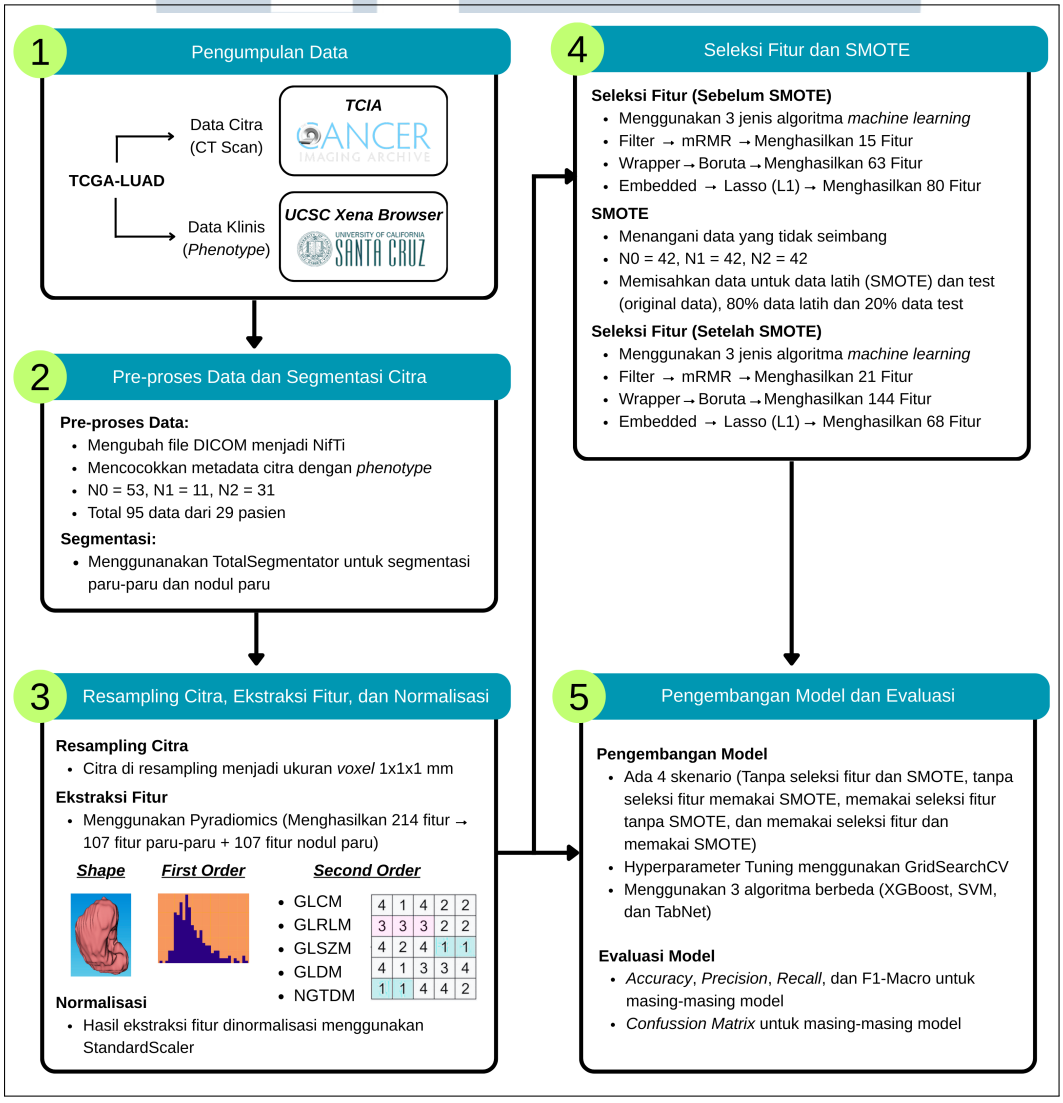
BAB 3
METODE PENELITIAN

Bab ini menguraikan desain dan prosedur teknis pengembangan sistem klasifikasi N-stage berbasis fitur radiomik citra CT dari dataset TCGA-LUAD. Metodologi yang diterapkan mentransformasikan citra medis menjadi fitur numerik untuk kebutuhan model machine learning dan deep learning. Struktur bab disusun berdasarkan alur pipeline penelitian yang mencakup tahap pengambilan data hingga evaluasi model. Alur penelitian secara menyeluruh digambarkan pada Gambar 3.1.

