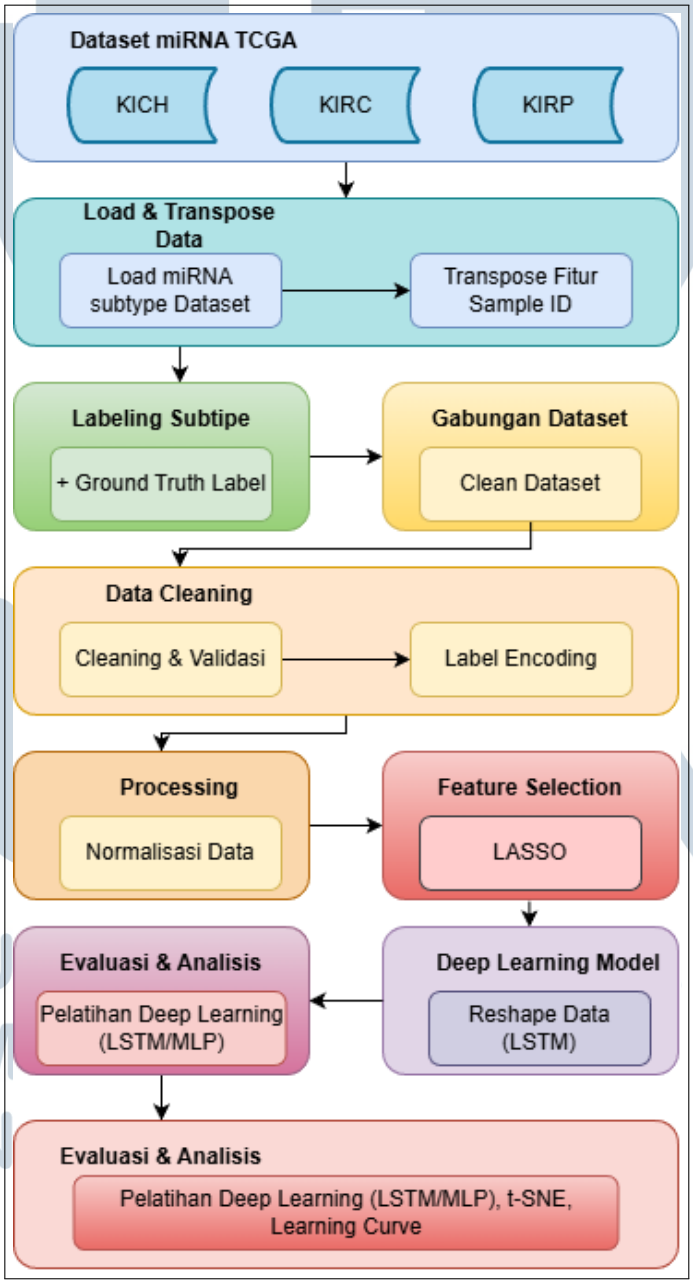
BAB 3
METODE PENELITIAN

Penelitian ini menggunakan rancangan metode sesuai dengan Gambar 3.1 Perangkat yang digunakan selama menjalankan proses adalah Windows 11 OS, Processor Ryzen 3 5800H (8 Cores 16 Threads), GPU NVIDIA RTX 3070 Laptop 8GB, dan 32GB 3200MT/s RAM.



Gambar 3.1. Alur Penelitian

Gambar 3.1 menunjukkan alur penelitian yang digunakan dalam pengembangan model klasifikasi subtipe kanker ginjal berbasis data ekspresi miRNA. Alur penelitian ini terdiri dari beberapa tahapan utama sebagai berikut:

1. Pengumpulan data ekspresi miRNA dan data klinis dari TCGA untuk tiga subtipe kanker ginjal, yaitu KICH, KIRC, dan KIRP.
2. Pra-pemrosesan data yang meliputi proses *load* dan transposisi data, harmonisasi fitur miRNA antar subtipe, penggabungan dataset, serta pembersihan data.
3. Pelabelan sampel berdasarkan subtipe kanker ginjal dan normalisasi data untuk memastikan keseragaman skala fitur.
4. Seleksi fitur menggunakan metode LASSO untuk mereduksi dimensi data dan memilih miRNA yang paling relevan.
5. Pembangunan model klasifikasi menggunakan *Long Short-Term Memory* (LSTM) sebagai model utama dan *Multi-Layer Perceptron* (MLP) sebagai pembanding.
6. Evaluasi dan analisis model menggunakan metrik klasifikasi dan visualisasi pendukung.

