk menangkap karakteristik alami goresan tangan anak dan telah mematuhi protokol serta persetujuan etis dari Kementerian Pendidikan Nasional Republik Turki dan dewan etik Universitas Harran. Sebagai korpus multimodal, KIDO menyediakan label afeksi biner yaitu antara *Happy* dan *Sad* yang ditentukan langsung oleh para siswa melalui teks refleksi subjektif supaya menjamin bahwa label emosi tersebut benar-benar menggambarkan niat asli pembuatnya. Berikut merupakan rincian karakteristik datasetnya yang disajikan pada Tabel 3.1.

Tabel 3.1 Karakteristik Dataset KIDO

Parameter	**Nilai**
Nama Dataset	KIDO (Children's Drawings Dataset)
Jumlah Subjek	5.430 siswa
Rentang Usia	6 – 14 tahun
Jumlah Gambar	10.860 gambar
Jumlah Kelas Emosi	2 kelas (Happy dan Sad)
Data Pendukung	Teks refleksi anak terkait alasan gambar merepresentasikan emosi tertentu
Pembagian Data	85% data latih dan 15% data uji

Dataset utama pada penelitian ini berbentuk pasangan modalitas yaitu citra coretan dan teks refleksi yang menjelaskan konteks emosional di balik gambar. Karakter visual coretan cenderung bersifat non-objektif dan abstrak sehingga makna semantik tidak selalu dapat diidentifikasi hanya dari pola piksel saja. Kondisi tersebut menuntut strategi ekstraksi fitur visual yang spesifik terhadap domain coretan anak dimana informasi tekstual berperan sebagai konteks pelengkap yang sering tidak tergambarkan secara langsung pada citra sehingga kombinasi dua modalitas ini menghasilkan representasi emosi yang lebih kaya dan lebih robust terhadap ambiguitas data visual.

Meskipun dikembangkan dalam konteks budaya tertentu, dataset seperti KIDO tetap relevan lintas budaya karena kosakata visual dalam gambar anak bersifat universal. Penelitian lintas budaya menunjukkan bahwa anak-anak dari berbagai latar belakang secara konsisten menggambar kategori objek yang serupa terlepas dari perbedaan budaya asal mereka (Restoy dkk., 2022) dan sejalan dengan pola perkembangan artistik yang berlaku universal termasuk di Indonesia (Sudanayasa dkk., 2026).

3.3.2.2 EmoBank

Sebagai penunjang pelatihan komponen VA Regressor, penelitian juga menggunakan dataset EmoBank yang dikembangkan oleh Buechel dan Hahn (2017). EmoBank merupakan korpus bahasa Inggris yang berisi anotasi kontinu dari *Valence*, *Arousal*, dan *Dominance* yang digunakan sebagai tolak ukur standar untuk tugas regresi emosi dimensional. Karakteristik EmoBank disajikan pada Tabel 3.2.

Tabel 3.2 Karakteristik Dataset EmoBank

Parameter	**Nilai**
Jumlah sampel	10.000 kalimat
Skala *Valence*	1 – 5 (negatif ke positif)
Skala *Arousal*	1 – 5 (tenang ke aktif)
Skala *Dominance*	1 – 5 (submisif ke dominan)
Bahasa	Inggris

Dalam penelitian ini hanya dimensi *Valence* dan *Arousal* yang digunakan sebagai fitur input dan dimensi *Dominance* tidak dimanfaatkan karena *Russell's Circumplex Model of Affect* yang menjadi landasan teori hanya mendefinisikan dua sumbu utama.

3.3.3 Analisis Kebutuhan Sistem

Dari studi literatur dan pengumpulan dataset, teridentifikasi tiga tantangan teknis yang menjadi dasar perancangan sistem. Pertama, Faster RCNN pada ITIN (Zhu dkk., 2022) dilatih pada *natural images* COCO sehingga tidak optimal untuk goresan abstrak anak, solusinya adalah *ink-based bounding box detection* berbasis OpenCV tanpa model deteksi objek berat. Kedua, ITIN hanya mengandalkan label emosi diskret sehingga kehilangan nuansa intensitas afektif, solusinya adalah integrasi VA Regressor berbasis RoBERTa yang dilatih pada EmoBank untuk

menghasilkan nilai *Valence* dan *Arousal* secara otomatis. Ketiga, *single-stage alignment* pada ITIN hanya memberikan satu kesempatan penyelarasan *cross-modal*, solusinya adalah mekanisme *two-stage cross-attention* dengan *residual connection* dan *summary attention* agar penyelarasan dapat disempurnakan secara bertahap.

Selain kebutuhan fungsional, sistem juga memerlukan antarmuka web berbasis FastAPI yang memungkinkan praktisi pendidikan dan kesehatan mental menggunakan hasil prediksi emosi beserta nilai *Valence-Arousal* secara interaktif.

3.4 Pengembangan Model VA

Tahap kedua adalah pengembangan model untuk menentukan nilai *Valence-Arousal* secara otomatis yang disebut dengan model VA Regressor. Model ini merupakan komponen yang bertugas memprediksi nilai *Valence* dan *Arousal* secara lanjutan dari teks refleksi anak. Tahap ini didahulukan sebelum pengembangan arsitektur ITIN utama karena fitur VA yang dihasilkan akan menjadi salah satu masukan pada Tahap 3.

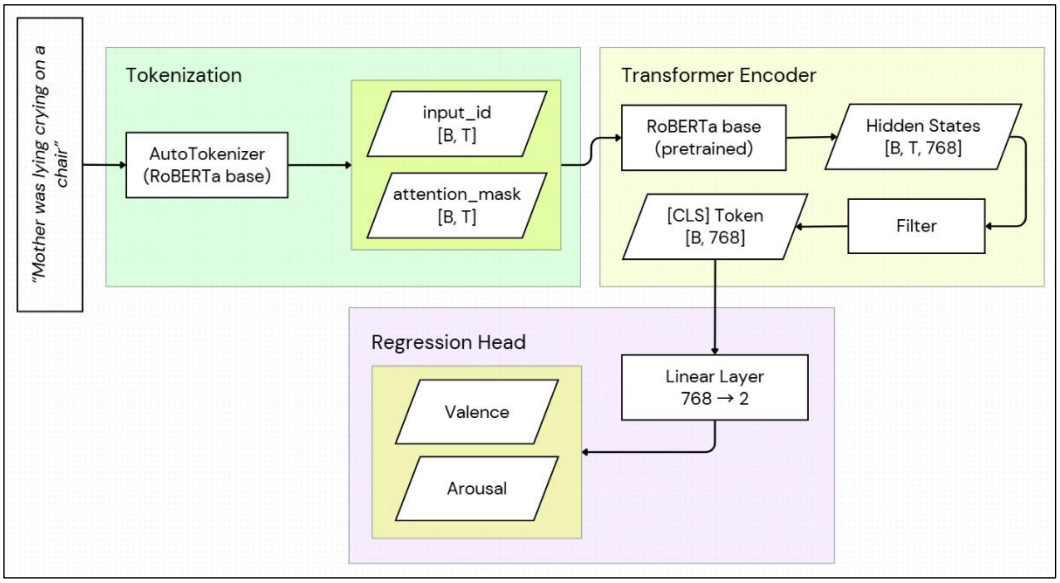
3.4.1 *Preprocessing* Dataset EmoBank

Sebelum digunakan untuk melatih VA Regressor, data teks dari EmoBank akan melewati tahapan pra-pemrosesan untuk menyesuaikan format input dengan arsitektur RoBERTa. Proses ini mencakup pembersihan teks dari karakter tidak relevan, normalisasi ke huruf kecil, tokenisasi menggunakan RoBERTa tokenizer dengan max_length sebesar 128, serta konversi ke tensor berupa input_ids dan attention_mask. Panjang maksimum 128 token dipilih berdasarkan distribusi panjang kalimat pada EmoBank yang sebagian besar tidak melampaui batas tersebut.

Setelah pra-pemrosesan, data dibagi menjadi tiga himpunan sesuai dengan pembagian bawaan dari EmoBank yaitu train, development atau validasi, dan test. Pembagian ini bersifat tetap mengikuti split resmi yang disediakan oleh dataset sehingga hasil eksperimen dapat dibandingkan secara langsung dengan penelitian lain yang menggunakan EmoBank (Buechel & Hahn, 2017).

3.4.2 Pengembangan Model

Model VA Regressor dikembangkan menggunakan arsitektur RoBERTa (Liu dkk., 2019) sebagai encoder teks dengan penambahan *Regression Head* di atasnya. RoBERTa dipilih karena optimasi pelatihannya yang lebih robust dibandingkan BERT standar terutama pada tugas regresi teks afektif. Arsitektur VA Regressor diilustrasikan pada Gambar 3.2.