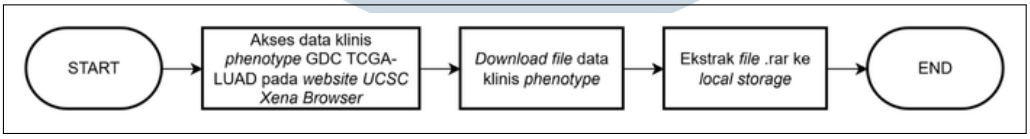


Gambar 3.1. Alur Penelitian Dampak SMOTE dan Komparasi Metode Seleksi Fitur pada Klasifikasi N-Stage Adenokarsinoma Paru Berbasis Radiomik

3.1 Pengumpulan Data

3.1.1 Data Klinis

Data klinis dan fenotipe untuk kohort Lung Adenocarcinoma (LUAD) diperoleh dari repositori publik UCSC Xena Browser. UCSC Xena adalah sebuah platform visualisasi dan eksplorasi online yang dirancang untuk meng-host dan mengintegrasikan dataset genomik dan fenotipe multi-omik dari berbagai repositori publik utama, termasuk Genomic Data Commons (GDC). Dataset spesifik yang digunakan dalam penelitian ini adalah GDC TCGA-LUAD. Ini merupakan dataset dari program The Cancer Genome Atlas (TCGA) yang telah diproses ulang secara seragam oleh GDC untuk standardisasi. Data ini diunduh dalam format tabular (file .csv) dan mencakup variabel fenotipe krusial yang diperlukan untuk pemodelan, seperti data demografis pasien, status survival (keseluruhan dan bebas penyakit), dan—yang paling penting untuk penelitian ini—informasi staging klinis dan patologis, termasuk data TNM (Tumor, Node, Metastasis). Proses pengambilan data klinis ditunjukkan di Gambar 3.2.