3.1 Pengumpulan Data

Data yang digunakan dalam penelitian ini merupakan data sekunder yang diperoleh dari *UCSC Xena Browser* dengan *cohort GDC TCGA Kidney Cancer* yang di download pada 20 Agustus 2025. Dataset ini mencakup tiga subtipe utama kanker ginjal, yaitu:

1. *Kidney Chromophobe* (KICH)
2. *Kidney Renal Clear Cell Carcinoma* (KIRC)
3. *Kidney Renal Papillary Cell Carcinoma* (KIRP)

Data yang digunakan berupa data ekspresi *microRNA (miRNA)* yang dihasilkan melalui teknologi *Illumina HiSeq*. Data tersebut telah melalui proses normalisasi dalam satuan

$$\log_2(\mathrm{RPM}+1)$$

di mana RPM (*Reads Per Million*) merupakan ukuran tingkat ekspresi miRNA. Transformasi logaritmik ini bertujuan untuk mengurangi variasi teknis antar sampel serta menstabilkan distribusi data sehingga lebih sesuai untuk proses analisis dan pemodelan pembelajaran mesin.

Selain data ekspresi miRNA, digunakan juga data klinis (*phenotype*) pasien *The Cancer Genome Atlas* (TCGA) yang memuat informasi subtipe kanker. Data klinis ini berfungsi sebagai ground truth label pada proses klasifikasi multikelas.

3.2 Pra-pemrosesan Data

Tahap pra-pemrosesan dilakukan untuk memastikan data berada dalam format yang sesuai serta memiliki kualitas yang baik sebelum digunakan dalam proses pemodelan. Pra-pemrosesan yang tepat diperlukan untuk meminimalkan kesalahan analisis dan meningkatkan kinerja model klasifikasi.

Langkah-langkah pra-pemrosesan data dalam penelitian ini meliputi:

1. Transposisi Data, dilakukan agar setiap baris merepresentasikan satu sampel pasien dan setiap kolom merepresentasikan satu fitur *miRNA*.

2. Harmonisasi Fitur miRNA, dengan mengambil irisan (*intersection*) fitur miRNA yang muncul secara konsisten pada seluruh subtipe kanker ginjal, yaitu *Kidney Chromophobe* (KICH), *Kidney Renal Clear Cell Carcinoma* (KIRC), dan *Kidney Renal Papillary Cell Carcinoma* (KIRP), sehingga diperoleh fitur yang seragam untuk seluruh data.

3. Penggabungan Data, yaitu menggabungkan data ekspresi miRNA dengan data klinis berdasarkan *sample ID* TCGA untuk memastikan kesesuaian antara fitur molekuler dan label klinis.

4. Pembersihan Data (*Data Cleaning*), dilakukan dengan menghapus sampel duplikat serta sampel yang tidak memiliki label subtipe kanker ginjal yang valid.

5. Pelabelan Kelas, di mana setiap sampel diberi label numerik menggunakan teknik *label encoding* untuk merepresentasikan kelas *Kidney Chromophobe* (KICH) = 0, *Kidney Renal Clear Cell Carcinoma* (KIRC) = 1, dan *Kidney Renal Papillary Cell Carcinoma* (KIRP) = 2.

Seluruh tahapan pra-pemrosesan data diimplementasikan menggunakan pustaka *Pandas* dan *NumPy* untuk manipulasi data, serta *scikit-learn* untuk proses pelabelan kelas dan persiapan data sebelum pemodelan.

