PERBANDINGAN PERFORMA MODEL *EXTRA TREES CLASSIFIER*, XGBOOST, DAN ETCXGB DENGAN *FEATURE SELECTION* DAN *FEATURE EXTRACTION* PADA DATA PENYAKIT JANTUNG

SKRIPSI

Program Studi Sarjana Informatika

Jurusan Informatika

Oleh:

Maharani Nurazizah

NIM D1041221005

FAKULTAS TEKNIK

UNIVERSITAS TANJUNGPURA

PONTIANAK

2026

HALAMAN PERNYATAAN

Yang bertanda tangan di bawah ini:

Nama : Maharani Nurazizah

NIM : D1041221005

menyatakan bahwa dalam SKRIPSI yang berjudul "Perbandingan Performa Model *Extra Trees Classifier*, Xgboost, Dan ETCXGB Dengan *Feature Selection* Dan *Feature Extraction* Pada Data Penyakit Jantung" tidak terdapat karya yang pernah diajukan untuk memperoleh gelar sarjana di suatu perguruan tinggi manapun. Sepanjang pengetahuan Saya, tidak terdapat karya atau pendapat yang pernah ditulis atau diterbitkan oleh orang lain, kecuali yang secara tertulis diacu dalam naskah ini dan disebutkan dalam Daftar Rujukan.

Demikian pernyataan ini dibuat dengan sebenar-benarnya. Saya sanggup menerima konsekuensi akademis dan hukum di kemudian hari apabila pernyataan yang dibuat ini tidak benar.

Pontianak,15 Juni 2026

Maharani Nurazizah
NIM D1041221005

KEMENTERIAN PENDIDIKAN TINGGI, SAINS, DAN TEKNOLOGI
UNIVERSITAS TANJUNGPURA
FAKULTAS TEKNIK
Jalan Prof. Dr. H. Hadari Nawawi Pontianak 78124
Telepon (0561) 740186, WA: +6282152280907
Email : ft@untan.ac.id Website : http://teknik.untan.ac.id

HALAMAN PENGESAHAN

PERBANDINGAN PERFORMA MODEL *EXTRA TREES CLASSIFIER*, XGBOOST, DAN ETCXGB DENGAN *FEATURE SELECTION* DAN *FEATURE EXTRACTION* PADA DATA PENYAKIT JANTUNG

Jurusan Informatika
Program Studi Sarjana Informatika

Oleh:

Maharani Nurazizah
NIM. D1041221005

Telah dipertahankan di depan Penguji Skripsi pada tanggal 15 Juni 2026
dan diterima sebagai salah satu persyaratan untuk memperoleh gelar sarjana

Susunan Penguji Skripsi

Dosen Pembimbing Utama	: Dr. Arif Bijaksana Putra Negara, S.T., M.T. NIP. 197208081998021002
Dosen Pembimbing Kedua	: Helen Sastypratiwi, S.T., M.Eng. NIP. 198601172012122004
Dosen Penguji Utama	: Prof. Dr. Herry Sujaini, S.T., M.T. NIP. 196806291997021001
Dosen Penguji Kedua	: Rina Septiriana, S.T., M.Cs. NIP. 198709232020122001

Pontianak, 15 Juni 2026

Dekan,

Dr.-Ing. Ir. Slamet Widodo, M.T., IPM
NIP. 196712231992031002

Pembimbing Utama

Dr. Arif Bijaksana Putra Negara, S.T., M.T.
NIP. 197208081998021002

Dipindai dengan CamScanner

HALAMAN PERSEMBAHAN

Skripsi ini didedikasikan untuk kedua orang tua saya (Bapak Barkah Bujiyan Tohatsu dan Ibu Tri Pamili Handayani) dan kedua saudaraku (Luqmanul hakim dan Nayla Humaira). Kebaikan, kasih sayang, dan dukungan tanpa henti dari kalian telah memberi saya lebih dari apa yang pernah saya mimpikan.