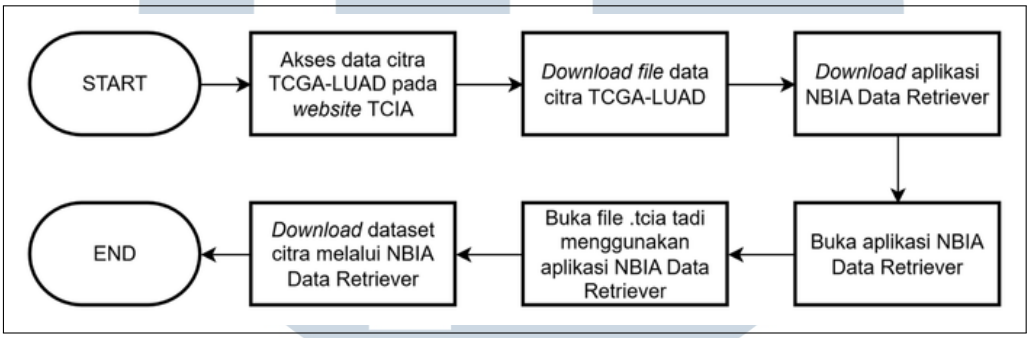


Gambar 3.2. Proses Pengambilan Data Klinis TCGA-LUAD

3.1.2 Data Gambar

Data pencitraan radiologis yang sesuai dengan pasien dalam kohort klinis TCGA-LUAD diperoleh dari The Cancer Imaging Archive (TCIA). TCIA adalah sebuah repositori publik yang didanai oleh National Cancer Institute (NCI) yang berfungsi untuk menyimpan dan melakukan de-identifikasi arsip besar gambar medis kanker untuk akses publik. Data di dalam TCIA diorganisasi berdasarkan ”koleksi” (collections). Koleksi yang dimanfaatkan dalam penelitian ini adalah TCGA-LUAD. Koleksi ini secara eksplisit dikurasi untuk menyediakan gambar klinis yang cocok (memiliki keterkaitan) dengan subjek dari dataset genomik TCGA. Inti dari desain penelitian ini terletak pada keterhubungan antara dua sumber data ini. Infrastruktur data publik ini menyediakan ”Matched TCGA patient identifiers” (Pengidentifikasi pasien TCGA yang cocok). Pengidentifikasi unik inilah yang memungkinkan para peneliti, termasuk dalam studi ini,

untuk secara valid menghubungkan dan mengeksplorasi korelasi antara fenotipe radiologis (dari GDC/Xena) dan citra radiologis (dari TCIA). Secara spesifik, penelitian ini secara eksklusif berfokus pada modalitas gambar CT-Scan dari koleksi TCGA-LUAD. Pemilihan CT-Scan didasarkan pada signifikansi klinisnya sebagai modalitas standar emas dalam diagnosis, staging, dan pemantauan kanker paru. Studi radiomik sebelumnya telah memberikan bukti kuat bahwa fenotipe radiomik berbasis CT (CT-based radiomic phenotypes) memiliki kemampuan untuk mengidentifikasi n-staging LUAD. Proses pengambilan data citra ditunjukkan di Gambar 3.3.



Gambar 3.3. Proses Pengambilan Data Citra TCGA-LUAD

3.2 Pre-proses data dan Segmentasi Citra

3.2.1 Pre-proses Data

Pada tahap awal pre-proses, langkah pertama adalah membuka data klinis phenotype dari GDC TCGA-LUAD yang berisi data klinis terkait dengan jenis kanker paru-paru dan data citra TCGA-LUAD. Data citra ini, yang awalnya dalam format DICOM, kemudian dikonversi ke dalam format NIfTI menggunakan perangkat lunak Slicer 3D. Proses konversi ini penting untuk mempermudah pemrosesan citra pada tahap-tahap berikutnya.

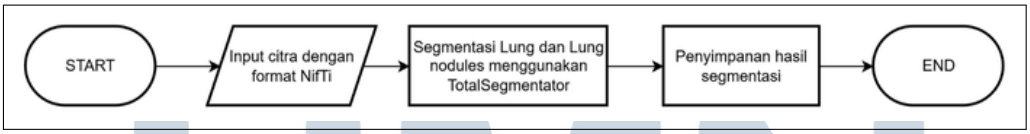
Setelah data citra berhasil dikonversi, langkah selanjutnya adalah melakukan pemetaan antara `ajcc_pathologic_n_diagnoses` dengan phenotype yang terkait dengan metadata citra, sehingga informasi medis dan citra dapat diintegrasikan untuk analisis lebih lanjut. Proses pengambilan data citra ditunjukkan di Gambar 3.4.

Gambar 3.4. Proses Preproses Data TCGA-LUAD

3.2.2 Segmentasi Citra

Pada tahap segmentasi citra, langkah pertama yang dilakukan adalah menginput citra dalam format NIfTI. Format NIfTI (Neuroimaging Informatics Technology Initiative) sering digunakan dalam pengolahan citra medis karena kemampuannya menyimpan data citra 3D. Setelah citra dimasukkan, proses selanjutnya adalah melakukan segmentasi organ paru-paru dan nodul paru-paru menggunakan alat TotalSegmentator. TotalSegmentator merupakan perangkat lunak yang dirancang untuk mendeteksi dan memisahkan area yang relevan pada citra medis, seperti paru-paru dan nodul.

