

BAB 3

METODE PENELITIAN

Bab ini menguraikan secara komprehensif desain, strategi, dan prosedur teknis yang diterapkan dalam penelitian untuk mengembangkan sistem prediksi status invasi otot pada kanker kandung kemih (*Bladder Cancer* - BCa) menggunakan pendekatan radiomics berbasis *Machine Learning*. Metodologi ini dirancang secara sistematis untuk menjembatani kesenjangan antara citra medis kualitatif dan diagnosis patologis kuantitatif melalui transformasi data citra *Magnetic Resonance Imaging (MRI)* menjadi fitur numerik. Struktur bab ini disusun mengikuti alur logika ilmiah, dimulai dari definisi metode, tahapan operasional, teknik pengumpulan data, hingga evaluasi model yang menggunakan algoritma *XGBoost*, *Support Vector Machine (SVM)*, dan *Neural Network*. Gambar 3.1 menjelaskan keseluruhan alur penelitian yang akan dilakukan, dari pengumpulan data hingga evaluasi kinerja model.



Gambar 3.1. Alur Penelitian

3.1 Pengumpulan Data

Penelitian ini menggunakan FedBCa (*Federated Bladder Cancer Dataset*), sebuah inisiatif data terbuka yang dirancang khusus untuk mengevaluasi algoritma *Federated Learning* dan analisis citra medis pada kanker kandung kemih. Dataset ini juga merepresentasikan tantangan dalam analisis citra medis, yaitu heterogenitas data akibat penggunaan perangkat yang berbeda-beda. Proses Pengumpulan data bisa dilihat pada Gambar 3.2.



Gambar 3.2. Pengumpulan Data

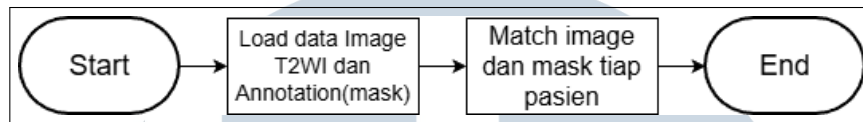
Dataset yang digunakan dapat diakses pada *link* berikut <https://zenodo.org/records/10409145> dimana disediakan sebuah file .rar yang berisi folder dari masing-masing *Medical Center* dari *Center* satu sampai *Center* empat. Masing-masing folder *Center* memiliki dua *folder* yaitu *T2WI* yang berisi gambar *MRI* dari tiap pasien di *Medical Center* tersebut, *Annotation* yang berisi hasil mask atau segmentasi tumor dari masing2 pasien juga dan satu file dengan *extension XLSX* yang berisi *label* masing-masing pasien dimana, *label* nol berarti pasien tersebut merupakan pasien *NMIBC* (*Non Muscle Invasive Bladder Cancer*) dan *label* satu berarti pasien merupakan pasien *MIBC* (*Muscle Invasice Bladder Cancer*) dengan nama "*CenterX_label*" dimana *X* merupakan angka tiap *Medical Center*.

Dataset ini diambil dari beberapa tempat antara *Southern Medical University* (*Center* satu), *Sun Yat-Sen University Cancer Centre* (*Center* dua), *Zhuhai Hospital affiliated with the Jinan University* (*Center* tiga), dan *Fifth Affiliated Hospital of Sun Yat-Sen University* (*Center* empat) pada periode antara November 2019 hingga Juli 2022. Tapi pada penelitian kali ini yang di pakai hanya *Center* dua sampai empat saja dikarenakan adanya ketidakseimbangan distribusi *label* pada *Center* satu. Dari *Center* dua sampai empat didapatkan total 102 pasien.

3.2 Pre-processing Data

Setelah mendownload dataset *FedBCa*, maka akan dilakukan pengekstrakan file .rar untuk mendapatkan *folder T2WI* dan *Annotation* setelah itu akan dilakukan

pencocokan antara *T2WI* dan *Annotation* untuk masing-masing pasien. Proses pada tahap ini bisa di lihat pada Gambar 3.3.



