BAB I
PENDAHULUAN

1.1 Latar Belakang

Jenis kanker yang paling umum ditemukan pada populasi wanita di seluruh dunia adalah kanker payudara. Data pada tahun 2020 mengungkapkan bahwa secara global diketahui bahwa 2,3 juta perempuan menerima diagnosis kanker payudara yang menyebabkan 685.000 kematian. Pada akhir tahun yang sama, sebanyak 7,8 juta perempuan masih hidup setelah didiagnosis sebagai penderita kanker payudara dalam lima tahun terakhir [1]. Kasus kanker payudara di setiap wilayah di seluruh dunia bervariasi tetapi semuanya mengalami peningkatan. Berdasarkan tren morbiditas dan mortalitas terkait kanker payudara saat ini, diperkirakan bahwa pada tahun 2030, jumlah kasus dan kematian akibat kanker payudara akan mencapai 2,64 juta dan 1,7 juta [2]. Peluang dan probabilitas kelangsungan hidup dapat meningkat secara signifikan melalui diagnosis dini kanker payudara karena hal ini memungkinkan pasien untuk menerima perawatan klinis tepat waktu. Salah satu metode untuk mendeteksi kanker payudara adalah dengan mengklasifikasikan tumor. Pada kasus kanker payudara, tumor dibagi menjadi dua kategori yaitu tumor ganas dan tumor jinak. Tumor ganas memiliki kecenderungan untuk menyebar dengan tingkat yang lebih tinggi dibandingkan dengan tumor jinak [3].

Banyak penelitian telah dijalankan untuk mengembangkan model klasifikasi berdasarkan analisis data medis dalam usaha meningkatkan diagnosis kanker payudara dengan akurasi yang lebih tinggi. Dataset WDBC (*Wisconsin Diagnostic Breast Cancer*) adalah salah satu sumber data yang digunakan dalam penelitian tersebut. Dataset ini berisi beragam fitur yang menggambarkan karakteristik sel-sel yang ditemukan dalam sampel biopsi kanker payudara, digunakan untuk membedakan kelas antara kanker payudara yang bersifat jinak (*benign*) dan yang bersifat ganas (*malignant*) [4].

Akan tetapi, dataset tersebut memiliki jumlah fitur yang cukup besar, sehingga ada kebutuhan untuk menerapkan metode seleksi fitur yang efisien guna mengenali fitur-fitur yang paling berpengaruh untuk proses klasifikasi.

Recursive Feature Elimination (RFE) adalah salah satu teknik seleksi fitur yang menggunakan metode rekursif untuk menghilangkan fitur-fitur yang kurang relevan, sehingga diperoleh fitur terbaik untuk membangun model [5]. Penelitian [6] mengungkapkan dampak signifikan penggunaan RFE sebagai alat seleksi fitur dalam menganalisis sentimen e-wallet di Twitter. Metode *Support Vector Machine* (SVM) menghasilkan evaluasi dengan tingkat akurasi sebesar 74%, yang kemudian meningkat menjadi 81% ketika SVM digabungkan dengan RFE. Penelitian [7] juga menunjukkan pengaruh RFE dalam meningkatkan akurasi model prediksi penyakit hati dengan menggunakan algoritma *Artificial Neural Network* (ANN) dan *AdaBoost*. Model *AdaBoost* tanpa RFE mencapai akurasi sebesar 86.74%, sementara dengan RFE, akurasinya meningkat menjadi 89.15%. Demikian pula, model ANN tanpa RFE memiliki akurasi sebesar 90.36%, namun dengan RFE, akurasinya meningkat menjadi 92.77%. Tidak hanya itu, penelitian [8] membandingkan metode seleksi fitur RFE dengan *Principal Component Analysis* (PCA) dan *Kernel PCA* (KPCA) yang diaplikasikan untuk memilih fitur TfEn (*Time-frequency Entropy*) yang paling optimal. Hasilnya menunjukkan keunggulan RFE dibandingkan dengan PCA dan KPCA. Terlihat dari nilai rata-rata akurasi total fitur TfEn mentah sebesar 98.80%, yang meningkat menjadi 100% dengan penerapan RFE. Sementara itu, PCA dan KPCA masing-masing hanya mencapai 88.80% dan 82.40%. Berdasarkan tinjauan literatur tersebut, penelitian ini akan menerapkan metode RFE untuk meningkatkan akurasi model klasifikasi kanker payudara pada dataset WDBC.

