

BAB I

PENDAHULUAN

1.1 Latar Belakang

Perkembangan teknologi informasi dan komputasi telah mendorong pemanfaatan machine learning di berbagai bidang, termasuk kesehatan. *Machine learning* digunakan untuk mendukung proses pengambilan keputusan melalui kemampuan mengenali pola dari data yang tersedia. Dalam bidang medis, penerapan machine learning dimanfaatkan untuk membangun model klasifikasi penyakit yang membantu tenaga kesehatan melakukan diagnosis secara lebih cepat, akurat, dan objektif. Model klasifikasi ini dikembangkan berdasarkan karakteristik data pasien sehingga dapat mengidentifikasi pola yang berkaitan dengan kondisi kesehatan tertentu (Kourou dkk., 2015).

Kanker payudara (breast cancer) merupakan salah satu penyakit yang menjadi perhatian global. Berdasarkan laporan *World Health Organization* (WHO), kanker payudara adalah jenis kanker dengan jumlah kasus tertinggi pada perempuan di dunia dan menjadi salah satu penyebab utama kematian akibat kanker. Deteksi dini dan diagnosis yang tepat sangat penting untuk meningkatkan peluang keberhasilan pengobatan serta menurunkan angka kematian. Oleh karena itu, berbagai penelitian telah dilakukan untuk mengembangkan model klasifikasi berbasis machine learning yang dapat mendukung proses diagnosis kanker payudara secara lebih efektif (WHO, 2026).

Breast Cancer Wisconsin Dataset (Diagnostic) merupakan salah satu dataset yang sering dijadikan acuan dalam penelitian klasifikasi kanker payudara. *Dataset* ini menjadi benchmark karena memiliki kualitas data yang baik dan sering digunakan untuk mengevaluasi berbagai algoritma klasifikasi serta teknik prapemrosesan data (Dua & Graff, 2019; Kourou dkk., 2015). *Dataset* ini terdiri dari 30 fitur numerik yang merepresentasikan karakteristik inti sel hasil analisis citra mikroskopis (Street dkk., 1993). Meskipun jumlah fiturnya tidak terlalu besar, beberapa fitur memiliki tingkat korelasi tinggi yang mengindikasikan adanya

redundansi informasi (Ibrahim dkk., 2021). Redundansi tersebut dapat meningkatkan kompleksitas model, memperbesar risiko overfitting, dan menurunkan kemampuan generalisasi model klasifikasi (Chandrashekar & Sahin, 2014; Guyon & De, 2003; Saeys dkk., 2007).

Pendekatan yang umum digunakan untuk mengatasi permasalahan tersebut meliputi feature selection dan feature extraction. Feature selection bertujuan memilih fitur yang paling relevan terhadap proses klasifikasi dengan menghilangkan fitur yang tidak memberikan kontribusi signifikan. Sementara itu, feature extraction mentransformasikan fitur asli menjadi representasi baru yang lebih ringkas tanpa kehilangan informasi penting. Penerapan kedua pendekatan ini dapat meningkatkan efisiensi komputasi, mengurangi redundansi data, dan mendukung peningkatan performa model klasifikasi (Guyon & De, 2003).

Principal Component Analysis (PCA) merupakan salah satu metode *feature extraction* yang paling banyak digunakan. PCA mentransformasikan sekumpulan fitur yang saling berkorelasi menjadi sejumlah komponen utama (*principal components*) yang ortogonal, sehingga dapat mereduksi dimensi data sambil mempertahankan sebagian besar variasi informasi yang ada (Jolliffe & Cadima, 2016). Metode ini telah diterapkan pada berbagai permasalahan klasifikasi, termasuk klasifikasi kanker payudara, karena mampu mengurangi kompleksitas data, mempercepat pelatihan model, dan meningkatkan efisiensi komputasi tanpa menurunkan performa klasifikasi secara signifikan (Ibrahim, dkk., 2021; Esen, dkk., 2024). Namun, PCA merupakan metode *unsupervised feature extraction* yang hanya mempertimbangkan variasi data tanpa memperhatikan hubungan fitur terhadap variabel target, sehingga komponen yang dipertahankan belum tentu merupakan fitur yang paling relevan untuk klasifikasi (Guyon & Elisseeff, 2003; Chandrashekar & Sahin, 2014).