Gambar 3.3. Preprocessing Data

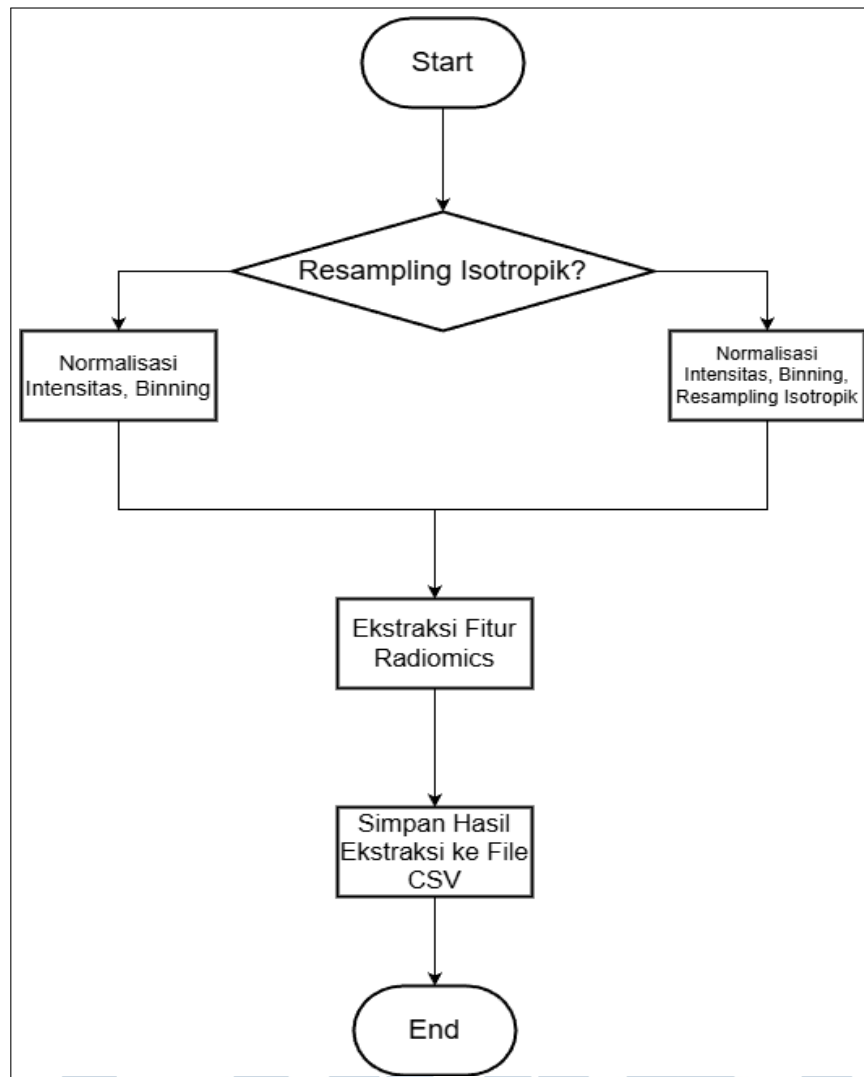
Dari total 102 pasien berpasangan dengan masing-masing *annotation*, enam pasien dari *Center* tiga dengan (*ID* : Tujuh(07), 10, 14, 21, 22, 23) tidak dapat diproses dikarenakan ketidaksesuaian letak geometri antara *image MRI T2WI* dan *ROI mask (Annotation)*. Ketidaksesuaian ini meliputi perbedaan ukuran dimensi, resolusi spasial (*spacing*), dan posisi koordinat antara *T2WI* dan *Annotation*.

Meskipun telah dilakukan upaya koreksi otomatis menggunakan *resampling*, *PyRadiomics* tetap menolak *file-file* tersebut karena standar validasi geometri yang ketat. Untuk menjaga kualitas data, enam sampel ini di-*exclude* sehingga total data akhir yang akan digunakan untuk penelitian ini berjumlah 96 Pasien dari *Center* dua sampai *Center* empat.

3.3 Ekstraksi Fitur Radiomics

Pada tahap ini dilakukan proses ekstraksi fitur *radiomics* dari citra *MRI T2-Weighted Image (T2WI)* yang telah melalui proses *preprocessing*. Ekstraksi fitur bertujuan mendapatkan informasi kuantitatif terkait karakteristik intensitas, tekstur, serta bentuk jaringan tumor yang tidak dapat diamati secara langsung secara visual. Pada penelitian ini, ekstraksi fitur dilakukan menggunakan library *PyRadiomics*.

U N I V E R S I T A S
M U L T I M E D I A
N U S A N T A R A



Gambar 3.4. Proses Ekstraksi Fitur Radiomics

Seperti ditunjukkan pada Gambar 3.4, proses ekstraksi fitur dilakukan melalui dua jalur, yaitu *tanpa resampling* dan dengan *resampling* isotropik. Pemisahan jalur ini dilakukan untuk mengatasi permasalahan heterogenitas citra medis yang umum ditemukan pada *dataset* multi-sumber.