Gambar 3.2 Arsitektur VA Regressor

Model VA Regressor dikembangkan menggunakan arsitektur RoBERTa (Liu dkk., 2019) sebagai *encoder* teks dengan penambahan *Regression Head* di atasnya, terinspirasi dari pendekatan regresi berbasis model *Transformer* pra-latih untuk memprediksi nilai *Valence* dan *Arousal* secara kontinu dari teks (Mendes &

Martins, 2023). Pada Gambar 3.2 diatas, arsitektur VA Regressor terdiri dari tiga blok utama yaitu Tokenisasi, *Encoder Transformer*, dan *Regression Head*. Proses dimulai dari Teks Input yang merupakan kalimat teks refleksi anak dalam bahasa Inggris misalnya *"I drew this because I feel very happy today"*. Teks mentah ini belum dapat langsung diproses oleh model transformer dan perlu diubah ke dalam format numerik terlebih dahulu.

Tahap selanjutnya adalah AutoTokenizer RoBERTa-base yaitu proses pemecahan teks menjadi unit-unit token menggunakan algoritma *Byte-Pair Encoding* (BPE). *Tokenizer* secara otomatis menambahkan token khusus [CLS] di awal dan [SEP] di akhir sekuens. Keluaran *tokenizer* berupa dua tensor yaitu input_ids yang merepresentasikan indeks numerik setiap token dan attention_mask yang membedakan token nyata dari token *padding*. Panjang maksimum sekuens dibatasi 128 token sesuai konfigurasi pada Tabel 3.3.

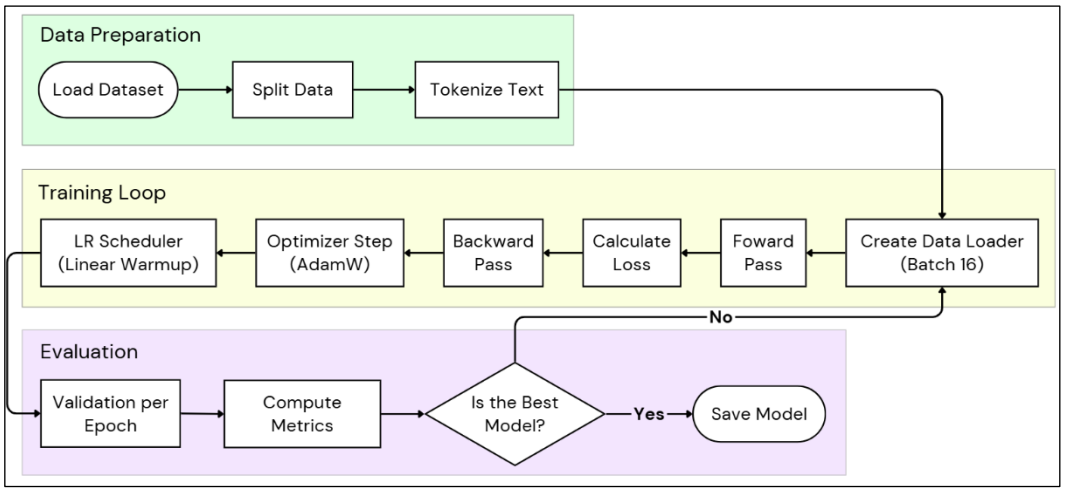
Kedua tensor tersebut kemudian dilemparkan ke RoBERTa-base, model transformer berbasis *encoder-only* dengan 12 lapisan *self-attention* dan dimensi hidden sebesar 768. RoBERTa-base merupakan versi yang dioptimalkan dari BERT dengan strategi pelatihan yang lebih ketat termasuk penghapusan tugas *Next Sentence Prediction* dan pelatihan pada korpus yang lebih besar. Setiap token diproses secara paralel menghasilkan *Hidden States* berdimensi [B, T, 768]. Dari seluruh hidden states tersebut hanya representasi token [CLS] berdimensi [B, 768] yang diambil karena posisinya di awal sekuens memungkinkan *attention* model mengakumulasi informasi dari seluruh token lainnya.

Representasi [CLS] kemudian diproyeksikan melalui *Linear Layer* dengan dimensi input sebesar 768 dan menghasilkan output 2 dimensi pada *Regression*

Head yaitu dua nilai skalar kontinu yaitu *Valence* dan *Arousal* dalam tensor [B, 2] dengan rentang skala 1 hingga 5. VA Regressor hanya akan menghasilkan nilai mentah [B, 2] saja sebagai outputnya.

3.4.3 Pelatihan Model

Pelatihan VA Regressor dilakukan menggunakan data EmoBank yang telah diproses pada tahap sebelumnya. Pada Gambar 3.3 alur pelatihan VA *Regressor* terdiri dari tiga fase utama yaitu persiapan data, training loop, dan evaluasi. Pada fase *Data Preparation*, tahap pertama adalah pembacaan dataset EmoBank dan pembagiannya menjadi tiga himpunan yaitu train, dev (validasi), dan test. Himpunan dev digunakan untuk memantau performa selama pelatihan sedangkan himpunan test digunakan untuk evaluasi akhir setelah pelatihan selesai. Setiap kalimat kemudian ditokenisasi menggunakan AutoTokenizer RoBERTa-base dengan panjang maksimum 128 token.



Gambar 3.3 Alur Pelatihan VA Regressor

Pada fase *Training Loop*, data latih diumpankan melalui DataLoader dalam batch berukuran 16. Setiap iterasi terdiri dari *forward pass* yang menghasilkan

prediksi VA dan penghitungan loss gabungan dibuat menggunakan Persamaan (14) dan Persamaan (15) yang dikombinasikan dengan cara:

$$Loss = \alpha \cdot \rho_c + (1 - \alpha) \cdot MSE \quad \textbf{(18)}$$

Selain *forward pass* ada juga *backward pass* yang berfungsi untuk menghitung gradien dan pembaruan bobot menggunakan optimizer AdamW. Setelah pembaruan bobot, LR Scheduler dengan *linear warmup* menyesuaikan *learning rate* secara bertahap dimulai dari nilai kecil pada awal pelatihan lalu meningkat secara linear hingga mencapai nilai target kemudian menurun kembali. Strategi ini membantu model melakukan adaptasi awal yang lebih stabil sebelum memasuki fase pembelajaran penuh.

Pada fase *Evaluation*, model dievaluasi pada himpunan dev di akhir setiap epoch menggunakan lima metrik secara bersamaan yaitu MSE, RMSE, *Pearson r* untuk mengukur korelasi linear, *Spearman ρ* untuk mengukur korelasi peringkat, dan CCC untuk mengukur kesesuaian distribusi secara keseluruhan. Keputusan penyimpanan checkpoint didasarkan pada nilai CCC validasi terbaik. Strategi *best-checkpoint saving* ini memastikan bahwa model yang tersimpan adalah versi dengan kemampuan generalisasi terbaik bukan sekadar model dari epoch terakhir yang berpotensi mengalami *overfitting*. Checkpoint terbaik inilah yang kemudian digunakan pada tahap 3.4.5 untuk mengekstraksi fitur VA dari seluruh data KIDO.

Tabel 3.3 Hyperparameter Pelatihan VA Regressor

Parameter	**Nilai**
Base Model	RoBERTa-base
Learning Rate	1e-5 atau 0.00001
Batch Size	16
Epochs	10
Loss Function	Persamaan (18)

CCC Weight	0.5
Max Sequence Length	128
Optimizer	AdamW
Weight Decay	0.01
LR Scheduler	*Linear warmup*
Seed	42

Beberapa keputusan desain kunci pada Tabel 3.3 perlu digarisbawahi. *Learning Rate* 1e-5 dipilih lebih konservatif dari nilai umum *fine-tuning* transformer agar bobot *pretrained* RoBERTa tidak terdegradasi dan konvergensi berlangsung stabil. *Loss Function* Persamaan (17) dengan *CCC Weight* 0.5 menyeimbangkan ketepatan nilai absolut (MSE) dan kesesuaian distribusi prediksi (CCC) secara proporsional. Optimizer AdamW dengan *Weight Decay* 0.01 dan *LR Scheduler linear warmup* dipilih untuk stabilitas fase adaptasi awal. *Normalize Labels* dinonaktifkan untuk mempertahankan rentang skala *Circumplex* asli agar sinyal VA tidak terkompresi saat disalurkan ke ITIN. *Seed* 42 memastikan *reproducibility* eksperimen.