3.3 Struktur Data untuk Klasifikasi

Dalam penelitian ini, dataset yang digunakan untuk proses klasifikasi disusun dalam bentuk matriks fitur, di mana setiap baris merepresentasikan satu sampel pasien kanker ginjal, dan setiap kolom merepresentasikan satu fitur miRNA hasil seleksi. Label kelas menunjukkan subtipe kanker ginjal, yaitu *Kidney Chromophobe* (KICH), *Kidney Renal Clear Cell Carcinoma* (KIRC), dan *Kidney Renal Papillary Cell Carcinoma* (KIRP), yang digunakan sebagai *ground truth* dalam proses pelatihan dan evaluasi model.

3.4 Pembentukan dan Pembagian Dataset

Dataset hasil pra-pemrosesan selanjutnya digabungkan menjadi satu dataset multikelas yang terdiri dari seluruh sampel dari tiga subtipe kanker ginjal, yaitu *Kidney Chromophobe* (KICH), *Kidney Renal Clear Cell Carcinoma* (KIRC), dan *Kidney Renal Papillary Cell Carcinoma* (KIRP). Penggabungan ini bertujuan untuk membangun model klasifikasi multikelas yang mampu membedakan ketiga subtipe kanker ginjal tersebut secara akurat.

Dataset yang telah terbentuk kemudian dibagi menjadi data latih (*training set*) dan data uji (*testing set*) menggunakan metode *stratified split*. Metode ini dipilih untuk memastikan bahwa proporsi masing-masing kelas tetap terjaga pada setiap subset data, sehingga distribusi kelas pada data latih dan data uji merepresentasikan distribusi kelas pada dataset asli.

Selain pembagian data, penelitian ini juga menerapkan teknik *k-fold cross validation* pada data latih. Penerapan teknik ini bertujuan untuk mengevaluasi kestabilan dan kemampuan generalisasi model yang dibangun, serta mengurangi ketergantungan model terhadap satu skema pembagian data tertentu. Dengan demikian, performa model yang diperoleh diharapkan lebih robust dan dapat digeneralisasikan dengan baik pada data yang belum pernah dilihat sebelumnya.

3.5 Seleksi Fitur Berbasis LASSO

Data ekspresi *microRNA (miRNA)* memiliki karakteristik berdimensi tinggi, sehingga berpotensi menimbulkan permasalahan *curse of dimensionality* dan *overfitting* pada proses pemodelan. Oleh karena itu, penelitian ini menerapkan metode seleksi fitur menggunakan *Least Absolute Shrinkage and Selection Operator* (LASSO).

Metode LASSO menggunakan regularisasi L1 yang bekerja dengan memberikan penalti terhadap nilai absolut koefisien fitur. Regularisasi ini secara otomatis mengecilkan koefisien fitur yang kurang relevan hingga bernilai nol, sehingga hanya fitur miRNA yang memiliki kontribusi signifikan terhadap proses klasifikasi subtipe kanker ginjal yang dipertahankan.

Sebelum proses seleksi fitur dilakukan, data distandarisasi menggunakan *StandardScaler* agar seluruh fitur berada pada skala yang sebanding dan tidak didominasi oleh fitur dengan rentang nilai yang lebih besar. Proses seleksi fitur ini menghasilkan beberapa subset fitur dengan jumlah yang berbeda, yaitu 20, 35, dan 50 miRNA. Subset fitur tersebut selanjutnya digunakan untuk membandingkan performa model dan menentukan jumlah fitur yang paling optimal.

Untuk mencegah terjadinya *data leakage*, seleksi fitur berbasis LASSO hanya diterapkan pada data latih (*training set*), sementara data uji (*testing set*) hanya digunakan pada tahap evaluasi kinerja model.

3.6 Pembangunan Model

Pada penelitian ini, model *Long Short-Term Memory* (LSTM) digunakan sebagai model utama klasifikasi. Selain itu, model *Multi-Layer Perceptron* (MLP) digunakan sebagai model *baseline* pembanding untuk mengevaluasi efektivitas penerapan arsitektur *Long Short-Term Memory* (LSTM).