3.3.1 Ekstraksi Fitur Tanpa Resampling

Pada jalur pertama, citra *MRI T2WI* diproses tanpa dilakukan *resampling*. Tahapan yang dilakukan meliputi:

1. Normalisasi intensitas, yaitu penyetaraan skala nilai intensitas untuk mengurangi variasi nilai piksel antar citra.

2. *Binning*, digunakan untuk membatasi jumlah level intensitas ke dalam rentang tertentu sehingga perhitungan histogram menjadi lebih stabil.
3. Ekstraksi fitur *radiomics* menggunakan *PyRadiomics*, meliputi beberapa kategori fitur seperti:
 - A) *Shape Features*
 - B) *First Order Statistics*
 - C) *Gray Level Co-occurrence Matrix (GLCM)*
 - D) *Gray Level Run Length Matrix (GLRLM)*
 - E) *Gray Level Size Zone Matrix (GLSZM)*
 - F) *Gray Level Dependence Matrix (GLDM)*
 - G) *Neighborhood Gray Tone Difference Matrix (NGTDM)*

Hasil fitur yang diperoleh kemudian disimpan dalam file CSV sebagai *dataset* numerik untuk proses pelatihan model *machine learning* pada tahap selanjutnya.

3.3.2 Ekstraksi Fitur Dengan *Resampling*

Pada jalur kedua, citra *MRI T2WI* diproses dengan *resampling isotropik*, yaitu mengubah ukuran *voxel* menjadi dimensi yang sama (misalnya $1 \times 1 \times 1$ mm). *Resampling* bertujuan untuk meningkatkan konsistensi spasial antar citra, terutama jika berasal dari perangkat *MRI* dan parameter pemindaian yang berbeda.

Tahapan pada jalur ini meliputi:

1. *Resampling* isotropik
2. Normalisasi intensitas dan *binning*
3. Ekstraksi fitur *radiomics*
4. Penyimpanan hasil ekstraksi ke dalam *file* CSV

Langkah ini menghasilkan dataset fitur alternatif yang dapat dibandingkan dengan hasil jalur tanpa *resampling*.

3.3.3 Output Ekstraksi Fitur

Output dari proses ekstraksi fitur berupa dua file CSV:

1. *Dataset* fitur tanpa *resampling*
2. *Dataset* fitur dengan *resampling*

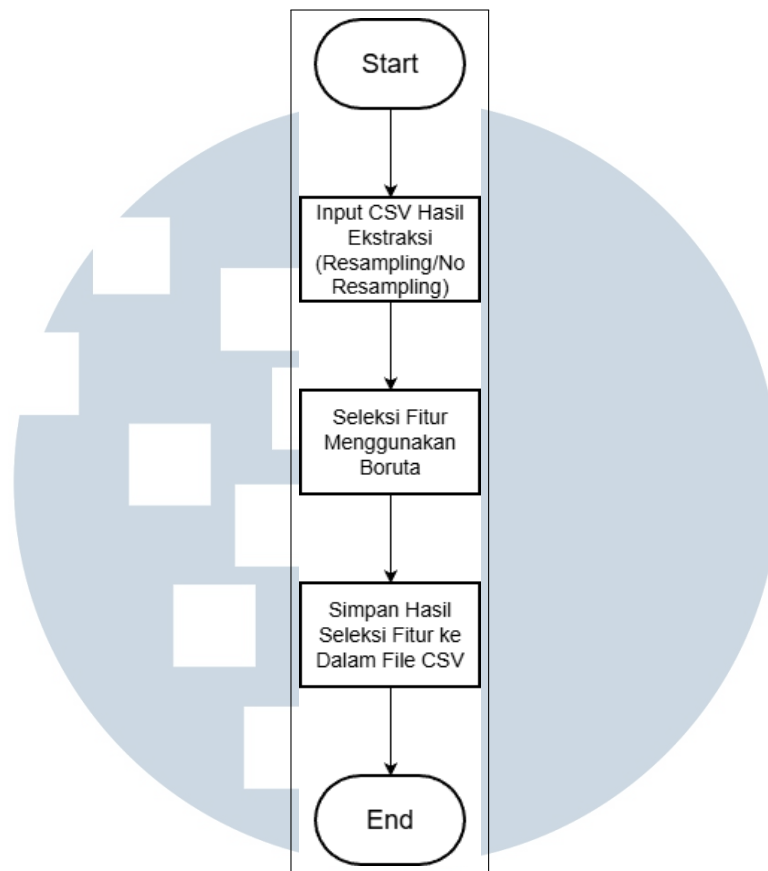
Kedua dataset tersebut digunakan untuk analisis komparatif dan pelatihan model klasifikasi tumor otak.

3.4 Seleksi Fitur

Tahap ini bertujuan untuk memilih fitur-fitur yang paling relevan dan signifikan dalam proses klasifikasi, sehingga dapat mengurangi kompleksitas model serta mencegah *overfitting*. Seleksi fitur dilakukan untuk menyaring fitur yang informatif dan menghilangkan fitur yang tidak relevan atau redundan, terutama karena proses ekstraksi fitur *radiomics* menghasilkan ratusan hingga ribuan fitur pada setiap sampel citra.

