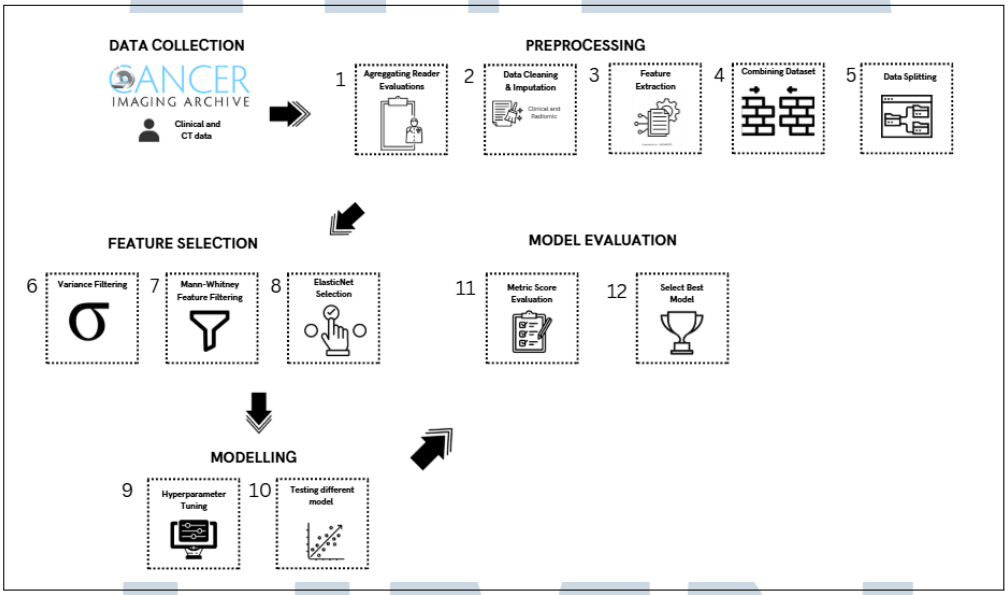
BAB 3
METODOLOGI PENELITIAN

Penelitian ini menggunakan rancangan metode sesuai dengan Gambar 3. Perangkat yang digunakan selama menjalankan proses adalah sebuah PC dengan spesifikasi sebagai berikut: sistem operasi Windows 10, prosesor Intel Core i7-8700K CPU @ 3.70GHz, GPU NVIDIA GTX 2080, dan RAM berkapasitas 32GB. Analisis data, ekstraksi fitur, serta pembuatan model dilakukan menggunakan editor kode Windsurf, dengan bahasa pemrograman Python versi 3.12 dan 3.7, serta *virtual environment* berbasis Anaconda.



Gambar 3.1. *Pipeline* Riset

Visualisasi dari alur penelitian dapat dilihat pada Gambar 3.1. Gambar tersebut menunjukkan *pipeline* dan alur eksperimen serta metodologi yang dilakukan secara garis besar, yaitu *data collection* atau pengumpulan data, *preprocessing* atau prapemrosesan data, *feature selection* atau seleksi fitur, *modelling* (pemodelan), dan *model evaluation* (evaluasi model).

