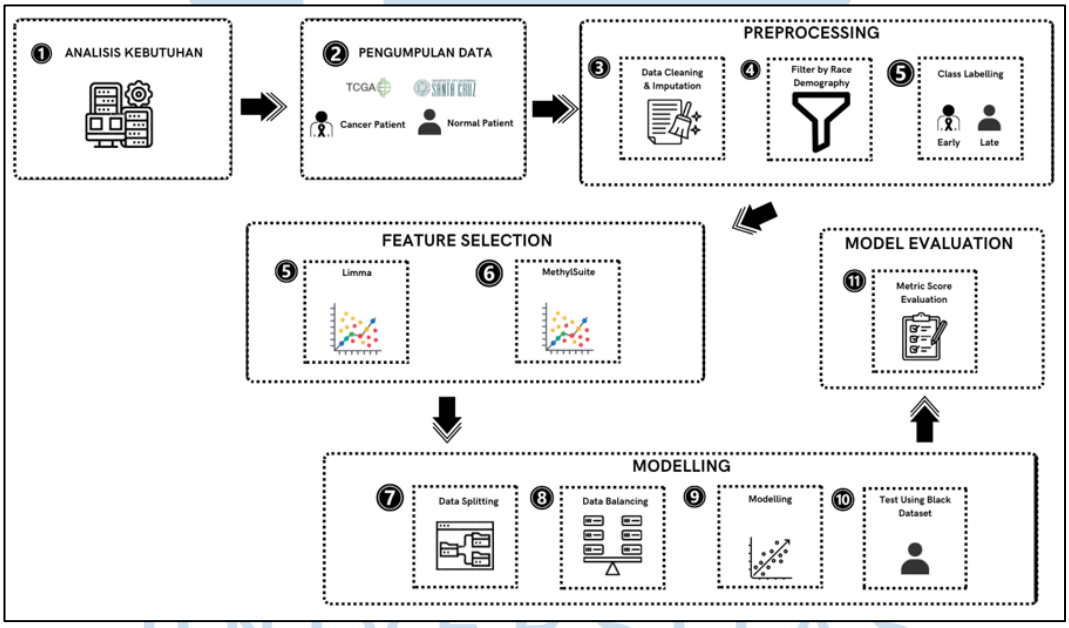
BAB III

METODE PENELITIAN

Dalam menyusun dan melaksanakan penelitian ini, objek yang menjadi fokus adalah prediksi kanker payudara berbasis informasi metilasi DNA. Penelitian ini bertujuan untuk membangun model prediktif yang akurat dalam mengidentifikasi kemungkinan terjadinya kanker payudara berdasarkan pola metilasi genomik.

Untuk mencapai tujuan tersebut, metode dan tahapan yang digunakan selama penelitian ini dapat dilihat pada Gambar 3.1.



Gambar 3. 1 Alur Penelitian yang digarap

3.1 Analisis Kebutuhan

Tahap ini bertujuan untuk menganalisis segala kebutuhan dan persiapan sebelum memasuki proses perancangan sistem prediksi kanker payudara berbasis metilasi DNA. Kebutuhan tersebut mencakup sistem pendukung yang digunakan dalam proses training, evaluasi, maupun pengembangan sistem prediktif. Pada tahap ini, digunakan perangkat lokal dengan spesifikasi prosesor AMD Ryzen 7 5800H, GPU NVIDIA GeForce RTX 3070 8GB, dan RAM sebesar 32 GB.

Kapasitas RAM yang besar digunakan untuk memastikan proses pelatihan model berjalan dengan lancar, mengingat dataset metilasi DNA yang digunakan memiliki ukuran besar dan memerlukan memori tinggi. Spesifikasi ini dinilai mampu mendukung proses pelatihan dan evaluasi model secara efisien karena didukung oleh GPU berperforma tinggi serta memori yang memadai, sehingga mampu menangani beban komputasi dari proses dan analisis data berskala besar.

3.2 Pengumpulan Data

Mencari dan menentukan *dataset* yang akan digunakan untuk proses training serta evaluasi dalam membentuk model yang diharapkan. Pada penelitian ini, digunakan total 2 buah *dataset* sebagai berikut:

1. *TCGA-BRCA.methylation450 dataset*

   Dataset ini didapatkan dari basis data *TCGA* yang dimiliki oleh *The University of California, Santa Cruz* dan terdiri dari 893 sampel pasien yang mengidap penyakit kanker payudara dengan format illumina-450k [29].