Setelah segmentasi selesai, hasil segmentasi akan disimpan untuk digunakan pada analisis atau tahapan pengolahan citra lebih lanjut. Gambar 3.5 alur proses segmentasi citra yang dilakukan pada tahap ini.



Gambar 3.5. Proses Segmentasi Data Citra TCGA-LUAD

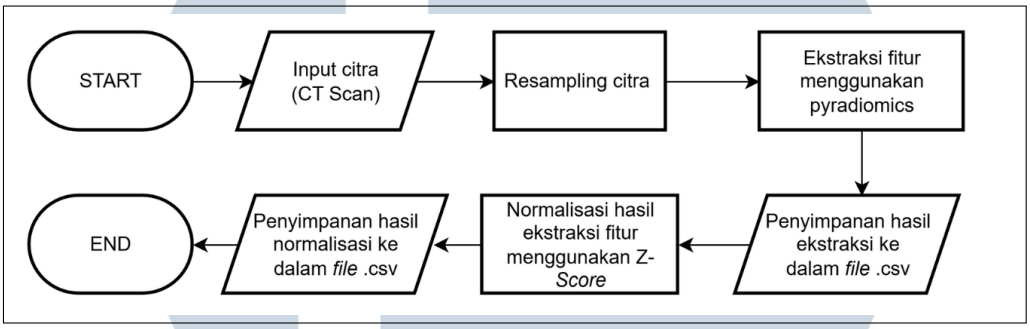
3.3 Resampling Citra, Ekstraksi Fitur, dan Normalisasi

Pada tahap ini, beberapa langkah penting dilakukan untuk menyiapkan data yang akan digunakan dalam analisis selanjutnya. Langkah pertama adalah resampling voxel citra 1x1x1 mm, yang bertujuan untuk mengubah ukuran citra agar sesuai dengan standar yang dibutuhkan untuk analisis berikutnya. Resampling ini penting untuk memastikan konsistensi dalam pengolahan citra yang akan diproses lebih lanjut.

Setelah itu, dilakukan ekstraksi fitur menggunakan metode pyRadiomics. Metode ini digunakan untuk mengekstrak berbagai fitur tekstur dari citra medis, seperti fitur bentuk, intensitas, dan statistik lainnya yang berguna dalam analisis lebih lanjut. Fitur-fitur yang diekstraksi akan disimpan dalam format file CSV, yang

kemudian dapat digunakan untuk berbagai jenis analisis statistik.

Langkah selanjutnya adalah normalisasi hasil ekstraksi fitur menggunakan metode Z-Score. Proses ini bertujuan untuk menormalkan data agar distribusinya memiliki rata-rata 0 dan deviasi standar 1, yang penting untuk memastikan konsistensi dan perbandingan antar fitur. Alur dari proses ini dapat dilihat pada Gambar 3.6.