PRAKATA

Puji syukur kehadirat Allah SWT atas kehendak-Nya maka penelitian dan penulisan skripsi dengan judul "Perbandingan Performa Model Extra Trees Classifier, Xgboost, Dan ETCXGB Dengan *Feature Selection* Dan *Feature Extraction* Pada Data *Heart Disease*". Skripsi ini ditulis untuk memenuhi syarat kelulusan untuk mendapat gelar pada prodi Informatika fakultas Teknik universitas Tanjungpura

Penulis menyadari bahwa pengerjaan penelitian ini, tidak terlepas dari bantuan yang diperoleh melalui bimbingan dan arahan dari berbagai pihak. Tanpa bimbingan dan arahan yang diberikan oleh kedua dosen pembimbing, penelitian ini tidak dapat terselesaikan. Oleh karena itu, penulis ingin mengucapkan terimakasih kepada: Bapak Dr. Arif Bijaksana Putra Negara, S.T. M.T. selaku Dosen Pembimbing Utama dan Ibu Helen Sastypratiwi, S.T. M.Eng selaku Dosen Pembimbing Pendamping

Penulis menyadari bahwa skripsi ini masih memiliki keterbatasan. Oleh karena itu, Penulis terbuka terhadap kritik dan saran yang membangun demi pengembangan penelitian selanjutnya. Penulis berharap penelitian ini dapat memberikan kontribusi akademik serta menjadi referensi bagi pengembangan model klasifikasi penyakit jantung di masa mendatang.

Pontianak,

Penulis

Maharani Nurazizah

ABSTRAK

Penyakit jantung merupakan penyebab kematian global tertinggi dengan total sekitar 17,9 juta jiwa per tahun. Keterlambatan diagnosis menjadi faktor utama yang memperburuk kondisi ini, sehingga diperlukan metode prediksi dini berbasis machine learning. Meskipun penelitian mengenai model hibrida telah banyak dilakukan, perbandingan komprehensif antara kinerja model tunggal dan model hibrida, serta pengaruh mendalam dari pengolahan fitur terhadap performa model, masih belum ditinjau secara saksama. Penelitian ini bertujuan untuk mengevaluasi pengaruh penerapan metode seleksi fitur (Fisher Score dan Mutual Information) serta ekstraksi fitur (Principal Component Analysis - PCA) terhadap kinerja tiga model klasifikasi: Extra Trees Classifier (ETC), XGBoost, dan model hibrida ETCXGB. Pengujian dilakukan melalui delapan skenario pemrosesan fitur pada dataset penyakit jantung dengan 13 variabel klinis. Evaluasi kinerja dilakukan berdasarkan metrik confusion matrix (akurasi, presisi, recall, F1-score, dan AUC), serta analisis stabilitas (standar deviasi dan gap akurasi). Hasil penelitian menunjukkan bahwa pengaruh metode pengolahan fitur bergantung pada kesesuaian karakteristik fitur terhadap mekanisme pembelajaran model. Model ETC mencapai performa paling optimal pada skenario Mutual Information dengan akurasi 83,44%, AUC 0,9099, dan standar deviasi terendah 0,0106 (menunjukkan stabilitas terbaik). Pada sisi lain, XGBoost dan model hibrida ETCXGB menunjukkan performa terbaik pada skenario Fisher Score + PCA, dengan nilai AUC 0,9085 dan akurasi tertinggi 83,78%. Secara keseluruhan, konfigurasi ETC dengan Mutual Information menjadi model paling optimal karena mampu memberikan keseimbangan terbaik antara akurasi klasifikasi, generalisasi, dan stabilitas pada dataset dengan karakteristik fitur yang beragam.

Kata kunci: *Machine Learning*, Prediksi Penyakit Jantung, Pemrosesan Fitur, Extra Trees Classifier, XGBoost