Tabel 3. 1 Sampel Data TCGA-BRCA.methylation450

Compost	TCGA-J9-A8CN-01A	TCGA-HC-7078-01A	TCGA-VN-A88K-01A
ENSG00000065534.17	12.09441	13.64971	13.7211
ENSG00000066468.19	7.84549	10.78872	7.257388
ENSG00000084207.14	9.848623	11.84784	10.53333
ENSG00000109846.6	10.45635	10.48784	11.45994
ENSG00000120885.18	12.29949	13.06525	13.4345

2. *TCGA-BRCA.clinical dataset*

   Dataset ini didapatkan dari basis data yang sama dengan data yang pertama dan terdiri dari data *medical record* para pasien. Pada penelitian ini *data clinical* digunakan untuk mengambil label *stage* kanker para pasien untuk melakukan *supervised classification*.

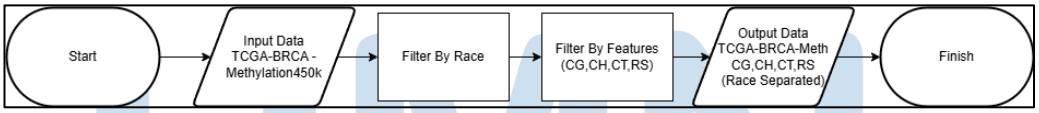
Kedua *dataset* ini diambil dari *database* TCGA milik *University of California, Santa Cruz* (UCSC). Kedua dataset yang digunakan memiliki konten yang saling berkorelasi satu sama lain di mana data pertama berisi informasi metilasi DNA pada Tingkat genom, sementara data kedua berisi informasi klinis dari masing-masing pasien. *Dataset TCGA-BRCA.methylation450* menyajikan nilai Tingkat metilasi dari berbagai *CpG sites* yang terletak di seluruh genom dalam format *illumina450k* yang merupakan *array mikrolarik* yang digunakan untuk mempelajari metilasi DNA pada manusia. Di sisi lain, *TCGA-BRCA.clinical* mencakup data rekam medis seperti usia, stadium kanker, status hidup, rekam terapi dan informasi klinis lainnya yang relevan. Penggabungan kedua dataset ini memungkinkan proses integrasi data untuk membentuk label klasifikasi (stadium kanker) berdasarkan pola metilasi. Proses ini penting guna membangun model prediksi yang tidak hanya berbasis data klinis, tetapi juga mempertimbangkan faktor epigenetik pada tubuh pasien yang memperdalam konteks prediksi.

3.3 Preprocessing Data

Pada tahap ini, langkah pertama yang dilakukan adalah memproses kedua dataset secara terpisah terlebih dahulu karena perbedaan format dan struktur datanya. Untuk data DNA methylation *(TCGA-BRCA.methylation450*), dilakukan proses pembersihan data untuk menghilangkan entri yang tidak relevan atau tidak lengkap. Sedangkan untuk data klinis *(TCGA-BRCA.clinical)*, dilakukan pengecekan menyeluruh terhadap fitur-fitur yang tersedia untuk memastikan bahwa fitur-fitur tersebut dapat diandalkan dan layak digunakan dalam proses penggabungan data selanjutnya. Setelah proses pembersihan data selesai, langkah berikutnya adalah mencocokkan data metilasi DNA dengan data klinikal melalui pencocokan ID pasien yang sesuai.

Kemudian, data *TCGA-BRCA.methylation450* akan difilter berdasarkan demografi sampel, yaitu hanya mencakup pasien dengan kategori ras White, Black, dan Asian. Pemilihan kategori ras ini didasarkan pada pertimbangan variasi biologis dan epigenetik yang dapat memengaruhi pola metilasi DNA. Studi sebelumnya menunjukkan bahwa perbedaan rasial dapat berkontribusi pada heterogenitas molekuler kanker payudara [30], termasuk dalam konteks ekspresi gen dan pola metilasi. Dengan demikian, filtrasi berdasarkan ras dilakukan untuk mendukung analisis *race-specific* yang lebih akurat dan relevan secara biologis, serta untuk mengevaluasi apakah terdapat pola metilasi yang khas pada kelompok ras tertentu yang mungkin berimplikasi pada diagnosis, prognosis, atau respons terapi.

Setelah proses filtrasi berdasarkan ras selesai, data metilasi kemudian dibagi menjadi empat partisi berdasarkan jenis fitur metilasi yang terkandung, yaitu: CG, CH, CT, dan RS. Pemisahan ini bertujuan untuk memungkinkan analisis yang lebih terstruktur dan mendalam terhadap masing-masing fitur yang memiliki karakteristik biologis dan klinis yang berbeda.



