

BAB 1

PENDAHULUAN

1.1 Latar Belakang Masalah

Kanker paru-paru merupakan salah satu tantangan kesehatan masyarakat yang paling utama dan mendesak pada era modern. Menurut data GLOBOCAN 2022, kanker paru-paru mencatat tingkat insidensi tertinggi dan menjadi penyebab utama kematian akibat kanker di seluruh dunia, dengan lebih dari 2,48 juta kasus baru dan 1,82 juta kematian yang didokumentasikan pada tahun 2022 saja [1]. Besarnya angka ini menuntut perhatian klinis dan upaya penanganan yang lebih intensif [2]. Subtipe yang dominan, yaitu *Non-Small Cell Lung Cancer* (NSCLC), mencakup sekitar 85% dari seluruh kasus kanker paru-paru, sebuah klasifikasi yang meliputi adenokarsinoma paru, karsinoma sel skuamosa, dan karsinoma sel besar [3]. Mengingat tingginya jumlah kasus baru dan tingkat kematian yang berkaitan dengan kanker paru, khususnya pada kasus yang didorong oleh NSCLC, subtipe ini telah menjadi titik fokus utama dalam penelitian onkologi yang mendalam.

Stadium kanker paru ditentukan melalui sistem klasifikasi TNM. Penentuan dilakukan dengan mempertimbangkan tiga poin penting: seberapa besar tumor utamanya (T), apakah sudah menyebar ke kelenjar getah bening di sekitarnya (N), dan apakah kanker telah menyebar jauh ke bagian tubuh lain (M). Meskipun kriteria penentuan stadium mengalami pembaruan berkala melalui pedoman *American Joint Committee on Cancer* (AJCC), parameter N (*Node*) tetap menjadi komponen yang sangat relevan dan stabil dalam praktik klinis di seluruh iterasi AJCC [4]. Akibatnya, *dataset* publik historis tetap relevan untuk penelitian klasifikasi stadium N, terutama karena mayoritas studi memprioritaskan pencitraan *Computed Tomography* (CT) untuk deteksi kanker paru [5]. Namun, penilaian stadium N pada NSCLC secara visual dan manual melalui pemindaian CT menyajikan tantangan yang signifikan. Hal ini disebabkan oleh keterlibatan kelenjar getah bening sering kali samar di dalam morfologi tumor primer, yang menyebabkan variabilitas antar-pengamat yang tinggi dan waktu analisis yang lama [6], terutama dalam kasus yang melibatkan nodul kecil atau infiltrasi halus yang tidak dapat dibedakan dari tampilan semu hasil pemindaian [7]. Sebagai respons, radiomik telah muncul sebagai solusi inovatif, yang mampu mengekstraksi ratusan hingga ribuan fitur kuantitatif secara otomatis dari gambar CT, sehingga dapat menjelaskan pola tekstur

dan heterogenitas tumor yang tidak tertangkap oleh mata manusia [8].

Terlepas dari kemajuan ini, sebagian besar penelitian radiomik sebelumnya masih terbatas pada fitur *first-order* dan berbasis bentuk (*shape-based*), serta mengabaikan fitur tingkat tinggi (*high-order*). Metrik tingkat tinggi ini sangat krusial untuk menangkap kompleksitas mikro-tektur yang mengindikasikan stadium N pada NSCLC [9]. Selain itu, segmentasi presisi yang membedakan parenkim paru dari nodul paru sangat penting sebelum ekstraksi fitur dilakukan, tingkat kebulatan benjolan (*sphericity*) tersebut merupakan petunjuk utama dalam menentukan stadium penyebaran kanker di kelenjar getah bening [10]. Dalam konteks ini, alat *Artificial Intelligence* modern seperti TotalSegmentator telah divalidasi oleh banyak ahli radiologi, mencapai koefisien *Dice* hingga 0,943 untuk 104 struktur anatomi seperti paru-paru dan nodul paru, sehingga berfungsi sebagai referensi yang andal dalam studi ini [11].

Setelah segmentasi yang akurat, proses ekstraksi radiomik menghasilkan ribuan fitur kuantitatif dengan rentang nilai yang sangat beragam, sehingga normalisasi menjadi tahap yang penting sebelum proses pemodelan. Namun, beberapa penelitian masih menerapkan normalisasi pada seluruh *dataset* sebelum pembagian data (*data splitting*), yang berpotensi menimbulkan kebocoran data (*data leakage*) dan menghasilkan estimasi performa yang terlalu optimistis [12]. Selain itu, karakteristik *dataset* klinis yang umumnya memiliki distribusi kelas tidak seimbang menuntut penerapan teknik *resampling* yang tepat [13]. Berbagai penelitian menunjukkan bahwa pendekatan *hybrid resampling*, khususnya SMOTE-Tomek, mampu meningkatkan representasi kelas minoritas sekaligus mengurangi sampel yang ambigu sehingga menghasilkan performa klasifikasi yang lebih stabil [14]. Setelah proses prapemrosesan, fitur radiomik yang berjumlah sangat besar masih memerlukan seleksi fitur untuk menghilangkan redundansi dan mempertahankan fitur yang paling informatif [15]. Dalam konteks ini, *Least Absolute Shrinkage and Selection Operator* (LASSO) telah terbukti efektif dalam menyusutkan koefisien fitur yang kurang relevan hingga nol, sehingga mampu menurunkan kompleksitas model, mengurangi *noise*, serta meningkatkan kemampuan generalisasi [16, 17].