Merujuk pada penelitian sebelumnya [9], dimana nilai akurasi klasifikasi dataset WDBC mencapai 92.228% menggunakan algoritma C4.5, penelitian ini bertujuan untuk mengeksplorasi metode alternatif dalam klasifikasi menggunakan *Support Vector Machine* (SVM). Pemilihan SVM

sebagai alternatif didasarkan pada penelitian lain yang menunjukkan bahwa SVM memiliki kinerja lebih unggul dibandingkan dengan algoritma C4.5. Sebagai contoh, penelitian [10] yang membandingkan algoritma C4.5 dan SVM dalam memprediksi ketepatan waktu kelulusan mahasiswa. Hasilnya menunjukkan tingkat presisi 81% dan tingkat akurasi 80% untuk Algoritma C4.5, sementara SVM lebih unggul dengan tingkat presisi 88% dan tingkat akurasi 85%. Penelitian [11] mengenai klasifikasi talenta karyawan dengan menggunakan algoritma C4.5, *K-nearest neighbors* (KNN), dan SVM menunjukkan bahwa metode SVM memiliki akurasi tertinggi sebesar 94.62%, diikuti oleh C4.5 dengan akurasi 93.87%, sementara KNN memiliki akurasi terendah, yaitu 87.37%. Penelitian [12] juga menunjukkan bahwa SVM mencapai performa terbaik dengan akurasi sebesar 90%, melebihi C4.5 yang mencapai 85% dan KNN dengan akurasi 88% dalam menentukan rata-rata kredit macet koperasi. Selanjutnya, penelitian [13] membandingkan kinerja algoritma SVM, C4.5, *Logistic Regression, Naive Bayes, Random Forest, XG Boost*, dan KNN dalam deteksi intrusi. Hasil eksperimen menunjukkan bahwa SVM secara signifikan melampaui algoritma-algoritma lainnya dalam hal akurasi, presisi, dan *recall*. Melihat hasil dari sejumlah penelitian tersebut, penelitian ini menerapkan metode seleksi fitur RFE yang diintegrasikan dengan metode klasifikasi SVM untuk mengoptimalkan kinerja dalam mengklasifikasikan kanker payudara pada dataset WDBC.

1.2 Rumusan Masalah

Penggunaan fitur atau atribut yang kurang relevan dapat menyebabkan kesalahan dalam klasifikasi kanker payudara sehingga berpotensi mengurangi akurasi hasil diagnosis.

1.3 Pertanyaan Penelitian

Pertanyaan penelitian yang bertujuan menjawab masalah yang telah dirumuskan sebelumnya diuraikan dalam poin-poin berikut:

1. Bagaimana pengaruh penggunaan RFE terhadap akurasi model SVM untuk klasifikasi kanker payudara?
2. Bagaimana perbandingan kinerja model SVM dengan seleksi fitur RFE dan model SVM tanpa seleksi fitur untuk klasifikasi kanker payudara?

1.4 Batasan Masalah

Agar penelitian sesuai dengan fokus masalah yang diidentifikasi, batasan penelitian diterapkan sebagai berikut:

1. Dataset yang digunakan berbentuk tabular dengan variabel fitur bertipe *continuous* dan variabel target bertipe *categorical*.
2. Seleksi fitur menggunakan metode RFE (*Recursive Feature Elimination*).
3. Klasifikasi menggunakan model SVM (*Support Vector Machine*).
4. Evaluasi model akan memanfaatkan metrik akurasi, presisi, *recall*, dan *F1-score*.

1.5 Tujuan Penelitian

Berdasarkan permasalahan yang telah diidentifikasi, tujuan dari penelitian ini dapat diuraikan sebagai berikut:

1. Menerapkan RFE pada model SVM untuk klasifikasi kanker payudara.
2. Mengukur kinerja model SVM menggunakan seleksi fitur RFE dan model SVM tanpa seleksi fitur.

1.6 Manfaat Penelitian

Berdasarkan urgensi penelitian, diharapkan penelitian ini dapat memberikan manfaat sebagai berikut:

1. Secara teoritis:
    a. Meningkatkan efisiensi dan kinerja model klasifikasi kanker payudara dengan meminimalkan dimensi data melalui seleksi fitur yang tepat.
    b. Memberikan wawasan tentang potensi penggunaan RFE dalam konteks medis dan bidang ilmu lainnya.
2. Secara praktis:
    a. Menyelesaikan syarat kelulusan penulis dalam menyelesaikan pendidikan di Institut Teknologi Telkom Purwokerto.
    b. Memberikan sumber referensi baru yang berguna untuk repositori Institut Teknologi Telkom Purwokerto.