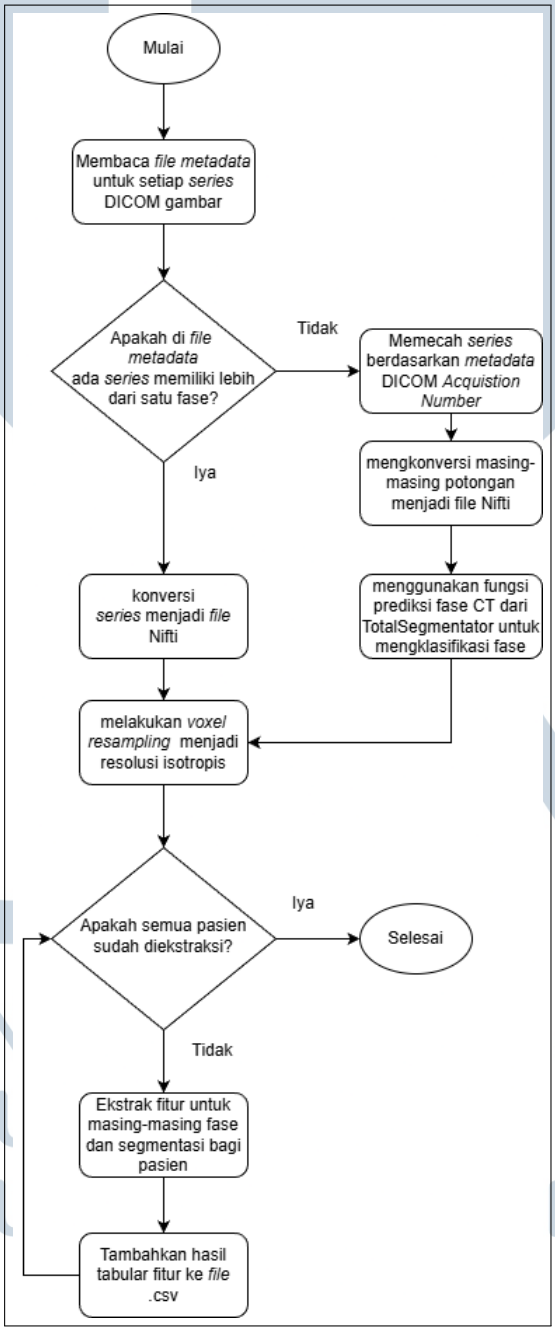
3.1 Pengumpulan Data

Data yang digunakan dalam penelitian ini merupakan koleksi HCC-TACE-SEG yang diperoleh dari situs *web* TCIA. Data yang digunakan terdiri atas data klinis dan citra pemindaian CT per fase dengan agen kontras dari 105 pasien, serta segmentasi tumor dan hepar dari citra tersebut. Data ini merupakan kompilasi pasien dari berbagai studi, sehingga bersifat *multi-study* dan *multi-center* [26]. Karena adanya kelalaian dari pihak kurator data, ditemukan 22 pasien yang memiliki fase CT campuran dalam satu *volume*. Seluruh hasil pemindaian disimpan dalam bentuk *folder* yang berisi puluhan hingga ratusan berkas DICOM, yaitu format berkas yang umum digunakan untuk

menyimpan metadata dari potongan *volume* CT. Karena tidak tersedianya keterangan mengenai regimen pengobatan untuk 17 pasien, maka pasien-pasien tersebut dieksklusikan dari penelitian.

3.2 Ekstraksi Fitur

Ekstraksi fitur merupakan proses di mana fitur-fitur radiomik kuantitatif dihitung dan diekstraksi dari citra CT. Alur dari proses ini dapat dilihat pada Gambar 3.2.



Gambar 3.2. *Flowchart* Proses Ekstraksi Fitur

Pertama, untuk memenuhi kebutuhan pustaka PyRadiomics versi 3.1.0, seluruh kumpulan berkas DICOM dikonversi menjadi berkas dengan format Neuroimaging Informatics Technology Initiative (NIfTI) menggunakan pustaka dicom2nifti versi 2.3.4. Konversi dilakukan dengan memproses setiap folder yang berisi berkas DICOM. Setiap berkas DICOM merepresentasikan potongan atau citra statis dari hasil akuisisi pemindaian CT.

Selanjutnya, untuk memastikan keseragaman posisi *voxel* antar *volume*, dilakukan *voxel resampling* untuk mengubah *spacing* antara *voxel* menjadi isotropik, yang didefinisikan sebagai 1 mm $\times$ 1 mm $\times$ 1 mm. Proses ini dilakukan menggunakan pustaka SimpleITK versi 2.5.2. Untuk memastikan bahwa fitur radiomik hanya dihitung pada jaringan *soft tissue*, nilai HU dari *pixel array* dilakukan proses *clipping* dengan batas bawah -100 HU dan batas atas 300 HU.