3.4.4 Evaluasi Model

Setelah pelatihan selesai, performa VA Regressor dievaluasi menggunakan himpunan dev dan test dari EmoBank. Lima metrik digunakan secara bersamaan untuk memberikan gambaran performa yang lebih lengkap. *MSE* dan *RMSE* mengukur besarnya error prediksi secara absolut di mana nilai yang lebih rendah menunjukkan prediksi yang lebih akurat. *Pearson r* mengukur kekuatan korelasi linear antara nilai prediksi dan nilai target, sedangkan *Spearman* ρ mengukur korelasi berbasis peringkat yang lebih robust terhadap outlier. *CCC* mengukur kesesuaian distribusi secara menyeluruh dengan mempertimbangkan korelasi, perbedaan rata-rata, dan perbedaan variansi secara

bersamaan. *Checkpoint* dengan nilai MSE total tertinggi disimpan sebagai model terbaik untuk digunakan pada tahap selanjutnya.

3.4.5 Ekstraksi Fitur

Setelah model VA *Regressor* terbaik diperoleh, tahap berikutnya adalah menggunakannya untuk mengekstraksi nilai VA dari seluruh teks refleksi pada dataset KIDO. Proses ini dijalankan sekali untuk setiap split data (latih dan uji) dan hasilnya disimpan dalam format .csv untuk efisiensi *loading* pada Tahap 3. Dengan demikian, komputasi inferensi VA Regressor tidak perlu diulang pada setiap *epoch* pelatihan ITIN.

3.5 Pengembangan Arsitektur ITIN

Tahap ketiga merupakan inti penelitian yaitu pengembangan arsitektur ITIN yang mengintegrasikan fitur visual, fitur tekstual, dan fitur VA menjadi satu sistem klasifikasi emosi yang terpadu. Tahap ini dibagi menjadi *preprocessing* data KIDO, perancangan arsitektur sistem, dan pelatihan model.

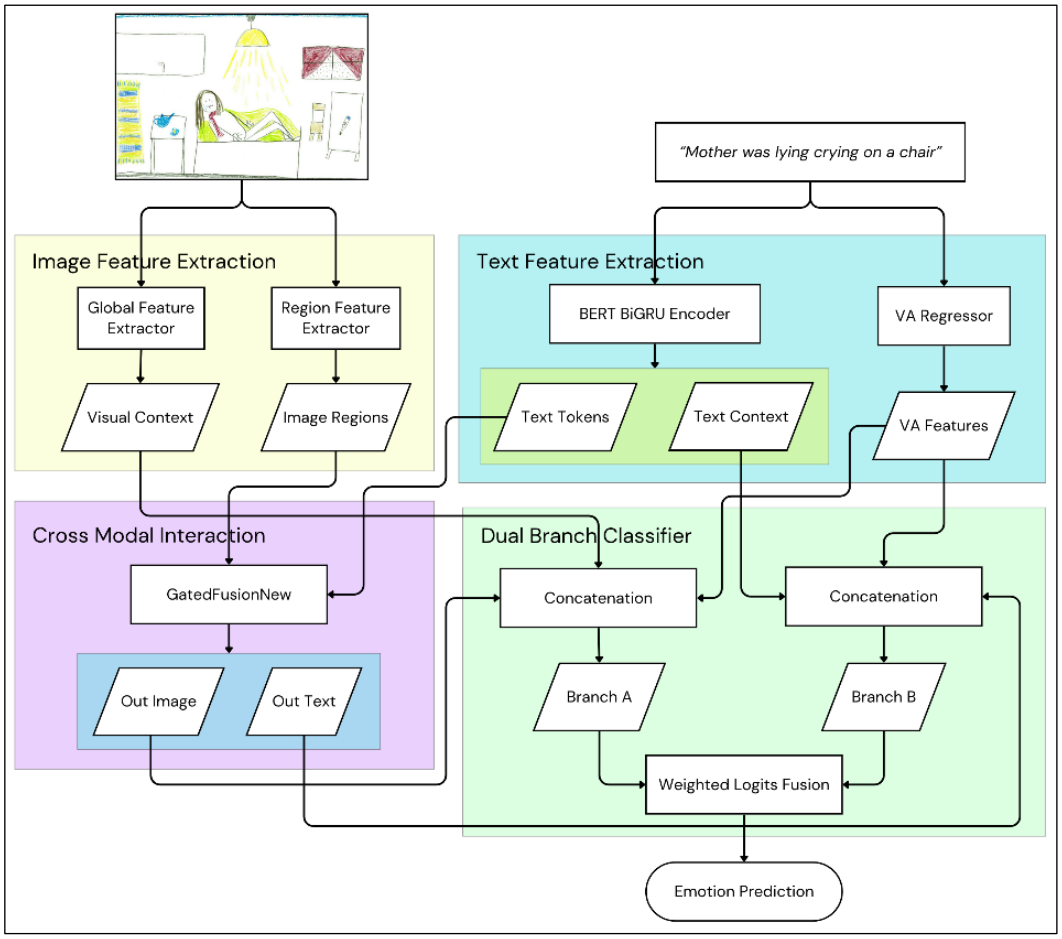
3.5.1 *Preprocessing* Dataset KIDO

Sebelum data KIDO dimasukkan ke dalam pipeline model, seluruh sampel melewati pra-pemrosesan yang berbeda untuk setiap modalitas. Jalur gambar menjalankan pipeline *ink-based detection* berbasis OpenCV yang menghasilkan region fitur tersimpan dalam format .npz, sedangkan jalur teks menjalankan tokenisasi BERT secara *inline* selama pelatihan. Penjelasan lengkap masing-masing pipeline diuraikan pada bagian Modul Ekstraksi Fitur Visual (Gambar 3.5) dan Modul Ekstraksi Fitur Teks (Gambar 3.6).

3.5.2 Pengembangan Arsitektur

3.5.2.1 Arsitektur Keseluruhan

Pengembangan model yang dibangun berdasarkan arsitektur ITIN karya Zhu dkk. (2022) yang telah dimodifikasi secara mendalam untuk mengakomodasi fitur *Valence-Arousal* dan mekanisme *gating* dua tahap. Modifikasi ini bertujuan untuk menghasilkan penyelarasan (*alignment*) fitur gambar dan teks yang lebih presisi. Proses kerangka kerja arsitektur secara menyeluruh dapat dilihat pada Gambar 3.4.



Gambar 3.4 Arsitektur ITIN dengan Integrasi VA

Hasil akhir dari model yaitu prediksi kelas emosi (*Happy* atau *Sad*) beserta skor probabilitasnya. Selain itu, nilai *Valence* dan *Arousal* yang dihasilkan oleh VA

Regressor juga diterjemahkan menjadi label emosi deskriptif agar lebih mudah diinterpretasikan oleh pengguna. Pemetaan nilai VA ke label emosi disajikan pada Tabel 3.4.

Tabel 3.4 Pemetaan Nilai *Valence-Arousal* ke Label Emosi Deskriptif

Valence	***Arousal***	**Label Emosi**
High	High	Excited
High	Mid	Happy
High	Low	Relaxed
Mid	High	Angry
Mid	Low	Calm
Low	High	Tense
Low	Mid	Sad
Low	Low	Depressed

Pada tabel tersebut nilai *Valence* dan *Arousal* masing-masing dikategorikan ke dalam tiga level yaitu *High*, *Mid*, dan *Low*. Proses pemetaan nilai dari skala regresi kontinu dengan skala 1–5 ke dalam tiga kategori diskret tersebut menggunakan transformasi *Z-Score* dengan setiap nilai prediksi (x) dikalibrasi terhadap nilai rata-rata (μ) dan standar deviasi (σ) dari keseluruhan himpunan data sebagaimana dinyatakan pada Persamaan (19) (Wongoutong, 2024).

$$Z = \frac{x - \mu}{\sigma} \qquad \textbf{(19)}$$

Persamaan (19) merupakan transformasi yang memastikan bahwa penentuan tinggi-rendahnya skor didasarkan pada sebaran nyata populasi anak secara spesifik. Selanjutnya, nilai dari hasil transformasi (Z) dikelompokan di mana nilai $Z \geq 0.5$ ditetapkan sebagai *High*, nilai $Z \leq -0.5$ sebagai *Low*, dan sisa rentang di antaranya yaitu $-0.5 < Z < 0.5$ sebagai kelompok *Mid*. Apabila parameter statistik tidak mencukupi untuk penghitungan *Z-Score*, sistem dirancang

untuk secara otomatis beralih menggunakan metode pembagian kuartil distribusi pada persentil ke-33 dan ke-66 sebagai cadangan (*fallback*).

Kombinasi kedua level sumbu tersebut kemudian disilangkan untuk menghasilkan delapan label emosi deskriptif yang mengacu pada *Russell's Circumplex Model of Affect* (Boateng dkk., 2022).

3.5.2.2 Justifikasi Modifikasi terhadap ITIN Referensi

Arsitektur yang diimplementasikan merupakan adaptasi dari ITIN (Zhu dkk., 2022) dengan sejumlah modifikasi yang disesuaikan terhadap karakteristik domain coretan anak dan tujuan penelitian ini. Ringkasan perbedaan antara ITIN referensi dan implementasi penelitian ini disajikan pada Tabel 3.5.

