

# BAB 1

## PENDAHULUAN

### 1.1 Latar Belakang Masalah

Kanker payudara merupakan salah satu kanker dengan angka kejadian tertinggi di dunia dan menjadi ancaman serius bagi kesehatan perempuan. Pada tahun 2022, sekitar 2,3 juta perempuan didiagnosis menderita kanker payudara dengan 670.000 kematian secara global [1]. Di Indonesia, tercatat 66.271 kasus baru atau sekitar 30,1% dari seluruh kasus kanker yang dilaporkan, sehingga kanker payudara menjadi kanker tersering sekaligus penyebab kematian tertinggi pada perempuan [2]. IARC (*International Agency for Research on Cancer*) memperkirakan jumlah kasus baru kanker payudara akan meningkat sekitar 46,5% pada tahun 2040 dibandingkan dengan tahun 2018 [3]. Deteksi dini memegang peran kunci karena kanker payudara stadium awal memiliki tingkat kelangsungan hidup lima tahun mencapai sekitar 90%, sedangkan pada stadium lanjut angka ini menurun tajam [4]. Oleh karena itu, upaya deteksi dan diagnosis dini sangat penting untuk meningkatkan harapan dan kualitas hidup pasien.

Mammografi merupakan modalitas skrining utama yang efektif untuk mendeteksi perubahan jaringan payudara bahkan lebih dari dua tahun sebelum gejala klinis muncul [5]. Citra mammografi memuat informasi detail mengenai kelenjar payudara, kulit, jaringan lemak, dan stroma yang berkaitan dengan risiko kanker payudara [6]. Namun, kompleksitas anatomi payudara, variasi kepadatan jaringan, serta perbedaan pengalaman radiolog dapat memengaruhi akurasi pembacaan, sehingga detail seperti mikrokalsifikasi berpotensi terlewat dan menimbulkan kesalahan diagnosis, terutama dalam membedakan lesi jinak (*benign*) dan ganas (*malignant*) [7]. Berbagai sistem CAD (*Computer-Aided Diagnosis*) berbasis *deep learning* telah dikembangkan sebagai solusi untuk pembacaan mammogram secara otomatis, tetapi masih terkendala kebutuhan data dalam jumlah besar, bersifat *black-box*, serta rentan terhadap *overfitting* ketika dataset terbatas [8]. Keterbatasan ini mendorong eksplorasi pendekatan alternatif yang lebih transparan dan *interpretable*, salah satunya adalah radiomik.

Radiomik (*radiomics*) merupakan metode komputasional non-invasif yang mengonversi citra medis menjadi data berdimensi tinggi dengan mengekstraksi fitur kuantitatif secara *high-throughput* dari area ROI (*Region of Interest*)

yang telah disegmentasi [9]. Dalam onkologi, radiomik mengekstrak fitur intensitas, bentuk, tekstur, dan pola *wavelet* dari berbagai modalitas pencitraan untuk mengungkap pola kompleks yang tidak terlihat secara visual, kemudian mengaitkannya dengan karakteristik biologis tumor dan luaran klinis [10]. Pendekatan ini berkontribusi pada karakterisasi tumor, pemantauan respons terapi, dan prediksi prognosis, termasuk kelangsungan hidup dan risiko kekambuhan pada berbagai jenis kanker [11]. Pada kanker payudara, radiomik mammografi dimanfaatkan untuk membedakan lesi jinak (*benign*) dan ganas (*malignant*), mengidentifikasi sub tipe molekuler, serta memprediksi luaran jangka panjang [12]. Ekstraksi fitur secara terstandar dapat dilakukan menggunakan PyRadiomics yang mengikuti pedoman IBSI (*Image Biomarker Standardisation Initiative*), sehingga meningkatkan *reproducibility* dan konsistensi antarstudi.