ABSTRACT

Heart disease is the leading cause of global mortality, accounting for approximately 17.9 million deaths annually. Delayed diagnosis is a primary factor exacerbating this condition, necessitating the development of early predictive methods based on machine learning. While numerous studies on hybrid models have been conducted, a comprehensive comparison between single and hybrid models, as well as an in-depth investigation into the impact of feature processing on model performance, remain insufficiently explored. This study aims to evaluate the influence of feature selection methods (Fisher Score and Mutual Information) and feature extraction (Principal Component Analysis - PCA) on the performance of three classification models: Extra Trees Classifier (ETC), XGBoost, and an ETCXGB hybrid model. Testing was performed across eight feature processing scenarios using a heart disease dataset consisting of 13 clinical variables. Performance evaluation was conducted using a confusion matrix (accuracy, precision, recall, F1-score, and AUC), alongside stability analysis (standard deviation and accuracy gap). The results indicate that the impact of feature processing methods varies depending on the alignment between feature characteristics and the model's learning mechanism. The ETC model achieved optimal performance under the Mutual Information scenario with an accuracy of 83.44%, an AUC of 0.9099, and the lowest standard deviation of 0.0106, indicating superior stability. Conversely, the XGBoost and ETCXGB hybrid models attained peak performance under the Fisher Score + PCA scenario, achieving an AUC of 0.9085 and a maximum accuracy of 83.78%. Overall, the ETC configuration with Mutual Information is identified as the most optimal model, as it provides the best balance between classification accuracy, generalization, and stability for datasets characterized by diverse linear and non-linear feature distributions.

Keywords: Machine Learning, Heart Disease Prediction, Feature Processing, Extra Trees Classifier, XGBoost

DAFTAR ISI

HALAMAN PERNYATAN ..... **Error! Bookmark not defined.**
HALAMAN PENGESAHAN i
PRAKATA iv
ABSTRAK v
ABSTRACT vi
DAFTAR ISI vii
DAFTAR GAMBAR ix
DAFTAR TABEL x
DAFTAR KODE PROGRAM xi
DAFTAR LAMPIRAN xii
GLOSARIUM xiii
BAB I PENDAHULUAN 1
1.1 Latar Belakang 1
1.2 Perumusan Masalah 4
1.3 Tujuan dan Sasaran Penelitian 4
1.4 Pembatasan Masalah 5
1.5 Sistematika Penulisan 6
BAB II TINJAUAN PUSTAKA 8
2.1 Studi Literatur 8
2.3 Machine learning 11
2.4 Fitur Seleksi 13
2.5 Fitur Ekstraksi 14
2.6 Python 15
2.7 One Hot Encoding 15
2.8 Metriks Klasifikasi 15
2.10 Uji Signifikansi 18
BAB III METODOLOGI PENELITIAN 20
3.1 Perangkat Penelitian 20
3.2 Metodologi penelitian 20
3.2.1 Penggumpulan Dataset 21

3.2.3 Fitur Seleksi 27
3.2.3.1 *Fisher Score* 28
3.2.4 Fitur Ekstraksi 32
3.2.4.1 *Principal Component Analysis* (PCA) 33
3.2.5 Implementasi Model 35
3.2.5.1 *Extra Trees Classifier* (ETC) 37
3.2.5.2 XGBoost 39
3.2.5.3 ETC-XGBoost 40
3.2.6 Evaluasi 41
3.2.7 Analisis Hasil 47
BAB IV HASIL DAN PEMBAHASAN 50
4.1 Pra-pemprosesan data 50
4.2 Fitur Seleksi 58
4.3 Implementasi Model 64
4.4 Evaluasi 69
4.5 Analisis Hasil 95
4.5.1 Pengaruh Pengolahan Fitur terhadap kinerja model 95
4.5.2 Perbandingan perfoma dan stabilitas model 106
4.5.3 Uji Signifikasi Hasil Evaluasi 116
4.5.4 Perbandingan Model 118
BAB V KESIMPULAN DAN SARAN 120
5.1 Kesimpulan 120
5.2 Saran 120
DAFTAR RUJUKAN 122
DAFTAR LAMPIRAN 127