Tabel 3.5 Perbandingan Arsitektur ITIN Referensi dan Implementasi

Aspek	ITIN Zhu dkk. (2022)	Implementasi Penelitian	Alasan Modifikasi
Region Detection	Faster RCNN (ResNet101, pretrained COCO)	*Ink-based Detection* (OpenCV)	Faster RCNN kurang cocok untuk goresan abstrak anak, ink-based lebih domain-spesifik.
Region Backbone	ResNet101 via Faster RCNN	ResNet50 standalone	Lebih efisien secara komputasi dengan representasi yang tetap memadai.
Cross-Modal Fusion	Single-stage attention	*Two-stage QKV attention*	Tahap *refinement* kedua meningkatkan presisi *cross-modal aligment*.
Fused Output	Satu vektor C	Dua vektor: out_img dan out_txt	Representasi terpisah mempertahankan kontribusi per-modalitas hingga level logits.
Final Fusion	Feature-level: $\lambda \cdot F1 + (1-\lambda) \cdot F2$	Persamaan (12)	*Late fusion* lebih fleksibel dan robust terhadap dominansi salah satu modalitas.

Valence-Arousal	Tidak ada	VA features dari RoBERTa	Novelty utama dan sinyal afektif kontinu yang tidak tertangkap fitur lain.
Sequence Reduction	*Average pooling*	SummaryAttn (*learned pooling*)	*Attention pooling* adaptif mempelajari bobot kepentingan setiap token.
Residual Connection	Tidak digunakan	*Residual* pada fusion *Multilayer Perceptron*	Menjaga aliran gradien dan stabilitas training.

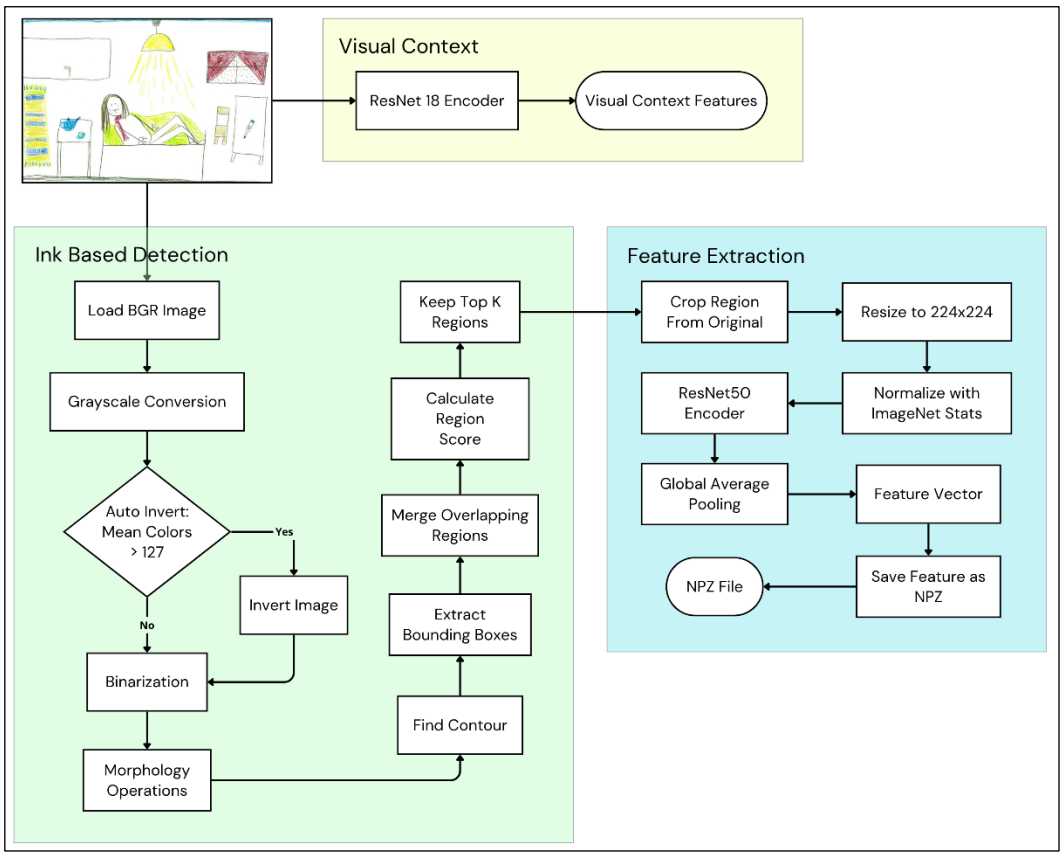
Perubahan arsitektur tersebut sudah dijelaskan sebagian besar di tahap awal yaitu analisis kebutuhan sistem. Intinya, modifikasi terpenting yang sekaligus menjadi novelty utama pada penelitian ini adalah integrasi fitur *Valence-Arousal* dengan menambahkan nilai VA tersebut dari RoBERTa regressor ke dalam *Dual Branch Classifier* sehingga model dapat memperoleh sinyal emosional secara langsung yang melengkapi representasi multimodal secara menyeluruh.

3.5.2.3 Modul Ekstraksi Fitur Visual

Ekstraksi fitur visual dilakukan melalui dua jalur paralel sebagaimana diilustrasikan pada Gambar 3.5. Jalur pertama menggunakan pendekatan *ink-based detection* dengan *backbone* ResNet50 untuk menghasilkan representasi lokal (*image regions*) sementara jalur kedua menggunakan ResNet18 secara langsung untuk mengekstraksi konteks visual global dari keseluruhan gambar. Pemisahan ini bertujuan agar model dapat menangkap detail goresan tinta sekaligus memahami suasana gambar secara menyeluruh.

Jalur pertama yaitu visual context yang memproses gambar coretan secara utuh tanpa segmentasi region. Gambar diubah menjadi 224×224 piksel dan dinormalisasi menggunakan mean dan standar deviasi ImageNet, kemudian dilemparkan ke ResNet18 yang secara langsung menghasilkan representasi visual

context yang berdimensi [B, 512]. Fitur global ini merangkum karakteristik keseluruhan gambar seperti komposisi umum, distribusi goresan, dan proporsi ruang kosong yang berfungsi sebagai konteks pelengkap pada cabang klasifikasi visual di Branch A (lihat Gambar 3.4).



Gambar 3.5 Kerangka Kerja Ekstraksi Fitur Visual

Selanjutnya pada jalur kedua, gambar coretan dibaca menggunakan cv2.imread() dalam format BGR, kemudian dikonversi ke *grayscale*. Langkah berikutnya adalah *Auto Invert*, dimana sistem secara otomatis memeriksa apakah rata-rata piksel gambar melebihi 127. Jika ya, berarti gambar memiliki latar belakang terang dengan goresan gelap sehingga gambar diinvert terlebih dahulu agar tinta menjadi piksel terang di atas latar gelap yang memudahkan proses binarisasi. Setelah binarisasi menggunakan metode *Otsu* atau *Adaptive*, dua operasi

morfologi diterapkan secara berurutan yaitu *Closing* untuk menutup celah kecil yang mungkin ada di dalam goresan yang terputus-putus, dan *Opening* untuk menghilangkan *noise* piksel kecil yang terisolasi.

Kemudian, kontur luar (*external contours*) diidentifikasi menggunakan cv2.findContours dengan *flag* RETR_EXTERNAL. Dari setiap kontur *bounding box* diekstraksi dan difilter berdasarkan min_area untuk menyingkirkan region yang terlalu kecil. *Bounding box* yang saling tumpang tindih kemudian digabungkan menggunakan *threshold Intersection over Union* atau IoU ≥ 0.15 yang dinyatakan dalam Persamaan (20) (Rezatofighi dkk., 2019).

$$\text{IoU}(A, B) = \frac{|\ A \cap B\ |}{|\ A \cup B\ |} = \frac{\text{luas irisan } A \cap B}{\text{luas } A + \text{luas } B - \text{luas irisan}} \qquad \textbf{(20)}$$

Setiap region yang tersisa diberi skor menggunakaan Persamaan (21), yang strukturnya terinspirasi dari pendekatan *objectness score* berupa kombinasi linear berbobot antara dua komponen skor ternormalisasi yang umum digunakan dalam tugas deteksi region pada citra (Choi dkk., 2019).

$$S_r = \alpha \cdot R_{ink} + (1 - \alpha) \cdot min(1.0,\ R_{area} \times \beta) \qquad \textbf{(21)}$$

Di mana S_r adalah skor akhir region, R_{ink} adalah rasio piksel tinta dalam *bounding box*, R_{area} adalah rasio luas region terhadap total gambar, α adalah bobot *ink ratio*, dan β adalah faktor normalisasi. Komponen pertama $\alpha \cdot R_{ink}$ mengukur kerapatan tinta pada setiap region, sedangkan komponen kedua $(1 - \alpha) \cdot \min(1{,}0,\ R_{area} \times \beta)$ memperhitungkan ukuran relatif region terhadap keseluruhan gambar yang dinormalisasi dengan faktor β dan dibatasi maksimum 1,0.