Meskipun menjanjikan, pendekatan radiomik menghadapi tantangan berupa dimensionalitas fitur yang sangat tinggi. Ekstraksi radiomik dari satu citra dapat menghasilkan ratusan hingga ribuan fitur sekaligus, sehingga menimbulkan masalah *curse of dimensionality* yang berpotensi menyebabkan *overfitting* dan menurunkan kemampuan generalisasi model. Kondisi ini membuat model menjadi kompleks, sulit diinterpretasikan, dan rentan terhadap *bias* jika tidak ditangani dengan baik. Oleh sebab itu, *feature selection* menjadi langkah prapemrosesan yang krusial untuk menentukan subset fitur yang paling informatif sekaligus menekan redundansi.

Dalam konteks ini, berbagai metode seleksi fitur umum digunakan untuk mengatasi dimensi tinggi pada data radiomik. mRMR (*Minimum Redundancy Maximum Relevance*) berfungsi sebagai metode *filter* yang memilih fitur dengan relevansi tinggi terhadap kelas target namun dengan redundansi minimal antarfitur berbasis *mutual information* [13, 14]. LASSO (*Least Absolute Shrinkage and Selection Operator*) adalah metode *embedded* yang memasukkan regularisasi  $L_1$  ke dalam model sehingga koefisien fitur yang kurang relevan terdorong menjadi nol [15, 16]. RFE (*Recursive Feature Elimination*) mewakili pendekatan *wrapper* yang secara iteratif menghilangkan fitur dengan kontribusi terendah hingga diperoleh subset yang dianggap optimal [17]. Di sisi lain, ETC (*Extra Trees Classifier*) dimanfaatkan sebagai metode *embedded* berbasis *ensemble* dengan membangun banyak pohon keputusan yang sangat teracak dan menghitung skor kepentingan fitur dari reduksi impuritas untuk mengidentifikasi prediktor yang paling informatif [18, 19].

Untuk memanfaatkan fitur radiomik yang telah melalui proses seleksi, algoritma seperti SVM (*Support Vector Machine*) dan XGBoost (*eXtreme Gradient*

*Boosting*) sering diterapkan karena kemampuannya dalam memodelkan hubungan kompleks antarfitur serta meningkatkan kinerja klasifikasi. SVM dikenal efektif pada *dataset* berdimensi tinggi karena dapat membentuk *hyperplane* pemisah yang optimal meskipun jumlah fitur lebih besar dibandingkan dengan jumlah sampel [20]. Di sisi lain, XGBoost unggul dalam menangkap pola nonlinier dan interaksi antarfitur melalui proses pembelajaran berbasis pohon yang dilakukan secara iteratif, serta dilengkapi dengan mekanisme regularisasi untuk mengurangi risiko *overfitting* [21]. Selain itu, penggunaan *stacking ensemble classifier* juga semakin populer karena mampu mengombinasikan prediksi dari berbagai algoritma untuk meningkatkan performa model [22]. Dalam pendekatan ini, beberapa model dasar (*base learners*) dengan karakteristik yang berbeda dilatih terlebih dahulu, kemudian prediksinya dikombinasikan oleh sebuah *meta-learner* untuk menghasilkan prediksi akhir. Dengan memanfaatkan keunggulan masing-masing model, *stacking* dapat menangkap pola yang lebih beragam dibandingkan dengan hanya menggunakan model tunggal [23].

Berdasarkan pemaparan sebelumnya, penelitian ini mengajukan pendekatan klasifikasi kanker payudara berbasis radiomik dengan memanfaatkan model *stacking ensemble*. Tahapan penelitian diawali dengan segmentasi ROI lesi pada citra mammogram, kemudian dilanjutkan dengan proses ekstraksi fitur radiomik menggunakan PyRadiomics. Fitur yang diperoleh kemudian diseleksi menggunakan empat metode *feature selection*, yaitu mRMR, LASSO, RFE, dan ETC, guna memperoleh subset fitur yang paling relevan. Subset tersebut selanjutnya digunakan sebagai input dalam arsitektur *stacking*, dengan SVM dan XGBoost bertindak sebagai *base learners*, sementara *Logistic Regression* berfungsi sebagai *meta-learner* untuk mengombinasikan probabilitas prediksi dari kedua model dasar. Seluruh pendekatan ini dievaluasi melalui pembagian data menjadi set pelatihan dan pengujian, serta penerapan *nested stratified 5-fold cross-validation* pada data pelatihan untuk proses optimasi hiperparameter. Rancangan tersebut diharapkan menghasilkan model klasifikasi kanker payudara yang akurat, lebih transparan, dan minim *bias* evaluasi, sehingga memberikan kontribusi nyata bagi pengembangan sistem CAD berbasis radiomik [24].