DAFTAR GAMBAR

Gambar 3. 1 Metodologi Penelitian..... 21
Gambar 3. 2 Diagram Alir Fitur Seleksi..... 28
Gambar 3. 3 Diagram Alir *Fisher Score* 30
Gambar 3. 4 Diagram Alir *Mutual Information* 32
Gambar 3. 5 Diagram Alir PCA 34
Gambar 3. 6 Proses Grid Search..... 36
Gambar 3. 7 Diagram Alir Implementasi ETC..... 38
Gambar 3. 8 Diagram Alir Implementasi Xgboost..... 39
Gambar 3. 9 Diagram Alir Implementasi ETCXGB 40
Gambar 4. 1 Visualisasi Boxplot untuk Pengecekan Outlier 52
Gambar 4. 2 Visualisasi Boxplot Fitur Kategorikal 55
Gambar 4. 3 Visualisasi Boxplot Fitur Numerik..... 56
Gambar 4. 4 Visualisasi Hubungan Fitur antar Fitur 57
Gambar 4. 5 Hasil Proses Pemilihan Fitur Numerik 60
Gambar 4. 6 Hasil Proses Pemilihan Fitur Kategorikal..... 62
Gambar 4. 7 Scree Plot Pencarian Ambang Varians Kumulatif..... 63
Gambar 4. 8 Confusion Matrix ETC 71
Gambar 4. 9 Confusion Matrix Xgboost 75
Gambar 4. 10 Confusion Matrix ETCXGB..... 78
Gambar 4. 11 ROC-AUC Skenario Seluruh Fitur..... 79
Gambar 4. 12 ROC-AUC Skenario *Fisher Score*..... 80
Gambar 4. 13 ROC AUC Skenario *Mutual Information*..... 82
Gambar 4. 14 ROC AUC Skenario PCA..... 83
Gambar 4. 15 ROC AUC Skenario *Fisher Score* dan *Mutual Information* 85
Gambar 4. 16 ROC AUC Skenario Implementasi *Fisher Score* dan PCA..... 86
Gambar 4. 17 ROC AUC Skenario *Mutual Information* dan PCA 87
Gambar 4. 18 ROC AUC Skenario *Fisher Score*, *Mutual Information* dan PCA..... 89
Gambar 4. 19 Visualisasi Perbandingan Metriks Klasifikasi Model ETC 99
Gambar 4. 20 Visualisasi Perbandingan Metriks Klasifikasi Model Xgboost 103
Gambar 4. 21 Visualisasi Perbandingan Metriks Klasifikasi ETCXGB 105
Gambar 4. 22 Hasil Evaluasi 24 Skenario..... 118

DAFTAR TABEL

Tabel 3. 1 Potongan Nilai Dataset pada penelitian 22
Tabel 3. 2 Fitur Pada Dataset Penyakit Jantung..... 22
Tabel 3. 3 Rentang Nilai Grid Search 37
Tabel 3. 4 Skenario Evaluasi Penelitian..... 42
Tabel 3. 5 Tabel Hasil Evaluasi Penelitian yang diharapkan..... 44
Tabel 3. 6 Hasil Uji Signifikansi pada 24 Skenario Eksperimen Melalui Wilcoxon Test..... 46
Tabel 4. 1 Hasil Pengecekan dan Penghapusan Duplikat..... 50
Tabel 4. 2 Hasil Penanganan Outlier..... 53
Tabel 4. 3 Tabel Skenario Fitur Seleksi 58
Tabel 4. 4 Tabel Implementasi Fisher Score..... 59
Tabel 4. 5 Hasil Implementasi Mutual Information..... 61
Tabel 4. 6 Analisis Threshold PCA..... 64
Tabel 4. 7 Hasil Grid Search untuk Mencari Hyperparameter..... 65
Tabel 4. 8 Tabel Hasil Evaluasi Penelitian 90
Tabel 4. 9 Hasil Uji Signifikansi pada 24 Skenario Eksperimen Melalui Wilcoxon Test..... 93

DAFTAR KODE PROGRAM

Kode Program 4. 1 Pembersihan data 50
Kode Program 4. 2 Penanganan Outlier menggunakan Median 52
Kode Program 4. 3 Implementasi *Fisher Score* 59
Kode Program 4. 4 Implementasi *Mutual Information* 60
Kode Program 4. 5 Implementasi PCA 63
Kode Program 4. 6 Potongan Kode Implementasi Gridseacrh 64
Kode Program 4. 7 Implementasi Model ETC 67
Kode Program 4. 8 Implementasi Xgboost 68
Kode Program 4. 9 Implementasi Model ETCXGB ETCXGB 69