Region dengan skor tertinggi dipilih sebagai Top-K sebesar 36 region yang akan diekstraksi fiturnya sebagai masukan *cross-modal attention* yang menjamin

model berfokus pada area goresan yang paling informatif. Setiap region dilakukan *crop* dari gambar asli, dilakukan *resize* ke 224×224, dan dinormalisasi menggunakan mean dan standar deviasi ImageNet, kemudian diproses oleh ResNet50 dan *Global Average Pooling* untuk menghasilkan vektor fitur [K, 2048]. Seluruh hasil disimpan dalam format .npz yang memuat lima jenis data yaitu *features* [K, 2048], *boxes* [K, 4], *scores* [K], image_h, dan image_w.

Tabel 3.6 Parameter *Ink-Based Detection*

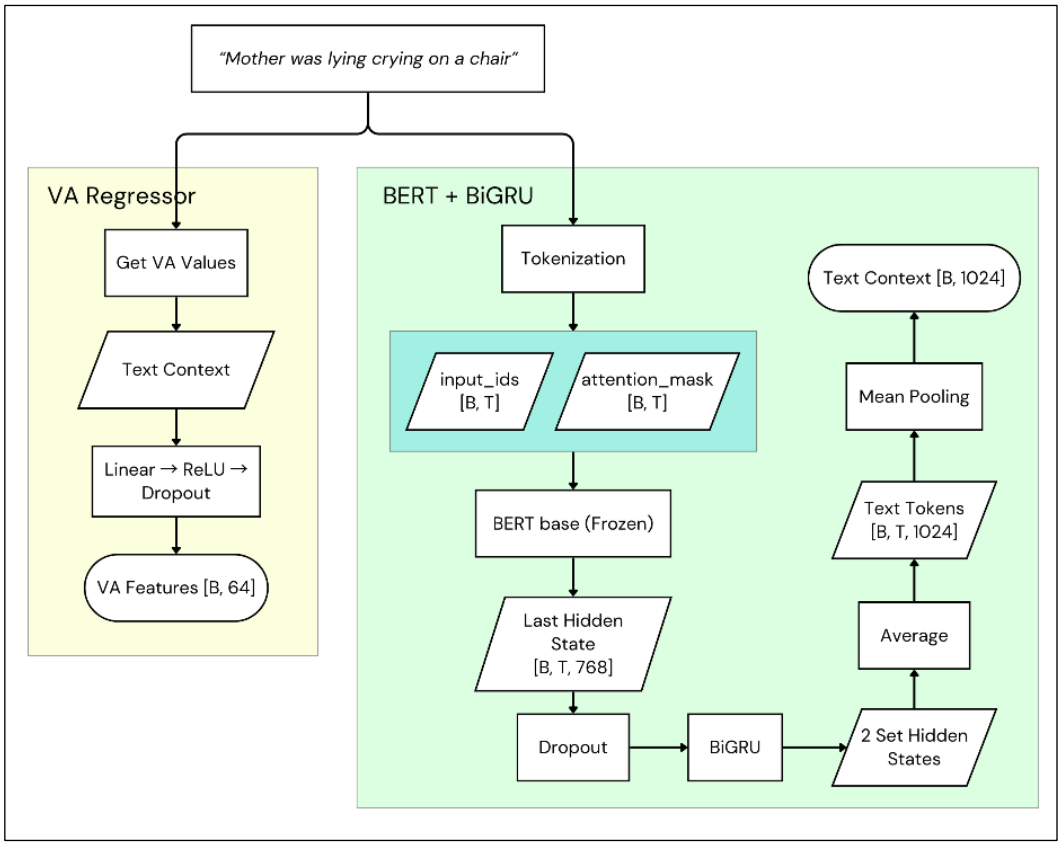
Parameter	Nilai	Deskripsi
k	36	Jumlah region per gambar (Top-K)
backbone	resnet50	*Backbone*
invert	auto	Mode *invert*: auto (cek mean>127 otomatis) / yes / no
binarize	otsu	Metode binarisasi: *Otsu* atau *Adaptive*
dilate	7	Ukuran kernel morfologi
close_iter	2	Jumlah iterasi operasi *Closing* (menutup celah pada goresan)
open_iter	0	Jumlah iterasi operasi Opening (menghilangkan noise)
min_area	300	Area minimum *bounding box* (piksel)
merge_iou	0.25	*Threshold* IoU untuk penggabungan bounding box

Nilai pada Tabel 3.6, dipilih berdasarkan karakteristik domain coretan anak. close_iter sebesar 2 memastikan celah goresan yang terputus-putus tertutup sepenuhnya, sementara open_iter bernilai 0 tidak diperlukan karena noise kecil sudah tersaring oleh min_area bernilai 300. Parameter invert dengan nilai "auto" mengeliminasi kebutuhan konfigurasi manual per gambar, dan merge_iou sebesar 0.15 mencegah duplikasi ekstraksi fitur pada area yang sama.

3.5.2.4 Modul Ekstraksi Fitur Teks

Encoding teks menggunakan arsitektur BERT dari penelitian Devlin dkk. (2019) yang dikombinasikan dengan BiGRU untuk menghasilkan representasi pada

dua level yaitu level token untuk proses *cross-modal interaction* dan level kalimat untuk konteks global. Arsitektur modul teks diilustrasikan pada Gambar 3.6.



Gambar 3.6 Kerangka Kerja Ekstraksi Fitur Teks

Terlihat pada Gambar 3.6, modul ekstraksi fitur teks terdiri dari dua cabang yang berjalan secara paralel dan keduanya menerima input berupa teks refleksi yang sama sebagai masukan.

Cabang pertama merupakan jalur VA Regressor yang menerima *VA Values* berdimensi [B, 2] berupa nilai *Valence* dan *Arousal* yang telah diproses sebelumnya menggunakan model VA Regressor berbasis RoBERTa pada Tahap 3.4. Nilai dua dimensi ini kemudian diproyeksikan melalui rangkaian *Linear Layer*, ReLU, dan *Dropout* untuk mengekspansi dimensinya dari 2 menjadi 64 dan menghasilkan *VA Features* berdimensi [B, 64]. Ekspansi dimensi ini diperlukan

agar fitur VA memiliki kapasitas representasi yang proporsional sebelum digabungkan dengan fitur modalitas lain pada *Dual-Branch Classifier*.

Cabang kedua merupakan jalur BERT dan BiGRU di mana teks refleksi terlebih dahulu diubah ke representasi numerik melalui *Tokenizer* yang menghasilkan input_ids dan attention_mask berdimensi [B, T]. Kedua tensor ini diumpankan ke BERT-base yang bobotnya dibekukan (*frozen*) selama pelatihan ITIN agar representasi *pretrained* tidak berubah. Output BERT berupa last_hidden_state berdimensi [B, T, 768] kemudian dilewatkan melalui lapisan *Dropout* sebagai regularisasi sebelum diteruskan ke BiGRU dengan hidden_size bernilai 1024. BiGRU menghasilkan dua *set hidden states* arah maju (*forward*) dan arah mundur (*backward*) yang kemudian dirata-ratakan elemen per elemen menjadi representasi berdimensi [B, T, 1024]. Dari representasi ini dihasilkan dua output secara paralel yaitu *Text Tokens* berdimensi [B, T, 1024] yang mempertahankan representasi per token untuk digunakan pada modul GatedFusionNew dalam proses alignment lintas modal, serta *Text Context* berdimensi [B, 1024] yang diperoleh melalui *mean pooling* pada dimensi waktu T sebagai representasi global kalimat yang digunakan pada Branch B *Dual-Branch Classifier*.