## 1.2 Rumusan Masalah

Berdasarkan latar belakang yang telah dijelaskan, rumusan masalah dalam penelitian ini dirumuskan sebagai berikut.

1. Bagaimana kinerja model klasifikasi kanker payudara yang dikembangkan menggunakan pendekatan *stacking ensemble*, dengan SVM dan XGBoost berperan sebagai *base learners*, sedangkan *Logistic Regression* digunakan sebagai *meta-learner*, berdasarkan fitur radiomik yang diekstraksi dari citra mammogram?
2. Apakah pendekatan *stacking ensemble* berbasis fitur radiomik mampu memberikan kinerja klasifikasi yang lebih unggul dibandingkan dengan penggunaan model SVM dan XGBoost secara terpisah?
3. Metode *feature selection* manakah di antara mRMR, LASSO, RFE, dan ETC yang menghasilkan subset fitur radiomik paling informatif untuk klasifikasi kanker payudara?

### 1.3 Batasan Permasalahan

Batasan dalam penelitian ini ditetapkan untuk memperjelas cakupan studi serta menjaga fokus agar tetap sesuai dengan tujuan yang telah ditentukan. Adapun batasan-batasan yang digunakan adalah sebagai berikut.

1. Dataset yang digunakan bersumber dari *The Cancer Imaging Archive* (TCIA), khususnya koleksi CBIS-DDSM, dengan jumlah total 1.566 pasien. Penelitian ini hanya memanfaatkan citra mammogram dua dimensi berjumlah 3.103 citra beserta *mask* ROI lesi yang memenuhi kriteria kelengkapan anotasi dan kualitas citra untuk proses ekstraksi fitur radiomik.
2. Data citra yang digunakan terbatas pada citra mammogram yang telah memiliki *mask* atau segmentasi ROI lesi teranotasi pada koleksi CBIS-DDSM. Proses segmentasi baru di luar anotasi tersebut tidak dibahas dalam penelitian ini.
3. Proses klasifikasi dalam penelitian ini dibatasi hanya pada dua kelas, yaitu *benign* (jinak) dan *malignant* (ganas), tanpa membedakan lebih lanjut sub tipe histologis maupun molekuler kanker payudara.
4. Fitur yang dimanfaatkan dalam penelitian ini berupa fitur radiomik yang diekstraksi dari ROI lesi pada citra mammogram menggunakan PyRadiomics, mencakup fitur *first-order*, *shape*, tekstur, serta *high-order*. Penelitian ini

tidak membahas penggunaan data klinis atau histopatologis lain di luar fitur radiomik tersebut.

5. Metode seleksi fitur dalam penelitian ini dibatasi pada empat pendekatan, yaitu *Minimum Redundancy Maximum Relevance* (mRMR), *Least Absolute Shrinkage and Selection Operator* (LASSO), *Recursive Feature Elimination* (RFE), dan *Extra-Trees Classifier* (ETC), sehingga metode lain di luar keempat pendekatan tersebut tidak dianalisis.
6. Model klasifikasi yang dikaji terbatas pada SVM, XGBoost, serta arsitektur *stacking ensemble*, dengan SVM dan XGBoost berperan sebagai *base learners*, sedangkan *Logistic Regression* digunakan sebagai *meta-learner*.
7. Skema pelatihan dan evaluasi model menggunakan pembagian data menjadi himpunan latih dan uji, yang dikombinasikan dengan *nested stratified 5-fold cross-validation* pada data pelatihan untuk penentuan hiperparameter. Kinerja model diukur menggunakan metrik ROC-AUC, akurasi, presisi, *recall*, dan F1-score.
8. Penelitian ini tidak mencakup aspek implementasi klinis secara langsung serta tidak melakukan validasi eksternal pada dataset atau institusi lain di luar CBIS-DDSM.