Seluruh informasi mengenai fase *volume*, lokasi *volume* dalam sistem, serta *metadata* lainnya disimpan dalam sebuah berkas *metadata.xlsx*. Data dari berkas ini dibaca pada saat proses ekstraksi dan digunakan untuk mengidentifikasi fase yang sedang diekstraksi serta lokasi berkas NIfTI yang akan diproses.

Dalam kasus ekstraksi fitur pada pasien dengan *volume* citra multifase, pasien-pasien tersebut diekstraksi secara terpisah dari kelompok utama. Karena setiap *volume* multifase memiliki informasi fase yang tercampur, seluruh berkas DICOM dipisahkan dan *volume* dikelompokkan berdasarkan metadata *AcquisitionNumber*, yang dibaca menggunakan pustaka pydicom. Setiap kelompok DICOM kemudian dikonversi menjadi berkas NIfTI, dan fase dari masing-masing kelompok ditentukan menggunakan fitur deteksi fase pemindaian dari pustaka TotalSegmentator.

Ekstraksi fitur *first-order*, *second-order*, dan *third-order* dari setiap fase dilakukan menggunakan PyRadiomics. Proses ini dilakukan untuk setiap pasien, dan fitur radiomik ditambahkan secara iteratif ke dalam sebuah berkas *.csv* hingga seluruh pasien selesai diproses. Gambar 3.2 memberikan visualisasi alur proses ekstraksi fitur. Karena adanya kesalahan pada proses *alignment* segmentasi, jumlah pasien yang berhasil diekstraksi menjadi 75.

3.3 Prapemrosesan Data

Setelah proses ekstraksi, data yang diperoleh menjalani tahap prapemrosesan. Tahap ini bertujuan untuk mengubah data yang masih mentah dan kurang optimal menjadi format yang sesuai untuk proses pelatihan model.

3.3.1 Data Klinis

Dikarenakan data yang digunakan mencakup lebih dari satu tipe TACE yang diadministrasikan, prapemrosesan data klinis dilakukan dengan langkah-langkah sebagai berikut:

1. Mengagregasikan hasil evaluasi mRECIST dari tiga radiolog berdasarkan nilai mayoritas (*mode*), karena tidak terdapat kondisi yang memerlukan *tie-breaker*.
2. Membuat kolom biner bernama *responder* dengan nilai 1 jika pasien memiliki nilai mRECIST hasil mode lebih dari 2, dan nilai 0 untuk kondisi lainnya. Kolom ini digunakan sebagai indikator keberhasilan respons pengobatan.
3. Menghapus pasien yang tidak memiliki keterangan regimen pengobatan (nilai NA).

4. Melakukan *one-hot encoding* untuk membinarisasikan tipe kemoterapi yang diberikan, sehingga dihasilkan dua kolom, yaitu `chemo_doxorubicin_LC_beads` dan `chemo_Cisplatin_doxorubicin_Mitomycin-C`, yang merepresentasikan regimen DEB-TACE serta cTACE dengan Cisplatin, Doxorubicin, dan Mitomycin-C.
5. Mengambil kolom regimen pengobatan, kolom *responder*, dan kolom identifikasi pasien (TCIA_ID), kemudian menggabungkannya dengan dataset fitur radiomik.