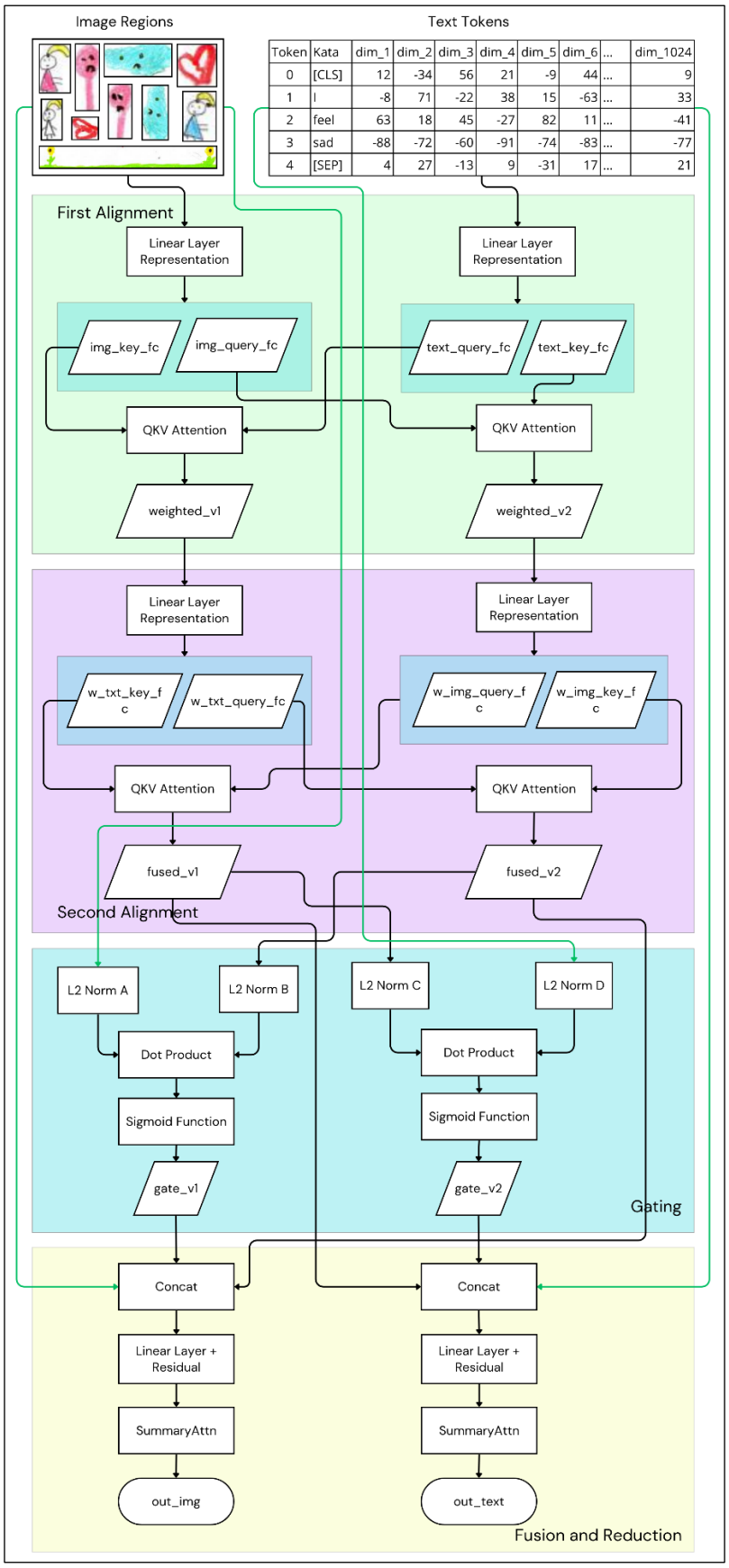
3.5.2.5 Modul *Cross-Modal Interaction* dan *Gated Fusion*

Modul ini merupakan implementasi dari mekanisme *cross-modal interaction*, proses di mana representasi dari dua modalitas yang berbeda yaitu region visual dan token teks saling bertukar informasi melalui mekanisme *attention* sehingga setiap modalitas diperkaya dengan konteks dari modalitas lainnya. Implementasi dilakukan dengan menggunakan modul GatedFusionNew yang

menggabungkan dua komponen utama yaitu *cross-modal alignment* berbasis *QKV attention* dengan dua tahap untuk menghitung korespondensi antara setiap region gambar dan setiap kata dalam teks serta *adaptive cross-modal gating* untuk menyaring pasangan region dan kata yang tidak selaras dan memperkuat interaksi yang relevan secara emosional.

Bisa dilihat pada Gambar 3.7 dimana GatedFusionNew terdiri dari empat stage yang berurutan dengan alur utama ditunjukan pada garis hitam dan garis hijau sebagai *get context data*. Pada Stage 1 atau *First Alignment*, setiap modalitas memiliki pasangan linear layer terpisah untuk menghasilkan representasi query dan key dari img_query_fc dan img_key_fc untuk modalitas visual, serta txt_query_fc dan txt_key_fc untuk modalitas tekstual. Dua operasi *QKV Attention* bidireksional kemudian dijalankan, yang pertama menggunakan *Query* dari teks dan *Key-Value* dari gambar untuk menghasilkan weighted_v1 (*word to region alignment*), sedangkan yang kedua menggunakan Query dari gambar dan *Key-Value* dari teks untuk menghasilkan weighted_v2 (*region* to *word alignment*). Formula Persamaan *attention* yang digunakan adalah *scaled dot-product attention* dari Persamaan (9).

Pada Stage 2 atau *Second Alignment*, hasil dari *weighted* dari Stage pertama menjadi input untuk tahap *refinement* dengan *set linear layer* baru yaitu w_img_query_fc, w_img_key_fc, w_txt_query_fc, dan w_txt_key_fc. Setelah itu, dua *QKV Attention* kembali dijalankan untuk menghasilkan fused_v1 dan fused_v2 yang merupakan representasi hasil penyempurnaan alignment tahap pertama.



Gambar 3.7 Mekanisme *GatedFusionNew* (*Two-Stage*)

Lanjut pada Stage 3 yaitu *Gating*, dimana nilai gate dihitung berdasarkan *dot product* antara representasi yang telah melalui *L2 Normalization*. Keempat tensor yaitu fused_v1, fused_v2, V1 sebagai *image region*s, dan V2 sebagai *text tokens* (dihasilkan dari Gambar 3.6) dimana semuanya dilakukan L2-normalize terlebih dahulu sebelum *dot product* dilakukan. Setelah itu gate_v1 dihitung dari *dot product* antara L2-norm (V1) dan L2-norm (fused_v2) kemudian dijalankan fungsi sigmoid dan begitupun pada gate_v2 dari *dot product* antara L2-norm (V2) dan L2-norm (fused_v1). L2 normalization sebelum *dot product* memastikan perhitungan skor keselarasan tidak dipengaruhi oleh besaran magnitudo vektor.

Operasi persamaan nilai gate terinspirasi dari mekanisme *adaptive cross-modal gating* pada arsitektur ITIN dari Zhu dkk. (2022) sebagaimana dinyatakan pada Persamaan (22).

$$g_{v1} = \sigma\left(\sum_{d=1}^{D} \hat{v}_{1,d} \cdot \widehat{fused}_{v2,d}\right) \qquad \textbf{(22)}$$

Di mana $\hat{\cdot}$ menandakan vektor yang telah di-L2-normalisasi. Nilai $g_{v1} \in [0,1]$ mengukur seberapa selaras region r dengan informasi teks yang relevan, nilai mendekati 1 berarti region tersebut sangat relevan secara *cross-modal*, sedangkan nilai mendekati 0 berarti pasangan tidak selaras dan pengaruhnya akan ditekan.

Terakhir yaitu Stage 4 atau *Fusion Reduction*, dimana pada jalur *image* terdapat tiga tensor digabungkan yaitu V1, fused_v2, gate_v1 yang secara berurutan merupakan fitur image asli, hasil fusi dari sisi teks, dan nilai gate sebagai bobot keselarasan. Hasil *concatenation* diproses oleh *linear layer* dengan *residual connection* lalu dirangkum menjadi satu vektor out_img [B, D] menggunakan SummaryAttn yang secara adaptif mempelajari bobot relevan

pada setiap token. Proses serupa berlaku untuk jalur teks menggunakan V2, fused_v1, gate_v2 dan menghasilkan vektor out_txt [B, D].

Residual connection pada Stage 4 mencegah degradasi gradien pada arsitektur yang dalam sebagaimana dinyatakan pada Persamaan (23).

$$\mathrm{co}_{v1} \leftarrow \mathrm{W}_{fuse\_img}\, \mathrm{co}_{v1} + \mathrm{v}_1 \quad \textbf{(23)}$$

$$\mathrm{w} = \mathrm{softmax}\,(\mathrm{MLP}(\mathrm{co}_{v1})) \in \mathbb{R}^R \quad \textbf{(24)}$$

$$\mathrm{out}_{img} = \sum_{r=1}^{R} w_r \cdot \mathrm{co}_{v1}[r] \in \mathbb{R}^D \quad \textbf{(25)}$$

SummaryAttn menghasilkan vektor ringkasan melalui *learned attention pooling*, mengadopsi prinsip *attention-based weighted sum* dari penelitian Bahdanau dkk. (2016) sebagaimana dinyatakan pada Persamaan (24) dan Persamaan (25), di mana MLP terdiri dari Linear$(D \rightarrow D)$ → ReLU → Dropout → Linear$(D \rightarrow 1)$.