Gambar 3. 2 Tahapan pemecahan data TCGA-BRCA-Methylation450k

Setelah dilakukan proses ekstraksi dan pemisahan fitur metilasi DNA, hanya fitur dengan konteks CpG (*Cytosine-phosphate-Guanine*) yang digunakan dalam analisis lanjutan. Hal ini disebabkan oleh tingginya proporsi data kosong pada fitur lain seperti CH (di mana H = A, C, atau T), CT, dan RS (*repeat sequence*), yang dapat menurunkan kualitas data dan memengaruhi akurasi model prediksi. Secara biologis, metilasi pada manusia secara dominan terjadi pada dinukleotida CpG, di mana sekitar 70–80% dari situs CpG mengalami metilasi pada jaringan somatik normal [31]. Sebaliknya, metilasi pada konteks non-CpG lebih jarang ditemukan dan umumnya terjadi pada tipe sel khusus seperti sel punca atau neuron. Selain itu, dalam konteks penelitian kanker, perubahan pola metilasi CpG baik dalam bentuk *hypermethylation* maupun *hypomethylation* telah banyak dikaitkan

dengan inaktivasi gen penekan tumor atau aktivasi onkogen, sehingga menjadikannya biomarker yang lebih relevan untuk diagnosis dan klasifikasi kanker [32]. Oleh karena itu, fokus pada fitur CpG dipilih tidak hanya karena kelengkapan data, tetapi juga karena signifikansinya secara biologis dan kemampuannya dalam mendukung efisiensi serta generalisasi model pembelajaran mesin yang dibangun.

3.4 Data Labelling

Data Setiap sampel dalam dataset telah diberi label stadium kanker payudara berdasarkan panduan dari AJCC Cancer Staging Manual (halaman 349). Pelabelan dilakukan dengan mempertimbangkan parameter utama yang tersedia dalam data fenotipe, yaitu T (Tumor), N (Nodus limfa), dan M (Metastasis) bisa dilihat pada Gambar 3.3. Ketiga parameter ini digunakan untuk menentukan stadium kanker sesuai dengan skema klasifikasi TNM yang ditetapkan oleh AJCC. Distribusi label stadium menunjukkan dominasi sampel pada stadium lanjut (misalnya Stadium III dan IV), yang kemungkinan mencerminkan bias populasi atau kecenderungan ketersediaan data yang lebih besar untuk kasus dengan kondisi klinis yang lebih berat. Pada penelitian ini, pelabelan tidak lagi dilakukan secara manual dikarenakan data yang digunakan telah memiliki label, namun staging stadium I dan II serta III dan IV diklasifikasikan atau dijadikan satu label yaitu *Early* untuk stadium I dan II dan *Late* untuk stadium III dan IV [33].

ANATOMIC STAGE/PROGNOSTIC GROUPS			
Stage 0	Tis	N0	M0
Stage IA	T1*	N0	M0
Stage IB	T0	N1mi	M0
	T1*	N1mi	M0
Stage IIA	T0	N1**	M0
	T1*	N1**	M0
	T2	N0	M0
Stage IIB	T2	N1	M0
	T3	N0	M0
Stage IIIA	T0	N2	M0
	T1*	N2	M0
	T2	N2	M0
	T3	N1	M0
	T3	N2	M0
Stage IIIB	T4	N0	M0
	T4	N1	M0
	T4	N2	M0
Stage IIIC	Any T	N3	M0
Stage IV	Any T	Any N	M1

Gambar 3. 3 Labelling Stage Kanker Payudara berdasarkan AJCC Manual [33]

Untuk keterangan lebih lanjut pada standar labelling ajcc berikut informasi tambahan mengenai apa maksud dari fitur T (0-4), N (0-3) dan Metastasis lihat Tabel 3.2 dan 3.3.

