BAB 1
PENDAHULUAN

1.1 Latar Belakang Masalah

Kanker merupakan salah satu penyakit yang menjadi perhatian utama dalam dunia kesehatan global. Kanker adalah penyakit serius yang dapat berakibat fatal, disebabkan oleh kerusakan genetik dan perubahan yang mengindikasikan adanya gangguan kesehatan dalam tubuh [1–3]. Sel-sel kanker merupakan pertumbuhan abnormal yang dapat muncul dimana saja dalam tubuh dan bisa menjadi sangat berbahaya. Penting untuk mendeteksi kanker sejak dini agar bisa menentukan langkah pengobatan yang efektif untuk menyembuhkannya [4–6].

Menurut data *Global Burden of Cancer* (GLOBOCAN), kanker menjadi salah satu penyebab kematian yang terbanyak diseluruh dunia [7–9]. Sekitar 10 juta kasus kematian terjadi pada tahun 2020 dan munculnya 19.3 juta kasus kanker baru [10, 11]. Walaupun jumlah kasus kanker payudara pada wanita telah melampaui kanker paru-paru, namun jika berdasarkan jumlah kasus kematian, kanker paru-paru masih menjadi penyebab utama kematian. [9, 12].

Kanker paru-paru menyumbang sekitar 18% dari total kasus kematian kanker secara keseluruhan [13, 14]. Merokok menjadi faktor utama penyebab kanker paru-paru di beberapa negara [15,16]. Jumlah perokok yang terus meningkat membuat risiko mengidap kanker paru-paru juga meningkat [17, 18]. Selain rokok, paparan polusi udara juga menjadi salah satu penyumbang risiko mengidap kanker paru-paru [19, 20].

Hal ini telah mendorong berbagai penelitian mengenai pengendalian kanker dan metode klasifikasi dini untuk mengambil tindakan pencegahan yang diperlukan dan dengan demikian mengurangi angka kematian akibat kanker paru-paru [21]. Klasifikasi tingkat risiko menjadi salah satu solusi dalam upaya memahami dan mengidentifikasi potensi risiko kesehatan yang mungkin dialami seseorang [22,23].

Salah satu algoritma yang digunakan untuk melalukan klasifikasi adalah *Random Forest* [1, 24]. Algoritma *Random Forest* membangun sejumlah pohon keputusan secara terpisah, lalu setiap pohon memberikan hasil keputusan [25–27]. Dari hasil setiap pohon dilakukan penggabungan hasil sehingga mencegah *overfitting* [28, 29]. Algoritma *Random Forest* juga dapat mengatasi masalah data dimensi tinggi (*high-dimensional data*) yang artinya algoritma ini mampu

menangani *dataset* yang memiliki banyak fitur atau atribut [30–32]. Algoritma *Random Forest* dapat digunakan dengan objek klasifkasi yang cukup kompleks, seperti prediksi kesehatan pasien *COVID-19* dengan *accuracy* 94% dan klasifikasi detak jantung antar pasien dengan *accuracy* 96% [33, 34].

Algoritma *Random Forest* memiliki *accuracy* 95%, *precision* sebesar 94%, *recall* sebesar 93%, dan *F1 score* sebesar 92% untuk melakukan prediksi risiko yang paling mempengaruhi kemungkinan seseorang mengidap kanker paru-paru berdasarkan kebiasaan dan gejala umum yang dialami menggunakan algoritma *Random Forest* [35]. Penelitan tersebut menggunakan *dataset* yang berjumlah 309 sample data dan 16 atribut. *Hyperparameter* yang digunakan adalah *maxDepth = 0, numIterations = 100, numFeatures = 0*.

Penelitian serupa lainnya adalah melakukan prediksi apakah seseorang memiliki risiko terkena kanker paru-paru berdasarkan kebiasaan dan gejala umum yang dialami menggunakan algoritma *Random Forest* dengan *dataset* yang berjumlah 309 sample data dan 16 atribut [36]. Hasil dari penelitian tersebut menunjukan tingkat *accuracy* sebesar 88%, *precision* sebesar 92%, *recall* sebesar 91%, dan *F1 score* sebesar 97%. *Hyperparameter* yang digunakan adalah *max_depth = 60, max_features = 'sqrt', n_estimators = 100*.