3.5.2.6 *Dual-Branch Classifier*

Tabel 3.7 Dimensi Input *Dual-Branch Classifier*

Branch	Komponen Input	Dimensi	Total
Image Branch (A)	*Fused Image* (out_img)	1024	1600
	Visual Context (ResNet18)	512	
	VA Features	64	
Text Branch (B)	*Fused Text* (out_txt)	1024	2112
	Text Context (mean pool)	1024	
	VA Features	64	

Klasifikasi akhir menggunakan arsitektur dual-branch di mana setiap cabang menerima fitur fusi modalitasnya masing-masing yang diperkaya dengan fitur VA dan konteks modalitas tunggal. Komposisi dimensi input setiap cabang disajikan pada Tabel 3.7.

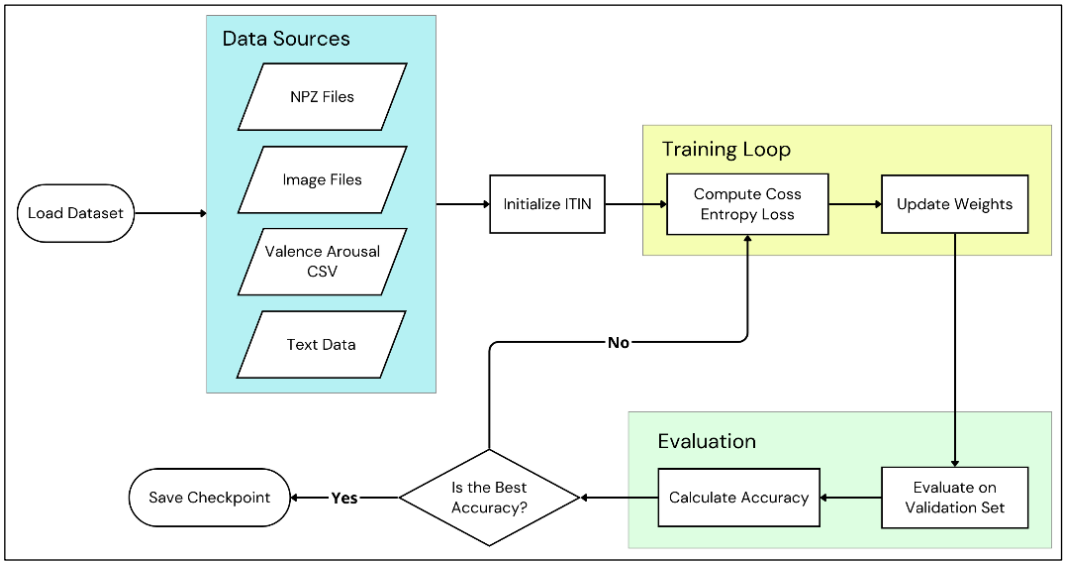
Arsitektur *dual-branch* ini merupakan modifikasi orisinal penelitian. Perbedaan total dimensi antara Branch A (1600) dan Branch B (2112) mencerminkan kapasitas representasi semantik teks yang lebih besar dari visual tepatnya *Text Context* BiGRU sebesar 1024 dimensi dengan *Visual Context* ResNet18 sebesar 512 dimensi. Injeksi *VA Features* sebesar 64 dimensi ke kedua branch secara bersamaan adalah *novelty* utama yang tidak terdapat pada ITIN bawaan, memastikan sinyal afektif kontinu memengaruhi keputusan klasifikasi dari sisi visual maupun tekstual.

Output akhir klasifikasi dihitung dengan *weighted logit fusion* sesuai pada Persamaan (12) dengan perhitungan $logits_{final} = 0.7 \cdot logits_A + 0.3 \cdot logits_B$. Bobot 0,7 pada *branch* visual dan 0,3 pada *branch* teks ditetapkan dengan dua pertimbangan. Pertama, bobot yang lebih besar pada *branch* visual mencerminkan prioritas pada informasi visual, mengingat gambar coretan merupakan modalitas utama ekspresi emosi anak, selaras dengan temuan pada penelitian klasifikasi emosi multimodal berbasis *weighted late fusion* yang juga menetapkan bobot *stream* visual jauh lebih besar dibandingkan *stream* tekstual (w_visual = 0,76 dan w_teks = 0,24) (Bose dkk., 2021).

Kedua, penetapan bobot ini juga berfungsi sebagai penyeimbang terhadap perbedaan kapasitas dimensi input antar *branch* pada Tabel 3.7, di mana Branch B (teks) memiliki total dimensi input yang lebih besar dibandingkan Branch A sebagai visual sehingga tanpa penyesuaian bobot, kontribusi *branch* teks berpotensi mendominasi keputusan akhir secara tidak proporsional.

3.5.3 Pelatihan Model

Setelah arsitektur sistem dirancang dan fitur-fitur telah diekstraksi, tahap berikutnya adalah pelatihan model ITIN secara multimodal. Alur pelatihan diilustrasikan pada Gambar 3.8.



Gambar 3.8 Alur Pelatihan ITIN *Classifier*

Pada fase ini, fitur region (.npz), fitur VA (.csv), dan fitur teks diproses secara terpadu melalui *DataLoader* agar model mempelajari relasi komplementer antar modalitas. Proses ini merupakan inti eksperimen utama dalam penelitian ini.

Pada setiap iterasi pelatihan, data batch diumpankan ke model melalui *forward pass* yang secara berurutan menjalankan ekstraksi fitur, *cross-modal fusion*, dan klasifikasi *dual-branch* hingga menghasilkan logits gabungan menggunakan Persamaan (12). Logits gabungan tersebut kemudian dihitung loss-nya menggunakan *Cross-Entropy Loss* sebagaimana didefinisikan pada Persamaan (13) terhadap label emosi *ground truth*. Setelah *loss* dihitung, *backward pass* menghitung gradien terhadap seluruh parameter yang dapat dilatih kemudian optimizer AdamW memperbarui bobot model berdasarkan gradien tersebut.

Tabel 3.8 *Hyperparameter* Pelatihan ITIN

Parameter	**Nilai**
Learning Rate	1e-4 atau 0,0001
Batch Size	16
Epochs	29
Weight Decay	1e-5 atau 0,00001
Dropout	0,2
Validation Ratio	0,15 atau 15% dari jumlah data *training*
Loss Function	*Cross-Entropy Loss (Categorical Cross-Entropy)*
Image Dimention	2048
Embed Size	1024
Hidden Dimention	1024
Optimizer	AdamW (Loshchilov & Hutter, 2019)
Scheduler	StepLR (gamma=0,1 dan step=10)
Seed	42

Konfigurasi *hyperparameter* pelatihan ITIN disajikan pada Tabel 3.8. *Learning Rate* 1e-4 lebih besar dari VA Regressor (1e-5) karena lapisan ITIN diinisialisasi acak dan membutuhkan pembaruan bobot yang lebih agresif. freeze_bert bernilai "True" membekukan encoder BERT untuk mencegah *catastrophic forgetting* sekaligus mengurangi parameter yang dioptimasi. *Dropout* 0,2 dipilih lebih tinggi dari nilai default karena ukuran dataset KIDO yang terbatas memerlukan regularisasi lebih ketat, berpasangan dengan *Weight Decay* 1e-5 sebagai mekanisme regularisasi ganda. *Scheduler* StepLR dengan gamma sebesar 0.1 dan step sebesar 10 menurunkan *learning rate* sepuluh kali pada epoch ke-10 dan ke-20 untuk transisi dari konvergensi awal ke penyesuaian halus.

3.6 Evaluasi dan Deployment

3.6.1 Pengujian Model

Pengujian dilakukan menggunakan himpunan data uji sebanyak 1.632 sampel yang telah ditetapkan oleh penyedia dataset KIDO dan tidak pernah terlihat selama pelatihan maupun pemilihan *checkpoint*. Evaluasi performa model secara kuantitatif menggunakan rangkaian metrik sebagaimana dirangkum pada Tabel 3.9.

Tabel 3.9 Metrik Evaluasi

Metrik	**Target Komponen**	**Keterangan**
Akurasi	Classifier ITIN	Proporsi prediksi benar terhadap seluruh sampel
Presisi	Classifier ITIN	Ketepatan prediksi kelas positif
Recall	Classifier ITIN	Kemampuan mendeteksi seluruh sampel kelas positif
F1-Score	Classifier ITIN	Harmonik mean presisi dan recall
CCC	VA Regressor	Keselarasan distribusi prediksi VA dengan *ground truth*
MSE	VA Regressor	Rata-rata kuadrat error prediksi VA
ECE	Classifier ITIN	Kalibrasi kepercayaan probabilitas keluaran model

Untuk memvalidasi kontribusi setiap komponen modifikasi secara independen, pengujian dilakukan melalui *ablation study* dengan enam skenario yang disajikan pada Tabel 3.10.