DAFTAR LAMPIRAN

LAMPIRAN A 127
LAMPIRAN B 127

GLOSARIUM

Akurasi (Akurasi)	Metrik evaluasi yang mengukur proporsi prediksi benar terhadap seluruh data. Dihitung sebagai (TP + TN) / (TP + TN + FP + FN).
Area Under Curve (AUC)	Luas area di bawah kurva ROC dengan rentang nilai 0–1 yang menunjukkan kemampuan model membedakan kelas.
Seluruh fitur	Skenario pengujian tanpa seleksi maupun ekstraksi fitur, menggunakan seluruh fitur asli dataset.
Boosting	Teknik ensemble yang membangun model secara bertahap untuk memperbaiki kesalahan model sebelumnya.
Confusion Matrix	Matriks evaluasi 2×2 yang berisi TP, TN, FP, dan FN untuk menilai performa klasifikasi biner.
Cross-Validation	Teknik evaluasi dengan membagi data menjadi beberapa *fold* untuk pelatihan dan pengujian secara bergantian.
Data Leakage	Kondisi ketika informasi data uji ikut memengaruhi proses pelatihan sehingga hasil menjadi bias.
Encoding	Proses mengubah data kategorikal menjadi bentuk numerik agar dapat diproses model.
Ensemble Learning	Pendekatan yang menggabungkan beberapa model untuk meningkatkan performa prediksi.
Extra Trees Classifier (ETC)	Model ensemble berbasis pohon keputusan dengan pemilihan split acak untuk meningkatkan variasi dan efisiensi.
Feature Extraction	Transformasi fitur asli menjadi representasi baru berdimensi lebih rendah, seperti PCA.
Feature Selection	Proses memilih fitur yang paling relevan terhadap variabel target.
Fisher Score	Metode seleksi fitur numerik berdasarkan perbedaan rata-rata antar kelas dibandingkan variansi dalam kelas.

F1-Score	Rata-rata harmonik dari presisi dan *recall*.
GridSearch	Metode pencarian kombinasi hyperparameter terbaik secara sistematis.
ETCXGB	Model stacking yang menggabungkan ETC sebagai base learner dan XGBoost sebagai meta-learner.
Hyperparameter	Parameter model yang ditentukan sebelum proses pelatihan.
Imputasi	Proses mengisi nilai yang hilang dalam dataset.
Iterative Imputer	Metode imputasi yang memprediksi nilai hilang secara iteratif berdasarkan hubungan antar fitur.
Machine Learning	Cabang kecerdasan buatan yang mempelajari pola dari data untuk membuat prediksi.
Multikolinearitas	Kondisi adanya korelasi tinggi antar fitur independen.
Mutual Information	Ukuran ketergantungan antara fitur dan variabel target yang mampu menangkap hubungan non-linear.
Nested Cross-Validation	Validasi silang dua lapis untuk evaluasi model dan pencarian *hyperparameter*.
One-Hot Encoding	Teknik encoding yang mengubah kategori menjadi vektor biner terpisah.
Outlier	Nilai ekstrem yang berada di luar pola umum data.
Overfitting	Kondisi ketika model terlalu menyesuaikan data latih sehingga gagal pada data baru.
Principal Component Analysis (PCA)	Metode reduksi dimensi dengan membentuk komponen utama yang tidak berkorelasi.
Precision	Proporsi prediksi positif yang benar terhadap seluruh prediksi positif.
Recall	Proporsi kasus positif yang berhasil dideteksi model.
ROC *Curve*	Kurva yang menunjukkan hubungan antara TPR dan FPR pada berbagai threshold.

Stacking Ensemble	Metode ensemble yang menggabungkan prediksi model sebagai input model lain.
Standardisasi	Proses menyeragamkan skala fitur numerik.
Stratified K-Fold	Teknik cross-validation yang menjaga proporsi kelas di setiap *fold*.
Threshold	Nilai ambang batas untuk pengambilan keputusan atau seleksi fitur.
UCI Hungarian Heart Disease	Dataset penyakit jantung dari UCI Machine Learning Repository yang digunakan dalam penelitian.
XGBoost	Algoritma gradient boosting yang efisien dan memiliki regularisasi untuk mencegah overfitting.