Tabel 3. 2 Penjelasan dari Fitur T

Kode	Keterangan
Tis	Karsinoma in situ (belum menyerang jaringan sekitar)
T0	Tidak ada tumor primer yang terdeteksi
T1	Tumor berukuran ≤ 2 cm
T2	Tumor berukuran > 2 cm hingga ≤ 5 cm
T3	Tumor berukuran > 5 cm
T4	Tumor menyebar ke dinding dada dan/atau kulit

Tabel 3. 3 Penjelasan dari Fitur N

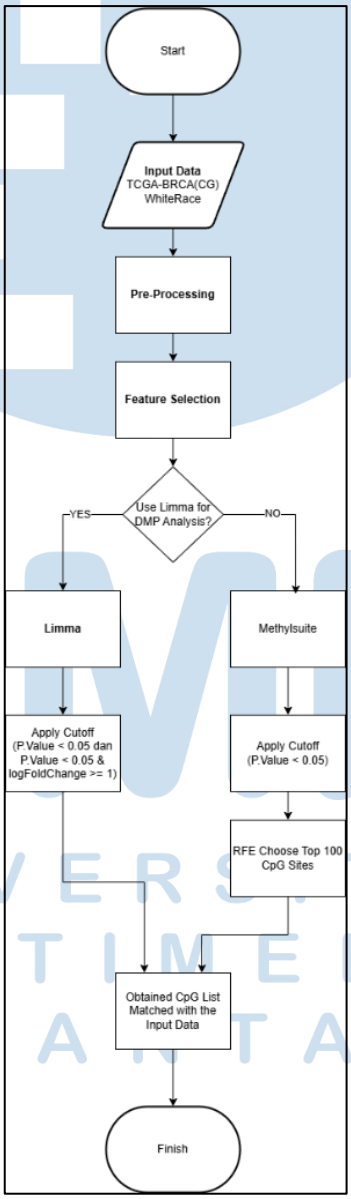
Kode	Keterangan
N0	Tidak ada metastasis ke kelenjar getah bening
N1mi	Metastasis mikro (diameter 0.2–2 mm)
N1	1–3 kelenjar aksila ipsilateral positif
N2	4–9 kelenjar aksila atau kelenjar internal mammary ipsilateral positif
N3	≥10 kelenjar aksila atau melibatkan kelenjar supraklavikula/internal mammary

3.5 Feature Selection

Fitur-fitur untuk membangun model *deep learning* dipilih melalui proses seleksi fitur yang komprehensif dengan menggabungkan analisis *Differentially Methylated CpG sites* (DMCpG), untuk mengidentifikasi gen-gen paling signifikan yang berperan dalam klasifikasi kanker payudara. Analisis DMCpG dilakukan menggunakan pendekatan *Linear Models for Microarray Data* (Limma), sebuah paket dari *R/Bioconductor* yang awalnya dirancang untuk menganalisis data ekspresi gen dari platform microarray dan PCR *throughput* tinggi. Seiring perkembangannya, Limma kini mendukung analisis data RNA-seq dan dapat diperluas untuk data metilasi DNA. Melalui pendekatan ini, situs-situs CpG yang mengalami hipermetilasi dan hipometilasi dapat diidentifikasi sebagai indikator adanya disregulasi genetik yang relevan dalam proses karsinogenesis.

Selain Limma, digunakan pula paket *MethylSuite* untuk menganalisis data metilasi DNA berdasarkan konteks genomik seperti CG, CH, CT, dan RS. Paket ini mendukung *pre-processing*, normalisasi, dan deteksi situs CpG signifikan berdasarkan kualitas data dan lokasi genom. Integrasi hasil dari Limma dan *MethylSuite* memungkinkan pemilihan fitur metilasi yang tidak hanya signifikan

secara statistik, tetapi juga bermakna secara biologis. Selanjutnya, metode *Recursive Feature Elimination* (RFE) digunakan untuk menyaring fitur-fitur terbaik dengan mengeliminasi variabel yang kurang berkontribusi terhadap kinerja model, sehingga menghasilkan subset fitur yang optimal untuk pembangunan model *ss* yang akurat dan efisien. Skema pemilihan fitur yang lebih detail dapat dilihat pada gambar 3.4



Gambar 3. 4 Flowchart pemilihan fitur

Fitur-fitur untuk pembangunan model dipilih menggunakan tiga metode utama, yaitu *Limma, MethylSuite,* dan *Recursive Feature Elimination (RFE).* Analisis awal dilakukan menggunakan *Limma*, di mana diterapkan dua skenario threshold yaitu skenario pertama menggunakan nilai $p < 0.05$, sedangkan skenario kedua menggunakan kombinasi nilai $p < 0.05$ dan $log_2FoldChange \geq 1$ untuk menyaring gen yang paling berbeda secara signifikan.

Metode kedua menggunakan *MethylSuite*, sebuah *package* yang dirancang khusus untuk mengidentifikasi gen-gen dengan tingkat signifikansi tinggi terhadap kanker. Setelah diperoleh daftar gen dari MethylSuite, diterapkan kembali *threshold* p-value < 0.05 guna memastikan bahwa hanya gen-gen yang paling signifikan yang dipertahankan.