3.3.2 Data Radiomik

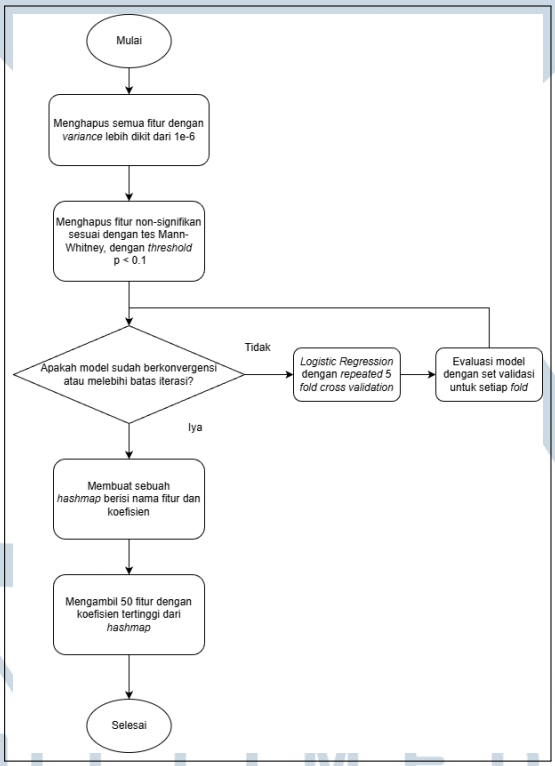
Prapemrosesan fitur radiomik dilakukan melalui tahapan berikut:

1. Menghapus fitur radiomik pada fase *delayed*, karena sebagian besar pasien tidak memiliki pemindaian CT dengan kontras fase tersebut.
2. Melakukan standardisasi fitur agar fitur dengan nilai rata-rata besar tidak mendominasi proses pemodelan.

Data klinis dan data radiomik kemudian digabungkan menjadi satu tabel, dan pembagian data untuk proses *training* dan *testing* dilakukan dengan rasio 70:30 [27], serta stratifikasi pasien dengan rasio 2 *non-responder* : 1 *responder* pada setiap pembagian data.

3.4 Seleksi Fitur

Seleksi fitur dilakukan untuk memperoleh *pool* fitur optimal dalam pelatihan model, menggunakan skema tiga langkah. Alur proses seleksi fitur ditunjukkan pada Gambar 3.3.



Gambar 3.3. *Flowchart* Proses Seleksi Fitur

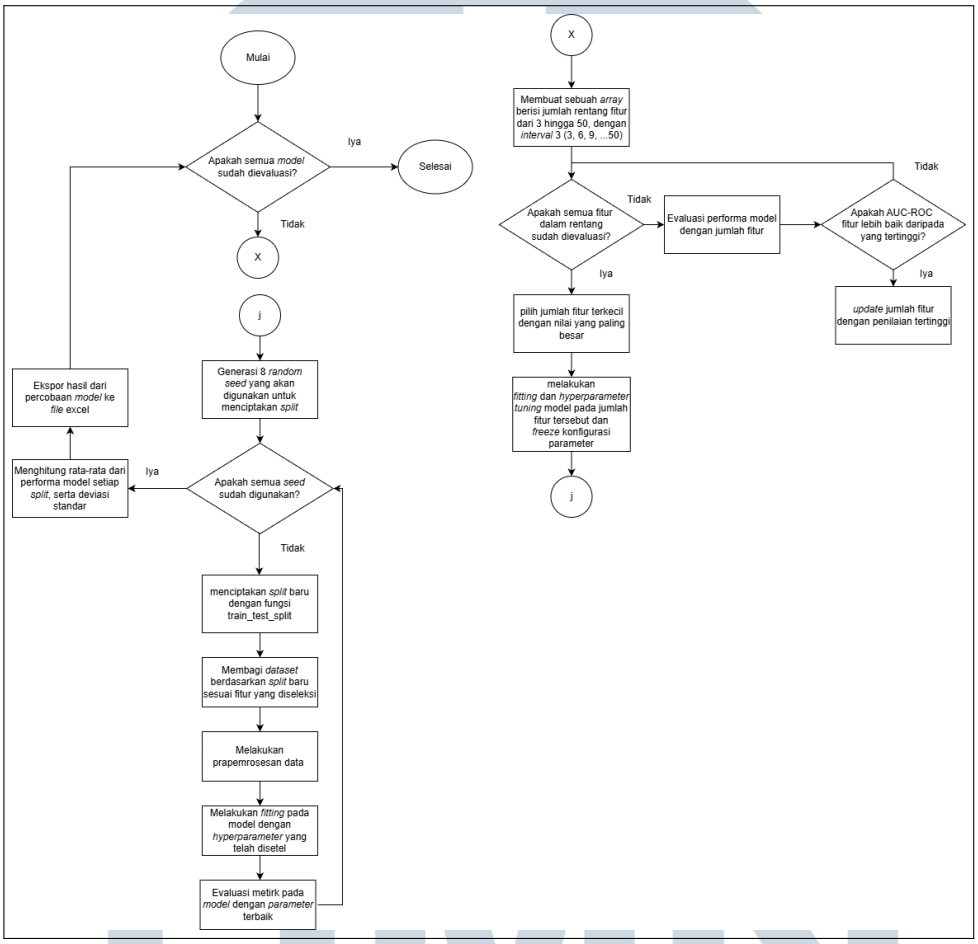
Pertama, dilakukan penyaringan awal dengan menghapus fitur yang memiliki nilai *variance* di bawah $1e^{-6}$ menggunakan teknik *variance thresholding*. Selanjutnya, dilakukan penyaringan lanjutan menggunakan uji statistik nonparametrik Mann–Whitney U-Test dengan *threshold* $p < 0.1$, karena nilai $p < 0.05$ dinilai terlalu restriktif dan tidak menghasilkan *pool* fitur yang memadai.