Pendekatan *correlation-based feature selection* (CFS) banyak digunakan untuk mengurangi redundansi dengan menghilangkan fitur-fitur yang memiliki tingkat korelasi tinggi, sehingga dapat menghasilkan subset fitur yang lebih representatif dan meningkatkan kemampuan generalisasi model (Hall, 1999; Saeys dkk., 2007). Namun, sebagian besar penelitian masih menggunakan nilai ambang

(*correlation threshold*) yang bersifat statis, seperti 0,80 atau 0,90, tanpa mempertimbangkan karakteristik distribusi data pada setiap *dataset*. Penggunaan nilai ambang tetap dapat berpotensi menghilangkan fitur yang masih informatif atau mempertahankan fitur yang redundan, sehingga proses seleksi fitur menjadi kurang optimal (Bolón-Canedo dkk., 2013; Ibrahim dkk., 2021). Oleh karena itu, diperlukan mekanisme penentuan *correlation threshold* yang lebih adaptif agar proses eliminasi fitur redundan dapat disesuaikan dengan karakteristik data dan menghasilkan subset fitur yang lebih optimal untuk klasifikasi.

Beberapa penelitian menunjukkan bahwa pendekatan *hybrid feature selection*, yaitu mengombinasikan dua atau lebih metode seleksi atau ekstraksi fitur, dapat menghasilkan representasi data yang lebih baik dibandingkan penggunaan satu metode secara terpisah (Chandrashekar & Sahin, 2014).). Namun, kajian mengenai integrasi *Principal Component Analysis* (PCA) dengan *Adaptive Correlation Threshold* (ACT) sebagai framework sistematis untuk klasifikasi kanker payudara masih terbatas. Selain itu, evaluasi terhadap kualitas fitur hasil seleksi dan pengaruhnya terhadap kinerja model klasifikasi pada *Breast Cancer Wisconsin Dataset* juga masih perlu diteliti lebih lanjut.

Berdasarkan permasalahan tersebut, penelitian ini mengusulkan *framework hybrid feature selection* yang mengintegrasikan *Principal Component Analysis* (PCA) dengan *Adaptive Correlation Threshold* (ACT). Framework ini dirancang untuk mengombinasikan kemampuan PCA dalam mereduksi dimensi data dengan mekanisme ACT yang mengurangi redundansi fitur secara adaptif sesuai karakteristik dataset. Selanjutnya, performa framework akan dievaluasi berdasarkan kualitas fitur hasil seleksi dan kinerja model klasifikasi pada *Breast Cancer Wisconsin Dataset*. Hasil penelitian ini diharapkan dapat berkontribusi pada pengembangan *Framework Hybrid Feature Selection berbasis PCA–Adaptive Correlation Threshold (Hybrid PCA–ACT Framework)* yang lebih efektif sebagai tahap prapemrosesan data pada permasalahan klasifikasi di bidang *machine learning*, khususnya untuk klasifikasi kanker payudara.

1.2 Identifikasi Masalah

Berdasarkan latar belakang yang telah diuraikan, beberapa permasalahan yang dapat diidentifikasi dalam penelitian ini adalah sebagai berikut:

1. *Breast Cancer Wisconsin Dataset* memiliki sejumlah fitur yang saling berkorelasi tinggi dan redundan, sehingga dapat memengaruhi efektivitas proses pembelajaran pada model klasifikasi.
2. Penggunaan metode feature selection secara tunggal belum mampu mengoptimalkan proses reduksi dimensi sekaligus mempertahankan informasi penting yang diperlukan untuk klasifikasi kanker payudara
3. Belum terdapat *framework hybrid feature selection* yang mengintegrasikan *Principal Component Analysis* (PCA) dengan *Adaptive Correlation Threshold* (ACT) untuk mengatasi permasalahan redundansi fitur pada *Breast Cancer Wisconsin Dataset*.
4. Belum diketahui sejauh mana penerapan *framework hybrid feature selection* berbasis PCA dan ACT dapat meningkatkan kualitas hasil seleksi fitur serta performa model klasifikasi dibandingkan penggunaan seluruh fitur.
5. Diperlukan evaluasi terhadap *framework* yang diusulkan untuk mengetahui efektivitasnya berdasarkan kualitas fitur hasil seleksi dan kinerja model klasifikasi.