Gambar 3.6. Proses Resampling Citra, Ekstraksi Fitur, dan Normalisasi

3.4 Seleksi Fitur dan SMOTE

3.4.1 Seleksi Fitur

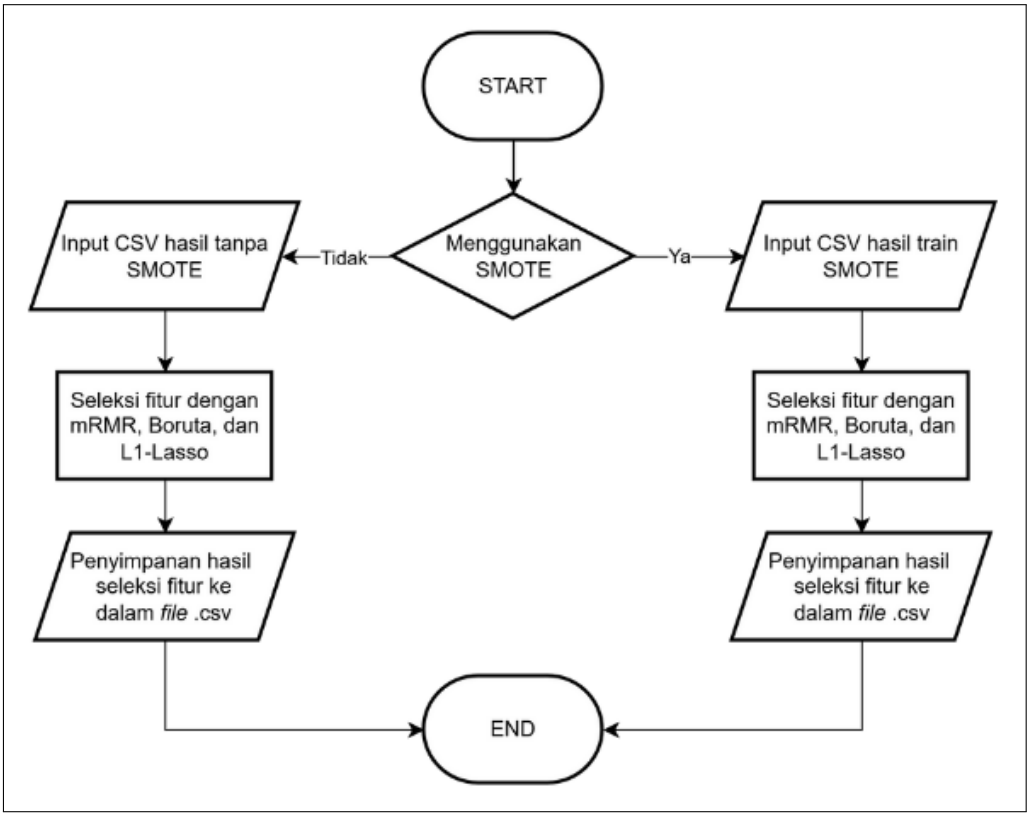
Pada tahap seleksi fitur, langkah pertama yang dilakukan adalah menginput file CSV yang berisi data hasil ekstraksi fitur sebelumnya. Pada tahap ini, terdapat dua kemungkinan jalur yang dapat diambil, yaitu dengan atau tanpa menggunakan SMOTE (Synthetic Minority Over-sampling Technique). SMOTE adalah teknik yang digunakan untuk mengatasi ketidakseimbangan kelas dengan cara menambah jumlah sampel minoritas melalui sintesis.

Jika data yang digunakan tidak menggunakan SMOTE, maka dilakukan seleksi fitur langsung pada data tersebut menggunakan beberapa metode, yaitu mRMR (minimum Redundancy Maximum Relevance), Boruta, dan L1-Lasso. Metode-metode ini digunakan untuk memilih fitur-fitur yang paling relevan dan berpengaruh dalam analisis selanjutnya.

Di sisi lain, jika SMOTE digunakan, maka data akan diproses terlebih dahulu dengan teknik ini untuk menambah jumlah data sebelum dilakukan seleksi fitur. Setelah itu, seleksi fitur dengan metode mRMR, Boruta, dan L1-Lasso tetap dilakukan pada data yang telah disintesis.

Hasil dari seleksi fitur ini akan disimpan kembali dalam file CSV untuk digunakan pada tahap analisis atau modeling lebih lanjut. Gambar 3.7

menggambarkan alur proses seleksi fitur yang dilakukan pada tahap ini.



