1. Pendahuluan

Kanker paru-paru adalah penyebab utama kematian terkait kanker di seluruh dunia dan membawa dampak sosial ekonomi yang signifikan bagi pasien, keluarga, dan masyarakat secara keseluruhan [1]. *Non-Small Cell Lung Carcinoma* (NSCLC) menyumbang sebagian besar tumor paru-paru [1]. Di antara *Non-Small Cell Lung Carcinoma*, *adenocarcinoma* adalah tipe histologis utama *lung carcinoma* di Taiwan (52,5%) [1]. Merokok merupakan faktor resiko utama untuk kanker paru-paru, meskipun ada faktor lain, seperti paparan lingkungan (misalnya, bahan kimia, agen fisik, dan radiasi), riwayat klinis penyakit paru-paru (misalnya, *bronchitis* kronis, *emfisema*, *pneumonia*, dan *tuberculosis*), riwayat tumor familial, atau diet juga dapat dikaitkan dengan perkembangan kanker paru-paru [1]. Di negara Barat 70% sampai 90% dari kanker paru-paru disebabkan oleh merokok, sedangkan di Taiwan hanya 7% dari kasus kanker paru-paru Wanita yang berhubungan dengan merokok [1]. Banyak gen (misalnya, TP53, EGFR, KRAS, PIK3CA, dan EML4-ALK) telah dilaporkan berhubungan dengan kanker paru-paru pada tidak pernah merokok, meskipun mekanisme molekuler NSCLC pada wanita tidak merokok masih belum jelas [1].

Pada tahun 2012, kanker paru-paru menyumbang 1,6 juta kematian dan 1,8 juta kasus baru [2]. Ini adalah jenis pembunuh kanker yang paling umum pada pria dan Wanita AS, dan menyebabkan lebih banyak kematian daripada gabungan kanker kolorektal, payudara, dan prostat [2]. Menurut ringkasan dari 10 tahun terakhir untuk diagnosis NSCLC kelenjar getah bening mediastinum menggunakan 18F-FDG PET/CT, sensitivitas median hanya 62% yang berarti sebagian besar metastasis dinilai negative palsu [3]. Untuk meningkatkan sensitivitas diagnosis NSCLC kelenjar getah bening mediastinum, diperlukan strategi klasifikasi yang lebih canggih dan algoritma pembelajaran mesin [3]. Dalam diagnosis kanker, klasifikasi berbagai jenis tumor sangat penting [4]. Prediksi akurat dari berbagai jenis tumor memungkinkan untuk pengobatan yang lebih baik dan meminimalkan toksisitas pada pasien [4]. Metode tradisional untuk mengatasi situasi ini terutama didasarkan pada karakteristik morfologi jaringan tumor [4]. Metode konvensional ini dilaporkan memiliki beberapa keterbatasan 7 diagnosis [4]. Untuk menganalisis masalah klasifikasi kanker menggunakan data ekspresi gen, pendekatan yang lebih sistematis telah dikembangkan [4].

Penggunaan pembelajaran mesin pada data ekspresi gen sangat diperlukan untuk mendeteksi adanya kanker paru-paru. Beberapa penelitian telah menggunakan metode tersebut, antara lain pada tahun 2015, Chen Yen dkk. melakukan kemoterapi adjuvant (ACT) untuk Non-Small Cell Lung Carcinoma dengan hasil akurasi klasifikasi adalah 65,71% [5]. Tahun 2021, Margarita Kirienko dkk. melakukan penelitian menggunakan 18FFDG PET/CT yang menghasilkan area di bawah kurva (AUC) sebesar 0,87 [6]. Tahun 2009, Peng Guan dkk. melakukan penelitian menggunakan Support Vector Machine (SVM) yang menghasilkan akurasi metode yang dimodifikasi meningkat dari 98,86% menjadi 100% pada set pelatihan dan dari 98,51% menjadi 99,06% pada set pengujian serta standar deviasi dari metode yang dimodifikasi meningkat dari 98,86% menjadi 100% pada set pelatihan dan dari 98,51% menjadi 99,06% pada set pengujian [7]. Tahun 2022, Jose Marcio Luna dkk. melakukan penelitian menggunakan pendekatan pembelajaran mesin mengasilkan 11,4% pasien mengalami esophagitis radiasi derajat 3 [8]. Tahun 2021, Nguyen Quoc Khanh Le dkk. melakukan penelitian menggunakan *genetic algoritm plus XGBoost classifier* menghasilkan akurasi 0,836 dan 0,86 [9].

Salah satu metode pembelajaran mesin yang biasa digunakan dalam tugas prediksi adalah metode ensemble [10]. Ensemble merupakan algoritma efektif yang menggabungkan semua algoritma pembelajaran untuk meningkatkan akurasi [10]. Keuntungan teknik algoritma ini yaitu dapat mengurangi masalah ukuran sampel yang kecil secara rata-rata dan menggabungkan dari model untuk mencegah overfitting dari data latih [10]. Oleh karena itu, metode ensemble menjanjikan untuk meningkatkan akurasi prediksi pada NSCLC [10]. Penelitian ini bertujuan untuk memprediksi NSCLC dengan menerapkan metode ensemble pada data *microarray*. Penulis menggunakan tiga metode ensemble untuk memprediksi NSCLC, yaitu *Random Forest*, *Adaptive Boosting*, dan *Extreme Gradient Boosting*.

Rumusan Masalah

Rumusan masalah dari tugas akhir ini yaitu:

a. Bagaimana pemilihan fitur berdasarkan parameter *chi-square*?
b. Bagaimana membangun model prediksi menggunakan *Ensemble*?
c. Bagaimana hasil kinerja metode *Ensemble*?

Tujuan

Tujuan dari tugas akhir ini adalah:

a. Melakukan pemilihan fitur berdasarkan parameter *chi-square*
b. Melakukan model prediksi menggunakan *Ensemble*
c. Mengukur performansi model prediksi *Ensemble*