

BAB 1

PENDAHULUN

1. Latar Belakang

Penyakit jantung merupakan salah satu penyebab utama kematian di dunia dan menjadi perhatian serius dalam bidang kesehatan global. Berdasarkan laporan *World Health Organization* (WHO), penyakit kardiovaskular menyebabkan lebih dari 17 juta kematian setiap tahun dan menyumbang sekitar sepertiga dari total kematian global. Kondisi ini menunjukkan urgensi pengembangan sistem deteksi dini yang akurat dan efisien guna menekan angka mortalitas akibat penyakit jantung (Islam, 2024 [1]). Perkembangan teknologi komputasi dan kecerdasan buatan telah mendorong pemanfaatan *machine learning* dalam bidang medis, khususnya untuk klasifikasi dan prediksi penyakit jantung. Berbagai algoritma telah digunakan seperti *support vector machine* (SVM), *naïve bayes*, *decision tree*, *k-nearest neighbor* (KNN), dan *random forest*. Di antara algoritma tersebut, *random forest* dikenal memiliki performa yang stabil, tahan terhadap *overfitting*, serta mampu menangani data berdimensi tinggi dan non-linear secara efektif (Rajendran & Karthi, 2022 [2]).

Random forest merupakan metode *ensemble learning* berbasis *decision tree* yang bekerja dengan membangun banyak pohon keputusan dan menghasilkan keputusan akhir berdasarkan *voting* mayoritas. Keunggulan *random forest* dalam klasifikasi data medis telah dibuktikan dalam berbagai penelitian, dengan tingkat akurasi yang relatif tinggi dibandingkan metode tunggal (Ullah, 2024 [3]). Namun demikian, performa algoritma klasifikasi sangat bergantung pada kualitas fitur yang digunakan. Salah satu permasalahan utama dalam klasifikasi penyakit jantung adalah adanya fitur yang redundan, tidak relevan, atau memiliki korelasi tinggi antar fitur. Kondisi ini dapat menurunkan akurasi model, meningkatkan kompleksitas komputasi, serta memperbesar risiko *overfitting* (Jain & Singh, 2018 [4]). Oleh karena itu, diperlukan teknik seleksi fitur yang mampu memilih subset fitur optimal sebelum proses klasifikasi dilakukan.

CFS merupakan metode seleksi fitur yang memilih subset fitur berdasarkan prinsip bahwa fitur yang baik harus memiliki korelasi tinggi terhadap kelas target dan korelasi rendah antar fitur. Dengan kata lain, CFS bertujuan memaksimalkan relevansi fitur terhadap variabel output serta meminimalkan redundansi antar fitur (Reddy et al., 2021 [5]). Beberapa penelitian menunjukkan bahwa penerapan CFS mampu meningkatkan performa klasifikasi penyakit jantung. Wosiak dan Zakrzewska (2018) menyatakan bahwa integrasi metode berbasis korelasi dalam proses seleksi fitur mampu meningkatkan akurasi dan efisiensi model klasifikasi kardiovaskular.

Penelitian lain oleh Vijayashree dan Sultana (2018) menunjukkan bahwa penerapan teknik seleksi fitur berbasis korelasi mampu meningkatkan akurasi hingga di atas 85% pada klasifikasi penyakit jantung.

Selain itu, Ghosh dan Islam (2023) membandingkan berbagai teknik seleksi fitur dan menunjukkan bahwa metode berbasis korelasi memberikan peningkatan stabilitas model serta mengurangi kompleksitas komputasi tanpa mengorbankan performa. Islam (2024) juga menegaskan bahwa *feature selection* merupakan komponen kritis dalam sistem prediksi penyakit jantung modern. Walaupun *random forest* secara inheren memiliki mekanisme pemilihan fitur melalui *random subspace*, penerapan teknik seleksi fitur eksternal seperti CFS sebelum proses klasifikasi masih menunjukkan peningkatan akurasi yang signifikan (Reddy, 2021 [5]). Hal ini menunjukkan bahwa optimasi akurasi melalui kombinasi CFS dan *Random Forest* masih menjadi ruang penelitian yang relevan.