#### 1.4 Tujuan Penelitian

Tujuan yang ingin dicapai dalam penelitian ini adalah sebagai berikut.

1. Membangun dan mengevaluasi model klasifikasi kanker payudara berbasis fitur radiomik dari citra mammogram menggunakan pendekatan *stacking ensemble* dengan SVM dan XGBoost berperan sebagai *base learners*, sedangkan *Logistic Regression* digunakan sebagai *meta-learner*.
2. Menganalisis dan membandingkan performa model *stacking ensemble* berbasis fitur radiomik dengan model SVM dan XGBoost yang digunakan secara individual dalam tugas klasifikasi kanker payudara.
3. Mengidentifikasi metode *feature selection* yang paling efektif di antara mRMR, LASSO, RFE, dan ETC dalam menghasilkan subset fitur radiomik yang informatif untuk klasifikasi kanker payudara.

## 1.5 Manfaat Penelitian

Penelitian ini diharapkan dapat memberikan beberapa manfaat sebagai berikut.

1. Manfaat teoritis: memberikan kontribusi pada pengembangan bidang radiomik dan *machine learning*, khususnya dalam penerapan pendekatan *stacking ensemble* dalam klasifikasi kanker payudara berbasis fitur radiomik dari citra mammogram.
2. Manfaat metodologis: menyajikan gambaran terkait pengaruh serta efektivitas berbagai metode *feature selection* (mRMR, LASSO, RFE, dan ETC) terhadap kinerja model klasifikasi kanker payudara dalam kerangka *stacking ensemble*, dengan SVM dan XGBoost sebagai *base learners* serta *Logistic Regression* sebagai *meta-learner*.
3. Manfaat jangka panjang: menjadi landasan bagi pengembangan sistem *computer-aided diagnosis* (CAD) berbasis radiomik yang lebih akurat, lebih *interpretable*, dan lebih *robust*, sehingga berpotensi mendukung proses deteksi dan diagnosis dini kanker payudara di lingkungan klinis di masa mendatang.

## 1.6 Sistematika Penulisan

Sistematika penulisan laporan ini disusun sebagai berikut:

1. Bab 1 PENDAHULUAN  
Bab ini memuat bagian pendahuluan yang meliputi latar belakang, rumusan masalah, batasan penelitian, tujuan penelitian, manfaat penelitian, serta sistematika penulisan.
2. Bab 2 LANDASAN TEORI  
Bab ini menguraikan landasan teori yang mencakup teori-teori pendukung, tinjauan penelitian terdahulu yang relevan, serta konsep-konsep terkait kanker payudara, mammografi, radiomik, *feature selection*, dan metode klasifikasi yang digunakan.
3. Bab 3 METODOLOGI PENELITIAN  
Bab ini menjelaskan metodologi penelitian, meliputi deskripsi *dataset*,

tahapan prapemrosesan dan segmentasi citra, proses ekstraksi fitur radiomik, penerapan metode seleksi fitur (mRMR, LASSO, RFE, dan ETC), perancangan model SVM, XGBoost, dan *stacking ensemble*, skema pembagian data menjadi set pelatihan dan pengujian, penerapan *nested stratified 5-fold cross-validation* pada data pelatihan untuk optimasi hiperparameter, serta metode evaluasi kinerja model.

#### 4. Bab 4 HASIL DAN DISKUSI

Bab ini menyajikan hasil eksperimen beserta pembahasannya, yang meliputi hasil seleksi fitur, kinerja masing-masing model klasifikasi, perbandingan antara *stacking ensemble* dan model individual, analisis metrik evaluasi, serta interpretasi hasil klasifikasi berdasarkan skenario eksperimen yang diterapkan.

#### 5. Bab 5 SIMPULAN DAN SARAN

Bab ini berisi simpulan dari penelitian yang telah dilakukan serta saran untuk pengembangan penelitian selanjutnya.