Penelitian serupa tentang klasifikasi risiko mengidap kanker paru-paru adalah menggunakan algoritma *Artificial Neural Network* untuk klasifikasi tingkat risiko mengidap kanker paru-paru dan menggunakan *dataset Cancer Patient Datasets* [37]. *Dataset* terdiri dari 1000 sample data dan 26 atribut. Penelitian tersebut menggunakan paramater tingkat risiko yang diklasifikasikan menjadi tingkat risiko rendah (*low*) dan tinggi (*high*). Hasil dari penelitian tersebut menunjukkan tingkat *accuracy* sebesar 91%, *precision* sebesar 92%, *recall* sebesar 91%, dan *F1 score* sebesar 91%.

Berdasarkan latar belakang yang dijelaskan, penelitian kali ini akan berfokus pada implementasi algoritma *Random Forest* untuk klasifikasi tingkat risiko mengidap kanker paru-paru menggunakan *dataset Cancer Patient Datasets*.

1.2 Rumusan Masalah

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

1. Bagaimana implementasi algoritma *Random Forest* untuk klasifikasi tingkat risiko mengidap kanker paru-paru?

2. Berapa tingkat *accuracy*, *precision*, *recall*, dan *F1 score* dalam klasifikasi tingkat risiko mengidap kanker paru-paru?

1.3 Batasan Permasalahan

Berdasarkan latar belakang yang telah dipaparkan diatas, maka batasan masalah yang didapat sebagai berikut:

1. *Dataset* yang digunakan adalah *Cancer Patient Datasets* yang berasal dari *Kaggle*.
2. Kelas yang digunakan dari atribut tingkat risiko mengidap kanker paru-paru adalah rendah (*low*), sedang (*medium*) dan tinggi (*high*).

1.4 Tujuan Penelitian

Berdasarkan latar belakang yang telah dipaparkan diatas, maka tujuan penelitian yang didapat sebagai berikut:

1. Mengimplementasi algoritma *Random Forest* untuk klasifikasi tingkat risiko kanker paru-paru.
2. Mengukur nilai *accuracy*, *precision*, *recall*, dan *F1 score* dari algoritma *Random Forest* untuk klasifikasi tingkat risiko mengidap kanker paru-paru.

1.5 Manfaat Penelitian

Berdasarkan latar belakang yang telah dipaparkan diatas, maka manfaat penelitian yang didapat sebagai berikut:

1. Model ini dapat dikembangkan lebih lanjut dan diimplementasikan ke dalam aplikasi di bidang kesehatan untuk membantu dokter dan tenaga medis dalam membuat keputusan yang lebih baik terkait dengan diagnosis, pengobatan, dan manajemen pasien kanker paru-paru.
2. Jika dikembangkan menjadi aplikasi, dapat digunakan oleh masyarakat untuk melakukan pencegahan dini terhadap risiko mengidap kanker paru-paru.

1.6 Sistematika Penulisan

Berisikan uraian singkat mengenai struktur isi penulisan laporan penelitian, dimulai dari Pendahuluan hingga Simpulan dan Saran.

Sistematika penulisan laporan adalah sebagai berikut:

- Bab 1 PENDAHULUAN
  Pada bab ini meliputi latar belakang masalah, rumusan masalah, batasan permasalahan, tujuan penelitian, manfaat penelitian, dan sistematika penulisan.

- Bab 2 LANDASAN TEORI
  Pada bab ini meliputi teori-teori maupun algoritma yang berkaitan dengan penelitian yang dilakukan antara lain penjelasaan mengenai Kanker Paru-Paru, *Ensemble Learning*, *Decision Tree*, *Random Forest*, dan *Confusion Matrix*.

- Bab 3 METODOLOGI PENELITIAN
  Pada bab ini berisi penjelasan mengenai metode penelitian yang digunakan, perancangan sistem, dan gambaran aplikasi yang dibangun.

- Bab 4 HASIL DAN DISKUSI
  Pada bab ini berisi *source code* dari model dan hasil pengujian dari aplikasi yang telah dibuat.

- Bab 5 KESIMPULAN DAN SARAN
  Pada bab ini berisi kesimpulan dari penelitian yang sudah dilakukan dan saran yang membangun untuk penelitian selanjutnya.