Meskipun proses prapemrosesan tersebut mampu meningkatkan kualitas data radiomik, pemodelan tetap menjadi tantangan karena karakteristik data radiomik yang berdimensi tinggi, heterogen, serta memiliki ukuran sampel yang relatif terbatas memerlukan algoritma yang mampu menangkap pola dari berbagai perspektif. Berbagai algoritma pembelajaran mesin menawarkan keunggulan

yang berbeda dalam memodelkan karakteristik data tersebut. *Support Vector Machine* (SVM) efektif untuk membangun *hyperplane* optimal pada ruang fitur yang kompleks [18]. *Random Forest* memiliki kemampuan yang baik dalam menghasilkan prediksi yang stabil melalui mekanisme *bagging* serta mampu menangani kompleksitas dan heterogenitas fitur secara efektif [19]. Sementara itu, XGBoost unggul dalam mempelajari hubungan non-linear dan interaksi antarfitur melalui mekanisme *gradient boosting* [20]. Perbedaan karakteristik pembelajaran tersebut menjadikan setiap algoritma mampu menangkap informasi yang berbeda dari data radiomik. Oleh karena itu, pendekatan *Stacking Ensemble* dipilih karena mampu mengintegrasikan prediksi dari beberapa *base learners* yang saling melengkapi ke dalam sebuah *meta-learner* [21]. Dengan memanfaatkan kekuatan masing-masing algoritma, *Stacking Ensemble* dapat menghasilkan model yang lebih stabil, memiliki kemampuan generalisasi yang lebih baik, serta lebih *robust* terhadap variasi data dibandingkan penggunaan satu algoritma secara terpisah [22, 23, 24]. Sebagai *meta-learner*, *Logistic Regression* dipilih karena mampu mempelajari kombinasi linier yang optimal dari probabilitas prediksi yang dihasilkan oleh setiap *base learner* [25]. Dengan kompleksitas model yang relatif rendah serta penerapan regularisasi, *Logistic Regression* dapat menggabungkan kontribusi dari setiap algoritma secara efektif, menghasilkan estimasi probabilitas yang terkalibrasi dengan baik, serta membentuk keputusan klasifikasi akhir yang lebih akurat dan konsisten.

Perkembangan terkini juga menunjukkan bahwa pendekatan *Stacking Ensemble* mulai diterapkan pada analisis radiomik kanker paru. Sebagai contoh, penelitian terbaru berhasil mengombinasikan fitur radiomik, *Stacking Ensemble*, dan *Explainable AI* (SHAP) untuk memprediksi *radiation pneumonitis* serta tingkat kelangsungan hidup pasien NSCLC [26]. Meskipun menunjukkan performa yang menjanjikan, penelitian tersebut masih berfokus pada prediksi prognosis pascaterapi, menggunakan skema evaluasi konvensional berupa pembagian data pelatihan-pengujian dan *cross-validation*, serta belum diarahkan untuk klasifikasi stadium N sebagai salah satu parameter penting dalam sistem TNM. Selain itu, hingga saat ini masih sangat terbatas penelitian yang secara bersamaan mengintegrasikan radiomik tingkat tinggi, *Stacking Ensemble*, evaluasi menggunakan *Nested Cross-Validation*, serta *Explainable AI* dalam satu kerangka kerja untuk klasifikasi stadium N pada NSCLC.

Dengan mempertimbangkan berbagai keterbatasan pada studi terdahulu, diusulkan kerangka kerja klasifikasi stadium N pada NSCLC yang lebih

komprehensif dan tangguh. Untuk mengatasi keterbatasan studi yang umumnya bergantung pada data pusat tunggal, dimanfaatkan *dataset* publik multisumber dari *The Cancer Imaging Archive* (TCIA) [27] dan *UCSC Xena Browser* [28], dengan total kohort sebanyak 88 pasien. Alur kerja yang dibangun mengintegrasikan segmentasi TotalSegmentator, ekstraksi fitur radiomik tingkat tinggi melalui PyRadiomics, serta penerapan LASSO pasca-pembagian data secara ketat. Pada tahap pemodelan, pendekatan *Stacking Ensemble* diimplementasikan dengan menggunakan *Support Vector Machine* (SVM), *Random Forest* (RF), dan *Extreme Gradient Boosting* (XGBoost) sebagai *base learners*, sementara *Logistic Regression* digunakan sebagai *meta-learner* untuk mempelajari kombinasi optimal dari probabilitas prediksi yang dihasilkan masing-masing *base learner*, sehingga mampu meningkatkan kemampuan generalisasi model. Untuk memperoleh estimasi performa yang lebih andal sekaligus mengurangi risiko *overfitting*, proses pelatihan dan evaluasi model dilakukan menggunakan *Nested Cross-Validation*, yang terdiri atas *5-fold outer cross-validation* yang diulang sebanyak 5 iterasi untuk evaluasi performa model, serta *3-fold inner cross-validation* untuk proses optimasi hiperparameter pada setiap *base learner*. Selain itu, untuk mengoptimalkan estimasi probabilitas pada klasifikasi stadium N yang memiliki distribusi kelas tidak seimbang, pendekatan *Stacking Ensemble* dipadukan dengan SMOTE-Tomek pada data pelatihan sehingga proses pembelajaran menjadi lebih *robust* terhadap ketidakseimbangan kelas. Sebagai kontribusi krusial untuk menjembatani tingginya kompleksitas komputasi model *Stacking Ensemble* dengan kebutuhan praktis di lapangan, diintegrasikan pula *Explainable AI* (SHAP). Langkah ini diharapkan mampu memberikan transparansi terhadap keputusan model, sehingga memfasilitasi interpretasi klinis dan menumbuhkan kepercayaan bagi para tenaga medis.