Langkah terakhir menggunakan hasil dari MethylSuite yang telah disaring berdasarkan *threshold* tersebut, kemudian diterapkan *RFE* untuk memilih 100 gen teratas yang paling berperan dalam perkembangan kanker. Pendekatan ini memastikan bahwa fitur-fitur yang digunakan dalam model tidak hanya secara statistik signifikan, tetapi juga relevan secara biologis terhadap jalur-jalur klinis kanker payudara, sehingga memberikan fondasi yang kuat untuk pembangunan model prediktif. Untuk Lebih detail, skenario *feature selection* dapat dilihat pada tabel 3.4

Tabel 3. 4 Skema Feature Selection Penelitian ini

No.	Metode	Deskripsi Tahapan	***Threshold* / Kriteria**	Hasil Akhir
1	*Limma*	Analisis diferensial ekspresi gen (DGE) untuk identifikasi gen signifikan.	p-value < 0.05 p-value < 0.05 & logFC ≥ 1	Daftar gen signifikan dari dua skenario

No.	Metode	Deskripsi Tahapan	*Threshold* / Kriteria	Hasil Akhir
2	*MethylSuite*	Identifikasi gen dengan tingkat metilasi signifikan terhadap kanker.	p-value < 0.05	Daftar gen yang sangat signifikan terhadap kanker
3	*RFE* (dengan data dari MethylSuite)	Pemilihan fitur menggunakan algoritma RFE	Top 100 gen berdasarkan ranking RFE	100 Cpg paling relevan

Setelah diperoleh hasil dari proses seleksi fitur, dilakukan validasi *biomarker* sebagai langkah lanjutan untuk memastikan relevansi biologis dari fitur yang terpilih. Validasi ini dilakukan melalui tiga tahap utama. Pertama, dilakukan pemetaan dari *probe* CpG ke nama gen yang sesuai menggunakan basis data *Illumina* [[33], guna mengidentifikasi representasi genetik dari fitur metilasi yang dipilih. Kedua, gen-gen yang telah diidentifikasi kemudian divalidasi melalui basis data *Gene Set Enrichment Analysis (GSEA)* [34] untuk menilai keterkaitan biologis dan fungsionalnya dalam konteks kanker, khususnya kanker payudara. Tahap ketiga adalah pemeriksaan potensi mutasi dari gen-gen tersebut menggunakan sumber data dari *cBioPortal* [35] dan *COSMIC (Catalogue of Somatic Mutations in Cancer)* [36]. Validasi ini bertujuan untuk memperkuat justifikasi biologis terhadap fitur-fitur yang dipilih dalam model, serta memastikan bahwa gen yang digunakan tidak hanya signifikan secara statistik, tetapi juga relevan secara klinis dan biologis.

3.6 Proses Training Model

Pada tahap pelatihan (*training*), digunakan arsitektur *Multi-Layer Perceptron (MLP)* yang dibangun secara *sequential* dengan tiga *hidden layer* dan

satu lapisan output. Lapisan pertama memiliki 256 neuron dan menggunakan fungsi aktivasi *ReLU* serta inisialisasi bobot *GlorotUniform* untuk mempercepat konvergensi [37]. Lapisan ini juga dilengkapi dengan *Batch Normalization* guna menstabilkan distribusi nilai aktivasi, serta *Dropout* sebesar 40% untuk mengurangi risiko *overfitting* pada data berdimensi tinggi seperti metilasi DNA [38],[39]. Lapisan kedua terdiri dari 128 neuron dengan konfigurasi serupa, yaitu aktivasi *ReLU*, *Batch Normalization*, dan *Dropout* sebesar 40%. Selanjutnya, lapisan ketiga terdiri dari 64 neuron dengan fungsi aktivasi *ReLU* dan *Dropout* sebesar 30%, namun tidak dilengkapi normalisasi. Lapisan output menggunakan fungsi aktivasi *sigmoid* dan jumlah neuron yang sesuai dengan jumlah kelas target, yang umum digunakan dalam klasifikasi biner. Arsitektur model secara detil dapat dilihat pada Tabel 3.5