BAB I
PENDAHULUAN

1.1 Latar Belakang

Penyakit jantung merupakan penyebab kematian tertinggi secara global dengan jumlah kematian mencapai sekitar 17,9 juta jiwa pada tahun 2024 (*World Health Organization*, 2024). Tingginya angka mortalitas tersebut salah satunya disebabkan oleh keterlambatan dalam proses penanganan. Kondisi ini menunjukkan pentingnya penerapan metode yang mampu mendukung diagnosis dini penyakit jantung [1]. Namun, proses diagnosis dini penyakit jantung umumnya memerlukan tahapan pemeriksaan yang relatif panjang serta biaya yang tidak sedikit [2]. Oleh karena itu, diperlukan solusi alternatif untuk mengatasi permasalahan tersebut.

Perkembangan teknologi memberikan peluang pemanfaatan metode komputasi sebagai solusi atas permasalahan tersebut, salah satunya melalui penerapan *machine learning*. *Machine learning* telah banyak digunakan dalam bidang kesehatan untuk menganalisis dan memprediksi data klinis. Kinerja metode *machine learning* sangat dipengaruhi oleh kualitas data dan fitur yang digunakan pada proses implementasi [3]. Data medis umumnya diperoleh dari hasil pemeriksaan klinis, sehingga cenderung bersifat kompleks dan memiliki hubungan antar fitur yang tinggi [4]. Hubungan antar fitur yang tinggi dapat menyebabkan redundansi informasi serta meningkatkan beban komputasi model apabila digunakan tanpa melalui proses pengolahan atau pemilihan fitur terlebih dahulu. Permasalahan tersebut sering ditemukan pada berbagai *dataset* penelitian medis

Salah satu *dataset* yang umum digunakan dalam penelitian *machine learning* di bidang medis adalah UCI *Hungarian Heart Disease*. *Dataset* ini diperoleh dari informasi medis pasien dan memuat sejumlah atribut pemeriksaan penyakit jantung yang digunakan untuk penilaian risiko penyakit jantung. Data dalam *dataset* ini terdiri dari dua jenis, yaitu data numerik dan data kategorikal [5]. Dataset *UCI Hungarian Heart Disease* yang digunakan dalam penelitian ini berasal dari repositori yang dilaporkan oleh [6]. Pada penelitian tersebut, proses klasifikasi dilakukan tanpa melalui tahapan pemilihan fitur, sehingga terdapat potensi penggunaan fitur yang bersifat redundan atau kurang relevan terhadap proses klasifikasi. Selain itu, jumlah data yang relatif terbatas serta kemungkinan adanya redundansi informasi dapat mempengaruhi performa model yang digunakan [7].

Pemilihan fitur merupakan tahapan penting dalam penerapan *machine learning*. Penggunaan seluruh fitur dalam suatu *dataset* tidak selalu menghasilkan peningkatan performa model, sehingga diperlukan pemilihan fitur yang memiliki relevansi tinggi terhadap target prediksi. Proses ini dapat dilakukan melalui metode seleksi fitur atau *feature selection* . Beberapa metode seleksi fitur yang umum digunakan antara lain *Fisher Score* daneman *Mutual Information*. *Fisher Score* bekerja dengan melakukan diskriminasi antar fitur berdasarkan perbandingan perbedaan rata-rata antar kelas terhadap varians setiap fitur [8]. Sementara itu, *Mutual Information* mengukur jumlah informasi yang dimiliki setiap fitur terhadap target, kemudian dilakukan pemeringkatan fitur berdasarkan nilai informasi tersebut [9]. Penerapan seleksi fitur berperan dalam mengidentifikasi hubungan antara fitur dan target serta mengurangi potensi redudansi fitur. Selain itu, data medis bersifat heterogen dan memiliki jenis tipe data lebih dari satu [10]. Kondisi tersebut menyebabkan kondisi fitur data yang berkolerasi satu sama lain. Korelasi antar fitur bisa menjadi permasalahan bagi model untuk bekerja [11]. Selain itu, dataset dengan jumlah kecil sering memiliki dimensi fitur yang tinggi, sehingga sulit untuk model bekerja dengan optimal [12].