1.2 Rumusan Masalah

Berdasarkan latar belakang penelitian yang telah dipaparkan, maka rumusan masalah dalam penelitian ini adalah sebagai berikut:

1. Bagaimana performa model *stacking* dalam memprediksi stadium penyebaran kelenjar getah bening (*N-Stage*) dibandingkan dengan model tunggal konvensional?
2. Bagaimana kontribusi fitur-fitur radiomik tertentu terhadap keputusan model

dalam membedakan antara stadium N0, N1, dan N2 berdasarkan analisis *Explainable AI* (SHAP)?

1.3 Batasan Permasalahan

Batasan masalah dalam penelitian ini ditetapkan agar fokus penelitian tetap terarah, yaitu:

1. Data yang digunakan terbatas pada citra *CT Scan* dengan label diagnosis patologis stadium *N-Stage* (N0, N1, N2).
2. Kondisi *dataset* klinis yang dievaluasi memiliki distribusi observasi yang tidak proporsional atau mengalami ketidakseimbangan kelas (*class imbalance*) yang signifikan antar tingkatan stadium.
3. Fitur yang diekstraksi hanya mencakup fitur radiomik kuantitatif (bentuk, intensitas, dan tekstur) menggunakan pustaka PyRadiomics.
4. Algoritma dasar (*base learners*) yang digunakan dalam arsitektur *stacking* terbatas pada SVM, Random Forest, dan XGBoost dengan *meta-learner* berupa Regresi Logistik.
5. Proses segmentasi *Region of Interest* (ROI) pada citra CT Scan dibatasi hanya menggunakan segmentasi otomatis berbasis kecerdasan buatan (*Artificial Intelligence*) melalui pustaka TotalSegmentator.

1.4 Tujuan Penelitian

Tujuan yang hendak dicapai melalui penelitian ini adalah:

1. Mengembangkan dan mengevaluasi model *machine learning* berbasis *stacking* untuk klasifikasi stadium tumor *N-Stage*.
2. Memberikan interpretasi yang transparan terhadap fitur-fitur radiomik yang paling berpengaruh secara klinis menggunakan metode SHAP.

1.5 Manfaat Penelitian

Penelitian ini diharapkan dapat memberikan manfaat sebagai berikut:

1. Bagi Praktisi Medis: Sebagai alat bantu pendukung keputusan (*Decision Support System*) dalam penentuan stadium tumor secara non-invasif dan objektif.
2. Bagi Ilmu Pengetahuan: Memberikan kontribusi pada pengembangan literatur mengenai implementasi *Explainable AI* pada model ansambel heterogen dalam domain medis.
3. Bagi Akademisi: Menambah pemahaman mendalam mengenai penerapan teknik *machine learning* tingkat lanjut dan pemrosesan citra medis secara terapan.

1.6 Sistematika Penulisan

Sistematika penulisan laporan penelitian ini disusun untuk memberikan gambaran menyeluruh mengenai alur penelitian, sebagai berikut:

1. Bab 1 PENDAHULUAN
Berisikan latar belakang, rumusan masalah, batasan penelitian, tujuan, manfaat, dan sistematika penulisan laporan.
2. Bab 2 LANDASAN TEORI
Membahas landasan teori mengenai stadium tumor (*N-Stage*), radiomik, algoritma pendukung (SVM, RF, XGBoost), *Stacking Model*, serta konsep *Explainable AI* (SHAP).
3. Bab 3 METODOLOGI PENELITIAN
Menguraikan alur pengerjaan penelitian mulai dari pengumpulan data, pra-pemrosesan citra, ekstraksi fitur, hingga rancangan arsitektur *Nested Cross-Validation*.
4. Bab 4 HASIL DAN DISKUSI
Menyajikan hasil eksperimen, perbandingan performa antar model, visualisasi stabilitas fitur, serta analisis interpretasi model menggunakan SHAP.
5. Bab 5 SIMPULAN DAN SARAN
Berisikan ringkasan temuan utama dari hasil penelitian serta saran pengembangan untuk penelitian selanjutnya.