Setelah itu, penalti ElasticNet pada model *Logistic Regression* digunakan untuk memangkas fitur-fitur yang tersisa dari hasil uji Mann-Whitney.

Untuk melakukan seleksi, model *Logistic Regression* melalui proses *grid search* dengan *repeated five-fold cross-validation* guna menghasilkan koefisien fitur yang stabil selama proses *fitting*. Pada akhir proses, sebuah *hashmap* yang berisi koefisien masing-masing fitur dikonstruksi dan diurutkan berdasarkan nilai koefisien secara *descending*. Jumlah fitur optimal ditentukan secara *heuristic* dan eksperimental dengan mempertimbangkan *trade-off* antara kebutuhan memori, waktu komputasi, dan presisi.

3.5 Proses Pemilihan Model

Proses pemilihan model mencakup penentuan fitur dan *hyperparameter* optimal untuk setiap model. Alur dari proses ini ditunjukkan pada Gambar 3.4.



Gambar 3.4. *Flowchart* Proses Pemodelan

Untuk setiap model, teknik *grid search* digunakan untuk mengevaluasi kombinasi parameter dengan *repeated stratified k-fold cross-validation*. Jumlah fitur yang optimum bagi setiap model, ditentukan menggunakan penambahan fitur secara sekuensial. Jumlah fitur divariasikan dari 3 hingga 50 dengan *interval* 3. Lalu, nilai AUC-ROC akan dievaluasi pada setiap jumlah fitur dalam rentang (dengan *hyperparameter tuning*), dan jumlah fitur terkecil dengan nilai AUC-ROC terbaik akan dipakai sebagai jumlah fitur optimum untuk model tersebut. *Hyperparameter* dan jumlah fitur terbaik bagi model akan di-*freeze* dan digunakan untuk evaluasi lanjut.

Untuk memastikan bahwa evaluasi metrik pembagian pasien awal tidak bersifat kebetulan (misalnya seluruh pasien dengan sinyal kuat berada pada *training set*), serta memastikan bahwa kumpulan fitur dan model yang dibangun bersifat informatif, pengujian model dilakukan pada 10 pembagian data yang berbeda dengan mengubah parameter *random_seed* pada fungsi *train_test_split* dari pustaka *scikit-learn*. Model dengan performa metrik tertinggi pada *cross-validation* kemudian

dievaluasi pada *test set*.

3.6 Evaluasi Model

Model yang terpilih dievaluasi dengan beberapa metrik yang disesuaikan untuk kondisi ketidakseimbangan kelas yang berada di setiap *split* data, untuk memfokuskan penilaian pada harmonisasi dan kemampuan model untuk melakukan *ranking* kasus secara tepat. Metrik yang digunakan untuk mengevaluasi adalah AUC-ROC, AUC-PRC, Akurasi, dan Nilai F1.