Tabel 3.10 Skenario *Ablation Study*

Komponen	**S1**	**S2**	**S3**	**S4**	**S5**	**S6**
Alignment	*Single-stage*	*Single-stage*	***Two-stage QKV***	*Single-stage*	*Single-stage*	***Two-stage QKV***
Residual connection	Tidak	Tidak	Tidak	Tidak	**Ya**	**Ya**
Sequence reduction	*Avg pooling*	*Avg pooling*	*Avg pooling*	*Avg pooling*	***Summary Attn***	***Summary Attn***

Fusion	*Feature-level* (λF)	*Feature-level* (λF)	*Feature-level* (λF)	***Dual Branch***	*Feature-level* (λF)	***Dual Branch***
VA *features* (64-d)	Tidak	**Ya**	Tidak	Tidak	Tidak	**Ya**
Region source	*Faster* R-CNN	***Ink-based***	***Ink-based***	***Ink-based***	***Ink-based***	***Ink-based***
Optimizer	Adam	Adam	Adam	Adam	Adam	**AdamW**
Jumlah region	m=2	**k=36**	**k=36**	**k=36**	**k=36**	**k=36**
Cross-modal gating	Sigmoid elem-wise	Sigmoid elem-wise	**L2-norm cosine**	Sigmoid elem-wise	Sigmoid elem-wise	**L2-norm cosine**
Output fusion	Satu Vektor	Satu Vektor	**Dua Vektor**	Satu Vektor	Satu Vektor	**Dua Vektor**

Keterangan:

Fokus Utama Skenario 1 (S1) : *Baseline* (Zhu dkk., 2022)

Fokus Utama Skenario 2 (S2) : *Valence Arousal*

Fokus Utama Skenario 3 (S3) : *Two-Stage Attention*

Fokus Utama Skenario 4 (S4) : *Dual Branch Classifier*

Fokus Utama Skenario 5 (S5) : *Residual Connection* (Persamaan (1)) dan *Summary Attention*

Fokus Utama Skenario 6 (S6) : *Full Model*

Nilai tebal menandai komponen yang berbeda dari Skenario 1 (Baseline). Notasi m sebesar 2 mengacu pada jumlah proposal Faster R-CNN (Zhu dkk., 2022), sedangkan k sebesar 36 adalah jumlah region *ink-based* yang konsisten di Skenario 2–6. Setiap skenario menguji satu modifikasi secara independen kecuali S6 yang mengaktifkan seluruh komponen sekaligus; semua skenario dilatih dengan hyperparameter identik sesuai Tabel 3.8 untuk menjamin perbandingan yang adil.

3.6.2 Analisis Hasil

Analisis hasil dilakukan secara kuantitatif dan kualitatif. Secara kuantitatif, hasil metrik dari setiap skenario pada Tabel 3.10 dibandingkan untuk menentukan

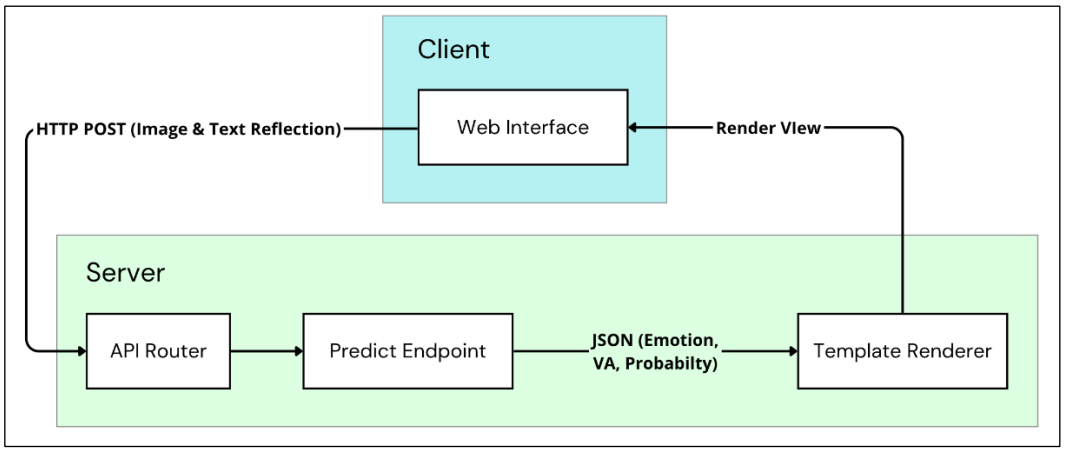
apakah integrasi VA dan modifikasi arsitektur memberikan peningkatan yang signifikan. Secara kualitatif, sistem menghasilkan beberapa visualisasi analitik sebagai berikut:

a. *Confusion Matrix*, yaitu distribusi prediksi dan *ground truth* per kelas emosi.
b. ROC Curve, yaitu kurva karakteristik operasi penerima untuk evaluasi diskriminabilitas.
c. *Precision* dan *Recall Curve*, merupakan trade-off presisi dan recall pada berbagai *threshold*.
d. *Training Curves*, adalah grafik loss dan akurasi pada set latih dan validasi per epoch.
e. *Calibration Curve*, merupakan *reliability diagram* untuk memverifikasi kalibrasi probabilitas (Guo dkk., 2017).
f. VA Scatter Plot, merupakan distribusi nilai *Valence-Arousal* per kelas emosi pada data uji
g. *Error Gallery*, adalah galeri sampel yang salah diklasifikasi untuk analisis penyebab kesalahan.

Hasil analisis secara keseluruhan kemudian dirangkum menjadi kesimpulan mengenai efektivitas pendekatan yang diusulkan, yang disajikan pada Bab 4 dan Bab 5.

3.6.3 Deployment Web API

Sebagai luaran tambahan, sistem dikemas ke dalam antarmuka web berbasis FastAPI agar dapat digunakan secara interaktif oleh praktisi pendidikan maupun kesehatan mental anak. Arsitektur *deployment* diilustrasikan pada Gambar 3.9.



Gambar 3.9 Arsitektur Sistem Antarmuka Web

Kedua model *ITIN* dan *VA Regressor* dimuat ke memori saat server diinisialisasi menggunakan pendekatan *caching* pra-operasional sehingga latensi inferensi per *request* tetap minimal. Sistem menerima unggahan gambar coretan dan teks refleksi melalui protokol *HTTP POST* lalu mengembalikan prediksi kelas emosi, skor probabilitas, dan nilai *Valence-Arousal* dalam format JSON yang kemudian dilakukan render menjadi tampilan HTML melalui templat Jinja2.

3.7 Lingkungan Pengembangan

Seluruh tahapan penelitian diimplementasikan menggunakan spesifikasi perangkat keras dan perangkat lunak yang telah ditampilkan pada Tabel 3.11 dan Tabel 3.12.

Tabel 3.11 Spesifikasi Perangkat Keras

Komponen	**Spesifikasi**
Laptop	ASUS TUF Gaming F15 FX506HC
Processor	Intel Core i5-11400H @ 2.70GHz (6 Core, 12 Thread)
RAM	16 GB
Storage	Intel SSD NVMe 512 GB (SSDPEKNU512GZ)
GPU	NVIDIA GeForce RTX 3050 Laptop GPU

Spesifikasi pada Tabel 3.11 mencerminkan perangkat tingkat menengah yang umum tersedia sehingga penelitian ini dapat direproduksi tanpa infrastruktur

khusus. RTX 3050 4 GB VRAM masih mencukupi untuk *batch size* 16 karena fitur region disimpan terlebih dahulu sebagai file .npz sehingga *bottleneck* komputasi ResNet50 tidak terjadi saat pelatihan berlangsung..

Tabel 3.12 Spesifikasi Perangkat Lunak

Komponen	**Spesifikasi**
Sistem Operasi	Windows 10/11
Bahasa Pemrograman	Python 3.8+
Framework Deep Learning	PyTorch
Library NLP	*Hugging Face Transformers*
Computer Vision	OpenCV, torchvision
Web Framework	FastAPI, Jinja2
IDE	*Visual Studio Code*

Seluruh komponen pada Tabel 3.12 dipilih dari ekosistem *open-source* standar *deep learning* modern untuk menjamin reproduksibilitas. *Hugging Face Transformers* digunakan khususnya untuk mengakses bobot *pretrained* BERT-base dan RoBERTa-base yang konsisten dengan penelitian sebelumnya, sedangkan OpenCV menangani pipeline morfologi *ink-based detection* yang menjadi komponen utama praproses visual.