3.6.1 *Multi-Layer Perceptron* (MLP)

Model *Multi-Layer Perceptron* (MLP) digunakan sebagai arsitektur jaringan saraf *feedforward* yang terdiri dari beberapa lapisan tersembunyi (*hidden layers*), sedangkan model *Long Short-Term Memory* (LSTM) merupakan salah satu varian dari *Recurrent Neural Network* (RNN) yang dirancang untuk menangkap ketergantungan jangka panjang pada data sekuensial.

3.6.2 *Long Short-Term Memory* (LSTM)

Model *Long Short-Term Memory* (LSTM) data hasil seleksi fitur diubah ke dalam bentuk tiga dimensi (*reshape*) agar sesuai dengan struktur input jaringan *Long Short-Term Memory* (LSTM), yaitu dalam format (*samples*, *timesteps*, *features*). Penyesuaian ini diperlukan agar model *Long Short-Term Memory* (LSTM) dapat memproses data sesuai dengan mekanisme memori internal yang dimilikinya.

Model *deep learning* dirancang untuk menangkap hubungan non-linear antar fitur miRNA yang tidak dapat dimodelkan secara optimal oleh algoritma *machine learning* konvensional. Proses pelatihan model dilakukan dengan menggunakan fungsi *loss categorical cross-entropy* yang sesuai untuk permasalahan klasifikasi multikelas, serta *optimizer Adam* untuk mempercepat konvergensi dan meningkatkan stabilitas proses pelatihan.

Selain itu, mekanisme *early stopping* diterapkan selama proses pelatihan untuk mencegah terjadinya *overfitting*, dengan menghentikan pelatihan model secara otomatis ketika performa pada data validasi tidak lagi mengalami peningkatan.

3.7 Evaluasi dan Analisis Model

Metode evaluasi yang digunakan dalam penelitian ini mengacu pada teori dan konsep evaluasi klasifikasi multikelas yang telah dijelaskan pada Bab 2. Evaluasi dilakukan untuk mengukur kinerja prediksi model terhadap data uji menggunakan metrik kuantitatif dan visualisasi pendukung.

Evaluasi kinerja model dilakukan menggunakan pendekatan *cross validation* untuk memperoleh hasil evaluasi yang lebih stabil dan representatif. Pendekatan ini memungkinkan pengujian model pada beberapa skema pembagian data, sehingga dapat meminimalkan bias akibat pemilihan data latih dan data uji tertentu.

Metrik evaluasi utama yang digunakan dalam penelitian ini meliputi:

1. *Accuracy*
2. *Precision*, *Recall*, dan *F1-score* untuk setiap kelas
3. *Matthews Correlation Coefficient (MCC)* pada setiap *fold*
4. *Confusion Matrix*

Selain evaluasi kuantitatif, dilakukan pula analisis kualitatif melalui beberapa visualisasi pendukung. Visualisasi *learning curve* digunakan untuk menganalisis proses konvergensi model selama pelatihan, serta untuk mengidentifikasi potensi terjadinya *underfitting* atau *overfitting*.

Selain itu, visualisasi *t-distributed Stochastic Neighbor Embedding (t-SNE)* diterapkan menggunakan fitur miRNA terpilih untuk mengamati pemisahan antar subtipe kanker ginjal (*Kidney Chromophobe* (KICH), *Kidney Renal Clear Cell Carcinoma* (KIRC), dan *Kidney Renal Papillary Cell Carcinoma* (KIRP).) secara visual pada ruang berdimensi rendah.

Hasil evaluasi dari berbagai skenario jumlah fitur, yaitu 20, 35, dan 50 miRNA, kemudian dibandingkan untuk menentukan konfigurasi model dan jumlah fitur yang memberikan kinerja terbaik dalam mengklasifikasikan subtipe kanker ginjal.