Berdasarkan uraian tersebut, penelitian ini berfokus pada optimasi akurasi klasifikasi penyakit jantung dengan menerapkan CFS sebelum proses klasifikasi menggunakan algoritma *Random Forest*. Penelitian ini diharapkan dapat menghasilkan model klasifikasi yang lebih akurat, efisien, dan *robust* terhadap variasi data.

2. Rumusan Masalah

Berdasarkan latar belakang yang telah diuraikan, ditemukan beberapa permasalahan yang menjadi fokus dalam penelitian ini, yaitu sebagai berikut:

1. Bagaimana mengimplementasikan teknik CFS untuk memilih subset fitur yang optimal pada dataset penyakit jantung?
2. Bagaimana performa algoritma *random forest* dalam mengklasifikasikan penyakit jantung tanpa penerapan teknik seleksi fitur?
3. Apakah penerapan teknik CFS sebelum proses klasifikasi *random forest* dapat meningkatkan nilai akurasi, *precision*, *recall*, dan *F1-score* dibandingkan model tanpa seleksi fitur?
4. Seberapa besar pengaruh pengurangan fitur terhadap kompleksitas model, waktu komputasi, serta stabilitas performa klasifikasi?
5. Bagaimana membandingkan hasil klasifikasi antara model *baseline* (*random forest* tanpa seleksi fitur) dan model optimasi (CFS + *Random Forest*) secara kuantitatif dan objektif?

3. Tujuan dan Manfaat

1. Tujuan

Berdasarkan rumusan masalah yang telah dipaparkan, maka penelitian ini bertujuan untuk:

1. Menerapkan teknik CFS untuk melakukan seleksi fitur pada dataset penyakit jantung guna memperoleh subset fitur yang optimal.
2. Mengimplementasikan algoritma *random forest* sebagai metode klasifikasi dalam mendeteksi penyakit jantung berdasarkan fitur yang tersedia.
3. Menganalisis dan membandingkan performa model klasifikasi *random forest* sebelum dan sesudah penerapan teknik CFS berdasarkan metrik evaluasi seperti *accuracy*, *precision*, *recall*, dan *F1-score*.
4. Mengoptimalkan tingkat akurasi klasifikasi penyakit jantung dengan meminimalkan fitur yang redundan atau tidak relevan melalui pendekatan seleksi fitur berbasis korelasi.
5. Mengevaluasi pengaruh reduksi dimensi fitur terhadap kompleksitas model, waktu komputasi, serta stabilitas performa klasifikasi.

2. Manfaat

Penelitian ini diharapkan memberikan manfaat sebagai berikut:

1. **Bagi Ilmu Pengetahuan.** Memberikan kontribusi dalam pengembangan metode optimasi klasifikasi penyakit jantung melalui integrasi teknik seleksi fitur Correlation-based Feature Selection (CFS) dengan algoritma Random Forest.
2. **Bagi Pengembangan Sistem Cerdas di Bidang Kesehatan.** Menyediakan model klasifikasi penyakit jantung yang lebih akurat dan efisien sehingga dapat dijadikan dasar dalam pengembangan sistem pendukung keputusan (*Decision Support System*) di bidang medis.
3. **Bagi Peneliti Selanjutnya.** Menjadi referensi ilmiah dalam penelitian terkait kombinasi *feature selection* dan *ensemble learning* untuk meningkatkan performa klasifikasi data medis.
4. **Bagi Institusi Pendidikan.** Menambah referensi penelitian dalam bidang *data mining*, *machine learning*, dan penerapannya pada analisis data kesehatan.

4. Batasan Masalah

Pada penelitian ini terdapat batasan masalah yang akan dibahas dan menjadi ruang lingkup penelitian, yaitu sebagai berikut:

1. Dataset yang digunakan merupakan dataset penyakit jantung publik (seperti UCI *heart disease dataset* atau dataset sejenis) yang memuat atribut klinis seperti usia, jenis kelamin, tekanan darah, kadar kolesterol, detak jantung maksimum, dan variabel medis lainnya.
2. Klasifikasi yang dilakukan bersifat biner, yaitu mengelompokkan data ke dalam dua kategori utama: tidak terdapat penyakit jantung (0) dan terdapat penyakit jantung (1).
3. Algoritma klasifikasi yang digunakan dalam penelitian ini adalah *random forest*, tanpa membandingkannya dengan algoritma *machine learning* lainnya seperti *support vector machine* (SVM), *k-nearest neighbor* (KNN), atau *neural network*.