Gambar 3.7. Proses Seleksi Fitur dengan SMOTE dan Tanpa SMOTE

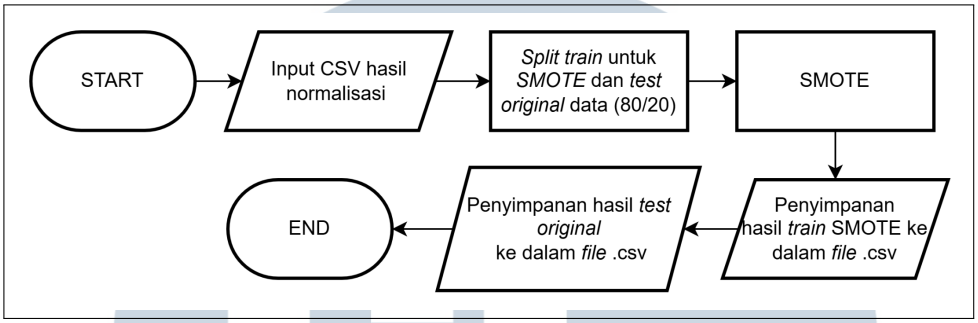
3.4.2 SMOTE

Pada tahap ini, dilakukan pemrosesan data untuk menangani ketidakseimbangan kelas menggunakan teknik SMOTE (Synthetic Minority Over-sampling Technique). Langkah pertama adalah menginput data CSV yang telah melalui tahap normalisasi sebelumnya. Data ini kemudian dibagi menjadi dua bagian, yaitu data untuk pelatihan (train) dan data untuk pengujian (test) dengan proporsi 80/20.

Selanjutnya, teknik SMOTE diterapkan pada data pelatihan (train) untuk mensintesis sampel dari kelas minoritas, sehingga distribusi kelas menjadi lebih seimbang. Dengan menggunakan SMOTE, data kelas minoritas diperbanyak dengan cara membuat contoh sintetik berdasarkan data yang ada.

Setelah proses SMOTE, hasil dari data test yang tidak diproses oleh SMOTE akan disimpan dalam file CSV terpisah, sementara data pelatihan yang telah diproses dengan SMOTE juga akan disimpan dalam file CSV untuk digunakan

dalam model selanjutnya. Gambar 3.8 menggambarkan alur proses penggunaan SMOTE dalam tahap ini.