1.3 Batasan Masalah

Agar penelitian ini lebih terarah dan fokus, maka batasan masalah dalam penelitian ini adalah sebagai berikut:

1. *Dataset* yang digunakan adalah *Breast Cancer Wisconsin Dataset (Diagnostic)* yang diperoleh dari *UCI Machine Learning Repository*.
2. Penelitian difokuskan pada pengembangan *framework hybrid feature selection* yang mengintegrasikan *Principal Component Analysis* (PCA) dan *Adaptive Correlation Threshold* (ACT).
3. Proses *feature selection* hanya diterapkan pada tahap prapemrosesan data sebelum proses klasifikasi dilakukan.

4. Evaluasi *framework* dilakukan menggunakan model klasifikasi machine learning yang telah ditentukan pada penelitian.
5. Pengukuran performansi *framework* dilakukan berdasarkan kualitas fitur hasil seleksi dan kinerja model klasifikasi menggunakan metrik evaluasi seperti *accuracy*, *precision*, *recall*, *F1-score*, *ROC Curve* dan *AUC*.
6. Penelitian tidak membahas aspek implementasi sistem klinis, diagnosis medis secara langsung, maupun interpretasi hasil oleh tenaga kesehatan.

Penelitian difokuskan pada analisis efektivitas *framework hybrid feature selection* dan tidak membahas pengembangan algoritma baru di luar integrasi PCA dan *Adaptive Correlation Threshold*.

1.4 Rumusan Masalah

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

1. Bagaimana merancang *framework hybrid feature selection* yang mengintegrasikan *Principal Component Analysis* (PCA) dengan *Adaptive Correlation Threshold* (ACT) dalam proses seleksi fitur untuk klasifikasi kanker payudara menggunakan *Breast Cancer Wisconsin Dataset*?
2. Bagaimana mengukur performansi *framework hybrid feature selection* yang diusulkan berdasarkan kualitas hasil seleksi fitur dan kinerja model klasifikasi pada *Breast Cancer Wisconsin Dataset*?

1.5 Tujuan Penelitian

Tujuan dari penelitian ini adalah sebagai berikut:

1. Merancang *framework hybrid feature selection* yang mengintegrasikan *Principal Component Analysis* (PCA) dengan *Adaptive Correlation Threshold* (ACT) pada proses seleksi fitur untuk klasifikasi kanker payudara.

2. Mengukur performansi *framework hybrid feature selection* yang diusulkan berdasarkan kualitas fitur hasil seleksi dan kinerja model klasifikasi pada *Breast Cancer Wisconsin Dataset*.

1.6 Manfaat Penelitian

Penelitian ini diharapkan dapat memberikan beberapa manfaat sebagai berikut:

1. Manfaat Teoritis

Penelitian ini diharapkan dapat memberikan kontribusi terhadap pengembangan ilmu pengetahuan di bidang *machine learning*, khususnya pada pengembangan *framework hybrid feature selection* yang mengintegrasikan *Principal Component Analysis* (PCA) dan *Adaptive Correlation Threshold* (ACT). Selain itu, penelitian ini diharapkan dapat memperkaya kajian mengenai strategi seleksi fitur yang mampu mengurangi redundansi fitur sekaligus mempertahankan informasi penting untuk meningkatkan kualitas proses klasifikasi.

2. Manfaat Praktis

Penelitian ini diharapkan dapat menghasilkan *framework hybrid feature selection* yang mampu meningkatkan kualitas fitur hasil seleksi serta performa model klasifikasi pada *Breast Cancer Wisconsin Dataset*. Framework yang diusulkan diharapkan dapat menjadi alternatif pendekatan dalam membangun sistem klasifikasi penyakit yang lebih efisien, akurat, dan memiliki kemampuan generalisasi yang lebih baik, khususnya pada kasus diagnosis kanker payudara berbasis *machine learning*.

3. Manfaat Akademis

Penelitian ini diharapkan dapat menjadi referensi bagi peneliti selanjutnya dalam pengembangan metode *feature selection*, *dimensionality reduction*, maupun *hybrid feature selection*. Selain itu, penelitian ini dapat dijadikan sebagai dasar untuk penelitian

lanjutan yang menguji framework yang diusulkan pada algoritma klasifikasi, *dataset* medis, maupun domain aplikasi lainnya.