4. Teknik seleksi fitur yang diterapkan adalah CFS sebagai metode utama untuk memilih subset fitur yang relevan, tanpa menggunakan metode seleksi fitur lain seperti *information gain*, *chi-square*, *relieff*, atau *principal component analysis* (PCA).
5. Proses pembagian data dilakukan menggunakan metode *train-test split* atau *k-fold cross validation*, dengan proporsi pembagian data yang ditentukan sebelum proses pelatihan model.
6. Evaluasi performa model dilakukan menggunakan metrik klasifikasi, yaitu: (a) *accuracy*, (b) *precision*, (c) *recall*, (d) *F1-score*, dan (e) *confusion matrix*.
7. Penelitian ini berfokus pada optimasi akurasi model melalui seleksi fitur, dan tidak mencakup optimasi lanjutan seperti hyperparameter tuning menggunakan grid search, Random Search, atau algoritma optimasi berbasis metaheuristik.
8. Penelitian ini tidak membahas aspek klinis secara mendalam maupun implementasi sistem dalam lingkungan rumah sakit, melainkan berfokus pada analisis komputasional dan peningkatan performa model klasifikasi.

5. Keterbaruan

Tabel 1.1. Keterbaruan Penelitian

No	Penelitian Terdahulu	Keterbaruan
1.	<i>Prediction of Heart Disease Risk Using Machine Learning with Correlation-based Feature Selection</i> (Reddy, 2021 [5]). Penelitian ini menerapkan CFS pada beberapa algoritma machine learning untuk meningkatkan performa prediksi penyakit jantung.	Penelitian ini secara spesifik memfokuskan pada optimasi akurasi Random Forest melalui penerapan CFS dan melakukan analisis komparatif kuantitatif antara model sebelum dan sesudah seleksi fitur menggunakan metrik evaluasi lengkap.
2.	<i>Integrating Correlation-Based Feature Selection and Clustering for Improved Cardiovascular Disease Diagnosis</i> (Wosiak & Zakrzewska, 2018). Penelitian mengintegrasikan CFS dengan teknik clustering untuk meningkatkan akurasi diagnosis.	Penelitian ini tidak menggunakan pendekatan clustering, tetapi secara langsung menguji pengaruh CFS terhadap performa algoritma ensemble Random Forest, sehingga memberikan analisis yang lebih terfokus pada klasifikasi supervised learning.
3.	<i>Heart Disease Prediction Using Entropy-Based Feature Engineering and Ensembling</i> (Rajendran & Karthi, 2022 [2]). Penelitian menggunakan teknik feature engineering berbasis entropy dan ensemble learning.	Penelitian ini menggunakan pendekatan seleksi fitur berbasis korelasi CFS yang bertujuan mengurangi redundansi fitur sebelum proses klasifikasi, serta mengevaluasi dampaknya terhadap kompleksitas model dan stabilitas performa.
4.	<i>A Machine Learning Framework for Feature Selection in Heart Disease Classification</i> (Vijayashree & Sultana, 2018). Penelitian menggunakan metode optimasi berbasis swarm dan SVM sebagai classifier.	Penelitian ini menggunakan Random Forest sebagai classifier utama yang memiliki mekanisme ensemble voting dan menganalisis peningkatan performa melalui pendekatan filter-based CFS, bukan wrapper atau metaheuristik.
5.	<i>Precision Healthcare: Machine Learning Algorithms and Feature Selection</i>	Penelitian ini memberikan kontribusi spesifik pada integrasi CFS dengan

No	Penelitian Terdahulu	Keterbaruan
	<i>Strategies for Heart Disease Prediction</i> (Islam, 2024 [1]). Studi ini membahas berbagai strategi seleksi fitur dan algoritma machine learning untuk prediksi penyakit jantung.	Random Forest dalam konteks evaluasi eksperimental yang terukur, serta menekankan pada optimasi akurasi dan efisiensi komputasi secara sistematis.