Selain seleksi fitur, pendekatan lain yang dapat digunakan untuk mengatasi multikolinearitas data adalah ekstraksi fitur. *Principal Component Analysis (*PCA*)* digunakan dalam penelitian ini untuk mereduksi dimensi data dengan mengubah

representasi fitur ke dalam bentuk yang lebih sederhana, namun tetap mempertahankan informasi utama yang terkandung di dalamnya. PCA membentuk fitur baru yang saling tidak berkorelasi sehingga dapat membantu meningkatkan kinerja model [13].

Beberapa penelitian telah dilakukan dengan menggunakan *dataset* UCI *Hungarian Heart Disease*, salah satunya oleh [6] yang menerapkan model hibrida *Extra Trees Classifier (*ETC*)* dan XGBoost. Hasil penelitian tersebut menunjukkan performa model yang signifikan dalam melakukan prediksi penyakit jantung. Namun, penelitian tersebut tidak melakukan evaluasi kinerja model tunggal, sehingga diperlukan kajian lebih lanjut terkait performa masing-masing model secara independen.

Penelitian ini, model *Extra Trees Classifier (*ETC*)*, XGBoost, serta model ETCXGB digunakan sebagai metode klasifikasi pada *dataset* UCI *Hungarian Heart Disease*. ETC merupakan metode *ensemble* yang membangun pohon keputusan dengan proses pembelahan acak untuk meningkatkan kemampuan prediksi [14]. Sementara itu, XGBoost merupakan metode berbasis pohon keputusan yang dibangun secara bertahap. Setiap pohon digunakan untuk memperbaiki kesalahan pada pohon sebelumnya [15]. Penelitian yang dilakukan oleh menunjukkan bahwa penggunaan fitur yang relevan berkontribusi terhadap peningkatan akurasi dan performa model XGBoost. Perbedaan karakteristik kedua model tersebut menjadi dasar penerapan model hibrida sebagaimana dilakukan oleh [6].

Berdasarkan perbedaan mekanisme kerja antara ETC dan XGBoost, diperlukan evaluasi kinerja secara independen pada masing-masing model. Meskipun memiliki pendekatan yang berbeda, kedua model tersebut sama-sama memanfaatkan pohon keputusan yang dibangun dari fitur, sehingga kinerjanya sangat bergantung pada kualitas data dan fitur yang digunakan. Oleh karena itu, penelitian ini bertujuan untuk menguji pengaruh penerapan metode *Fisher Score*, *Mutual Information*, dan PCA, serta penggunaan seluruh fitur tanpa proses pengolahan fitur, terhadap performa model ETC, XGBoost, dan ETCXGB. Hasil pengujian selanjutnya dievaluasi untuk meninjau dampak penerapan seleksi dan ekstraksi fitur terhadap performa model klasifikasi penyakit jantung.

1.2 Perumusan Masalah

1. Apa pengaruh penerapan metode *Fisher Score*, *Mutual Information*, PCA dan kombinasinya terhadap performa model *Extra Trees Classifier*, XGBoost, *atau* ETCXGB?

2. Model manakah di antara *Extra Trees Classifier*, XGBoost, ETCXGB yang menghasilkan performa optimal?

1.3 Tujuan dan Sasaran Penelitian

1. Untuk mengidentifikasi pengaruh penerapan metode *Fisher Score*, *Mutual Information*, PCA dan kombinasinya terhadap performa model *Extra Trees Classifier*, XGBoost, *atau* ETCXGB.
2. Untuk menentukan model manakah di antara *Extra Trees Classifier*, *XGBoost*, atau ETCXGB yang memberikan hasil performa terbaik dalam melakukan klasifikasi.