Pada penelitian ini digunakan metode *Boruta*, yaitu teknik seleksi fitur berbasis *wrapper* yang memanfaatkan algoritma *Random Forest*. *Boruta* bekerja dengan membandingkan pentingnya setiap fitur asli terhadap fitur *shadow* (fitur acak sebagai pembanding). Fitur yang memiliki tingkat kepentingan lebih tinggi dibandingkan *shadow features* akan dipertahankan sebagai fitur penting.

U N I V E R S I T A S
M U L T I M E D I A
N U S A N T A R A



Gambar 3.5. Proses Seleksi Fitur Menggunakan *Boruta*

Seperti ditunjukkan pada Gambar 3.5, proses seleksi fitur dilakukan pada dua *dataset* yang berbeda, yaitu *dataset* hasil ekstraksi fitur tanpa *resampling* dan *dataset* hasil ekstraksi fitur dengan *resampling* isotropik. Penggunaan dua jalur seleksi fitur bertujuan untuk mengevaluasi pengaruh *resampling* terhadap konsistensi dan kualitas fitur terpilih. Pendekatan ini diperlukan karena *dataset radiomics* memiliki tingkat heterogenitas yang tinggi akibat perbedaan perangkat *MRI* dan protokol akuisisi, sehingga perbandingan kedua jalur memungkinkan penilaian stabilitas fitur yang diperoleh.

3.4.1 Tahapan Seleksi Fitur

Adapun tahapan seleksi fitur adalah sebagai berikut:

1. Menginput file CSV hasil ekstraksi fitur (tanpa *resampling* atau dengan *resampling*)
2. Menjalankan seleksi fitur menggunakan algoritma *Boruta*

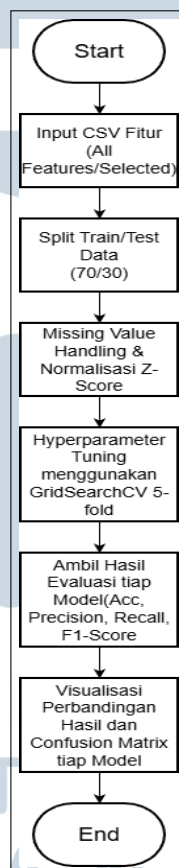
3. Menyimpan fitur terpilih ke dalam file CSV

Output dari proses ini berupa dua file CSV hasil seleksi fitur, yaitu:

- A) Hasil seleksi fitur tanpa *resampling*
- B) Hasil seleksi fitur dengan *resampling*

3.5 Pembangunan dan Evaluasi Model

Tahap ini merupakan proses pembangunan model klasifikasi berdasarkan fitur *radiomics* yang telah diperoleh dari tahap ekstraksi dan seleksi fitur. Tujuan tahap ini adalah untuk mengevaluasi performa model dalam membedakan kelas tumor berdasarkan fitur yang tersedia, serta menganalisis pengaruh seleksi fitur *Boruta* terhadap performa model.



Gambar 3.6. Flowchart Pembangunan dan Evaluasi Model Klasifikasi

Seperti ditunjukkan pada Gambar 3.6, pembangunan model dilakukan melalui dua jalur, yaitu menggunakan seluruh fitur hasil ekstraksi tanpa seleksi dan

menggunakan fitur terpilih hasil seleksi *Boruta*. Pemisahan jalur ini bertujuan untuk menilai apakah reduksi fitur dapat meningkatkan performa model dan mengurangi risiko *overfitting*, mengingat dataset *radiomics* umumnya memiliki dimensi fitur yang sangat tinggi.

3.5.1 Tahapan Pembangunan Model

Adapun tahapan pembangunan model meliputi:

1. Menginput dataset CSV (tanpa seleksi / dengan seleksi *Boruta*)
2. *Missing Value Handling* untuk mengatasi nilai kosong pada fitur
3. Normalisasi menggunakan metode *Z-Score*
4. Pembagian *dataset* menjadi *Train* dan *Test* sebesar 70 banding 30
5. *Hyperparameter tuning* menggunakan *GridSearchCV* dengan *5-fold cross-validation*
6. Evaluasi hasil model menggunakan metrik *Accuracy*, *Precision*, *Recall*, dan *F1-Score*
7. Visualisasi perbandingan performa model dan *Confusion Matrix*

Output dari proses ini berupa:

- A) Perbandingan performa model antara fitur tanpa seleksi dan dengan seleksi *Boruta*
- B) *Confusion Matrix* untuk masing-masing model
- C) Grafik perbandingan *Accuracy*, *Precision*, *Recall*, dan *F1-Score*