Tabel 3. 5 Arsitektur model Multi-Layer Perceptron pada penelitian ini

Layer	Tipe Layer	Jumlah Neuron	Fungsi Aktivasi	Inisialisasi Bobot	Batch Normalization	Dropout
Input + Hidden 1	Dense	256	ReLU	GlorotUniform	Ya	0.4
Hidden 2	Dense	128	ReLU	GlorotUniform	Ya	0.4
Hidden 3	Dense	64	ReLU	GlorotUniform	Tidak	0.3
Output	Dense	Jumlah kelas	Sigmoid	Default	Tidak	Tidak

Model dikompilasi menggunakan algoritma optimisasi Adam, dengan fungsi *loss* berupa *categorical crossentropy* yang sesuai untuk data target yang telah dikonversi ke format one-hot encoding. Proses *trainning* dilakukan selama maksimum 150 *epoch* dengan *batch size* sebesar 32. Untuk meningkatkan performa dan mencegah *overfitting*, digunakan beberapa *callback* seperti *ModelCheckpoint* untuk menyimpan bobot terbaik berdasarkan akurasi validasi, *EarlyStopping*

dengan *patience* selama 15 *epoch*, serta *ReduceLROnPlateau* untuk menurunkan *learning rate* apabila *val_loss* stagnan selama 5 *epoch.* Selain itu, digunakan juga model lain dengan arsitektur yang terdiri dari tiga lapisan, yaitu dua *hidden layer* dan satu *output layer*. Pada *hidden layer* pertama, digunakan *layer Dense* dengan 128 neuron, fungsi aktivasi *ReLU*, dan inisialisasi bobot *GlorotUniform*, dilengkapi dengan *batch normalization* dan *dropout* sebesar 0.4 untuk mengurangi *overfitting*. *Hidden layer* kedua menggunakan 64 neuron dengan fungsi aktivasi *ReLU* dan inisialisasi bobot *GlorotUniform*, namun tanpa *batch normalization* dan dengan *dropout* sebesar 0.3. Lapisan *output* menggunakan *layer Dense* dengan jumlah neuron sesuai jumlah kelas, fungsi aktivasi *Sigmoid*, serta inisialisasi bobot *default* tanpa penerapan *batch normalization* maupun *dropout*. Arsitektur ini dirancang untuk mengeksplorasi performa alternatif model dengan konfigurasi *regularisasi* dan jumlah *neuron* yang berbeda

Tabel 3. 6 Arsitektur model Multi-Layer Perceptron ke-2 pada penelitian ini

Layer	Tipe Layer	Jumlah Neuron	Fungsi Aktivasi	Inisialisasi Bobot	Batch Normalization	Dropout
Hidden 1	Dense	128	ReLU	GlorotUniform	Ya	0.4
Hidden 2	Dense	64	ReLU	GlorotUniform	Tidak	0.3
Output	Dense	Jumlah kelas	Sigmoid	Default	Tidak	Tidak

3.7 Evaluasi Model

Setelah proses pelatihan selesai, model dievaluasi menggunakan data validasi yang telah dipisahkan sebelumnya sebesar 10% dari total dataset. Evaluasi ini bertujuan untuk mengukur kemampuan generalisasi model terhadap data yang tidak pernah dilihat selama proses pelatihan. Pengukuran awal mencakup nilai *loss* dan *accuracy*, di mana *loss* mencerminkan tingkat kesalahan prediksi dan *accuracy* mengukur proporsi prediksi yang sesuai dengan label aktual. Untuk model ideal,

sebelum proses pelatihan dilakukan, data latih terlebih dahulu diolah menggunakan *Synthetic Minority Oversampling Technique* (SMOTE) guna mengatasi ketidakseimbangan kelas. Setelah pelatihan dan validasi, model akhir dievaluasi menggunakan data uji sebesar 10% untuk mengukur kinerja model terhadap data yang benar-benar belum pernah digunakan sebelumnya. Proses prediksi dilakukan dengan menerapkan fungsi aktivasi sigmoid pada lapisan output untuk menghasilkan distribusi probabilitas terhadap masing-masing kelas.

Selain itu, performa model juga ditinjau melalui *confusion matrix* dan *classification report*, di mana *confusion matrix* menunjukkan distribusi prediksi benar dan salah pada masing-masing label secara visual dengan matriks 2×2, sedangkan *classification report* memuat metrik evaluasi seperti presisi *(precision)*, *recall*, dan *F1-score* untuk setiap kelas. Melalui evaluasi ini, dapat disimpulkan seberapa baik model melakukan klasifikasi serta mengidentifikasi kemungkinan bias atau ketidakseimbangan kinerja antar kelas yang perlu diperbaiki pada iterasi model selanjutnya.