Gambar 3.8. Proses SMOTE

3.5 Pengembangan Model dan Evaluasi

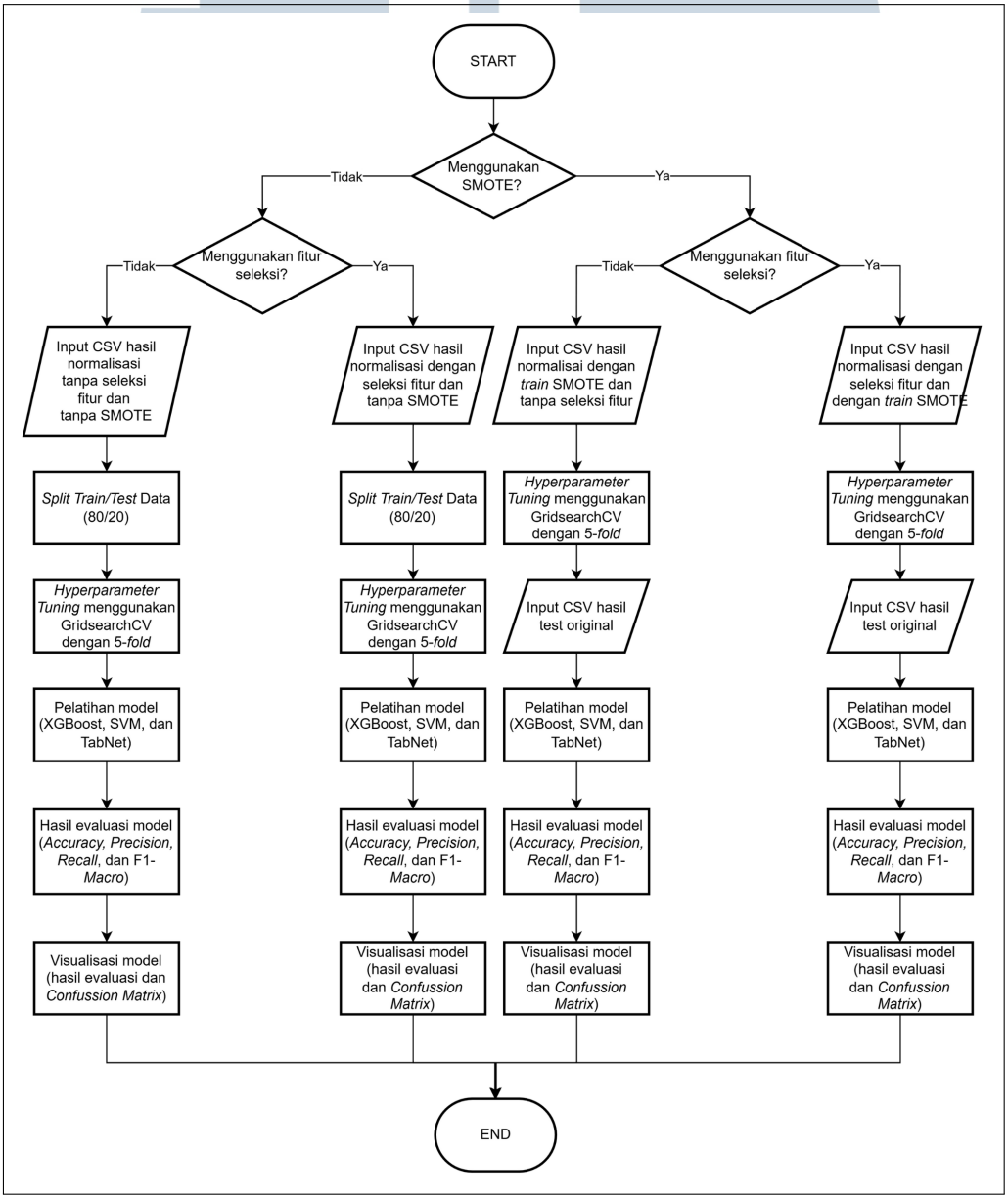
Pada tahap pengembangan model dan evaluasi, langkah pertama adalah menentukan apakah teknik SMOTE (Synthetic Minority Over-sampling Technique) akan digunakan atau tidak. Jika SMOTE digunakan, langkah berikutnya adalah memeriksa apakah seleksi fitur sudah dilakukan. Proses ini akan menentukan jalur alur berikutnya dalam pengembangan model.

1. Tidak menggunakan SMOTE:
    - Input data CSV hasil normalisasi tanpa seleksi fitur atau dengan seleksi fitur. Data tersebut kemudian dibagi menjadi data pelatihan (train) dan pengujian (test) dengan pembagian 80/20.
    - Setelah itu, dilakukan hyperparameter tuning menggunakan metode GridSearchCV dengan pembagian 5-fold untuk memilih parameter terbaik.
    - Model kemudian dilatih menggunakan algoritma seperti XGBoost, SVM, atau TabNet.
    - Hasil evaluasi model akan dievaluasi menggunakan metrik *accuracy, precision, recall, dan F1-macro*, dan dilanjutkan dengan visualisasi hasil evaluasi model serta matriks kebingungannya.
2. Menggunakan SMOTE:
    - Input data CSV hasil normalisasi dengan atau tanpa seleksi fitur dan lakukan proses yang serupa, yaitu membagi data menjadi pelatihan dan

pengujian, kemudian melanjutkan dengan hyperparameter tuning dan pelatihan model.

- Evaluasi model dilakukan dengan cara yang sama, menggunakan metrik dan visualisasi yang telah disebutkan.

Gambar 3.9 alur lengkap dari proses pengembangan model dan evaluasi yang melibatkan teknik SMOTE, seleksi fitur, pelatihan model, dan evaluasi kinerja model.



Gambar 3.9. Proses Pengembangan Model dan Evaluasi