Agar tujuan tersebut tercapai, sasaran yang ingin dicapai meliputi:

1. Menyediakan hasil evaluasi performa model *Extra Trees Classifier*, XGBoost, dan ETCXGB pada berbagai skenario preprocessing, yang diukur menggunakan metrik akurasi, presisi , *recall*, *F1-score*, ROC-AUC, dan standar deviasi.
2. Menyediakan hasil uji signifikansi yang digunakan sebagai alat untuk melihat pengaruh perubahan pada penerapan *Fisher Score*, *Mutual Information*, PCA, dan kombinasnya terhadap seluruh fitur pada model ETC, Xgboost, dan ETCXGB.
3. Menyediakan hasil implementasi secara kuantitatif antara *Fisher Score, Mutual Information* sebagai metode fitur seleksi, PCA sebagai metode fitur ekstraksi.
4. Memberikan kontribusi akademis berupa literatur perbandingan metode preprocessing pada prediksi penyakit jantung menggunakan model *ensemble*, serta kontribusi praktis berupa rekomendasi metode yang paling optimal untuk implementasi sistem prediksi penyakit jantung.
5. Memberikan pemahaman tentang peran setiap fitur dalam prediksi penyakit jantung secara kuantitatif berdasarkan hasil *feature selection* dan *feature extraction*, sebagai kontribusi terhadap pemanfaatan data medis dalam sistem

prediksi berbasis *machine learning.*

1.4 Pembatasan Masalah

1. Penelitian ini menggunakan dataset yang berasal dari penelitian yang dilakukan pada penelitian bilal (2024). Dataset tersebut menggunakan UCI Hungarian Heart Disease. Dataset berisi 13 variabel klinis. Analisis hanya dilakukan pada satu dataset ini tanpa penggabungan dengan dataset lain agar hasil pengujian tetap konsisten dan terkontrol.
2. Penelitian difokuskan pada tahap rekayasa fitur (feature engineering) untuk meningkatkan performa model prediksi penyakit jantung. Tahap ini mencakup dua bagian utama, yaitu *feature selection* dan *feature extraction*. Metode *feature selection* yang digunakan dibatasi pada dua pendekatan filter: *Fisher Score* dan *Mutual Information* . Metode *feature extraction* dibatasi pada *Principal Component Analysis* (PCA) sebagai teknik reduksi dimensi.
3. Pengujian model dibatasi pada tiga algoritma ensemble learning: *ETC, XGBoost*, dan ETCXGB. Model penelitian ini diadaptasi dari studi sebelumnya yang menggunakan model ETCXGB sebagai seluruh fitur. Dua algoritma penyusunnya, ETC dan *XGBoost*, turut digunakan secara mandiri sebagai model pembanding.
4. Penelitian ini menguji delapan skenario utama penerapan *feature selection* dan *feature extraction*, yaitu: seluruh fitur, *Fisher Score, Mutual Information,* PCA, *Fisher Score* + *Mutual Information*, PCA, *Fisher Score* + PCA, *Mutual Information* + PCA, dan *Fisher Score* + *Mutual Information* + PCA. Setiap skenario akan diuji pada tiga model ensemble learning ETC, XGBoost, dan ETCXGB, sehingga total pengujian adalah 24 kombinasi skenario pengujian.
5. Uji statistik wilcoxon testdigunakan sebagai alat pendukung analisis untuk mengevaluasi signifikansi terhadap perubahan yang dihasilkan pada penerapan metode penerapan fitur.
6. Evaluasi performa model dilakukan menggunakan metrik utama, yaitu *Akurasi, Precision, Recall, F1-score,* dan *ROC*-AUC serta standar deviasi. Metrik utama ini, digunakan sebagai alat penentu model optimal pada implementasi.

1.5 Sistematika Penulisan

Sistematika penulisan proposal skripsi ini disusun agar memudahkan pembahasan penelitian secara terstruktur. Adapun susunannya adalah sebagai berikut.

1. BAB I : Pendahuluan

   Bab ini memuat latar belakang penelitian, perumusan masalah, tujuan dan sasaran penelitian, pembatasan masalah, serta sistematika penulisan. Bagian ini memberikan gambaran umum mengenai urgensi deteksi dini penyakit jantung, keterbatasan metode diagnostik konvensional, peran *machine learning*, serta alasan dilakukannya analisis pengaruh *feature selection* dan *feature extraaction* menggunakan *Fisher Score*, *Mutual Information* dan PCA pada model ETCXGB.