

Analisis Komparasi Kinerja SVM, Random Forest, dan XGBoost untuk Diagnosis Penyakit Batu Empedu

Comparative Performance Analysis of SVM, Random Forest, and XGBoost for Gallstone Disease Diagnosis

Ananda Dika Erlangga, Erliyan Redy Susanto*

Fakultas Teknik dan Ilmu Komputer, Teknologi Informasi, Universitas Teknokrat Indonesia, Bandar Lampung, Indonesia

Email: ¹ananda_dika_erlangga@teknokrat.ac.id, ^{2,*}erliyan.redy@teknokrat.ac.id

Email Penulis Korespondensi: erliyan.redy@teknokrat.ac.id

Received: 16/08/2026, Revised: 30/08/2026, Accepted: 21/09/2026, Published: 22/09/2026

Abstrak—Penyakit batu empedu merupakan salah satu gangguan hepatobilier yang memerlukan diagnosis secara tepat untuk mendukung penanganan yang lebih cepat dan mengurangi risiko komplikasi. Ketersediaan data klinis dan laboratorium memberikan peluang untuk menerapkan *machine learning* sebagai pendekatan pendukung dalam proses klasifikasi penyakit. Penelitian ini bertujuan membandingkan kinerja algoritma *Support Vector Machine* (SVM), *Random Forest*, dan XGBoost dalam klasifikasi penyakit batu empedu menggunakan *Dataset Gallstone* dari UCI Machine Learning Repository. Dataset yang digunakan terdiri dari 319 sampel, sehingga evaluasi model perlu mempertimbangkan keterbatasan jumlah data dan potensi overfitting yang dapat memengaruhi kemampuan generalisasi model. Kontribusi penelitian ini adalah menyediakan evaluasi komparatif SVM, *Random Forest*, dan XGBoost pada *Dataset Gallstone* dengan menerapkan tahapan seleksi fitur, penanganan ketidakseimbangan kelas, standarisasi, dan validasi secara konsisten, sehingga perbedaan kinerja ketiga algoritma dapat dibandingkan secara lebih objektif. Tahapan penelitian meliputi seleksi fitur menggunakan ANOVA *F-Value*, penanganan ketidakseimbangan kelas menggunakan *Synthetic Minority Over-sampling Technique* (SMOTE), standarisasi fitur, pembentukan model, dan validasi menggunakan *Stratified 5-Fold Cross Validation*. Evaluasi dilakukan menggunakan *accuracy*, *precision*, *recall*, *F1-score*, dan *Area Under Curve* (AUC). Hasil penelitian menunjukkan bahwa XGBoost menghasilkan kinerja terbaik dengan *accuracy* 79,00%, *precision* 80,95%, *recall* 75,32%, *F1-score* 78,03%, dan AUC 85,04%. *Random Forest* memperoleh *accuracy* 78,06%, sedangkan SVM sebesar 74,92%. Analisis *feature importance* menunjukkan bahwa C-Reactive Protein (CRP), Vitamin D, *Hyperlipidemia*, *Bone Mass*, dan *Aspartate Aminotransferase* (AST) termasuk atribut yang memberikan kontribusi penting terhadap klasifikasi. Hasil penelitian menunjukkan bahwa XGBoost memiliki kinerja paling baik dibandingkan SVM dan *Random Forest* pada dataset yang digunakan dan berpotensi menjadi pendekatan pendukung dalam pengembangan sistem klasifikasi penyakit batu empedu, dengan pengujian pada dataset yang lebih besar dan beragam diperlukan untuk memastikan kemampuan generalisasinya.

Kata Kunci: Penyakit Batu Empedu; Machine Learning; SVM; Random Forest; XGBoost; Klasifikasi

Abstract—Gallstone disease is a hepatobiliary disorder that requires accurate diagnosis to support timely management and reduce the risk of complications. The availability of clinical and laboratory data provides an opportunity to apply machine learning as a supporting approach for disease classification. This study aims to compare the performance of Support Vector Machine (SVM), Random Forest, and XGBoost algorithms for gallstone disease classification using the Gallstone Dataset from the UCI Machine Learning Repository. The contribution of this study is to provide a comparative evaluation of SVM, Random Forest, and XGBoost on the Gallstone Dataset by consistently applying feature selection, feature standardization, class imbalance handling, and cross-validation to obtain a more objective comparison of the three algorithms. The dataset consists of 319 samples; therefore, model evaluation needs to consider the limited sample size and the potential risk of overfitting, which may affect model generalization. The research stages include feature selection using the ANOVA *F-Value*, handling class imbalance using the Synthetic Minority Over-sampling Technique (SMOTE), feature standardization, model development, and validation using Stratified 5-Fold Cross Validation. Evaluation was conducted using accuracy, precision, recall, F1-score, and Area Under the Curve (AUC). The results show that XGBoost achieved the best performance, with an accuracy of 79.00%, precision of 80.95%, recall of 75.32%, F1-score of 78.03%, and AUC of 85.04%. Random Forest achieved an accuracy of 78.06%, while SVM achieved 74.92%. Feature importance analysis indicated that C-Reactive Protein (CRP), Vitamin D, Hyperlipidemia, Bone Mass, and Aspartate Aminotransferase (AST) were among the attributes that contributed significantly to the classification. The results indicate that XGBoost outperformed SVM and Random Forest on the dataset used and has potential as a supporting approach for developing a gallstone disease classification system; however, evaluation using larger and more diverse datasets is required to confirm its generalization capability.

Keywords: Gallstone Disease; Machine Learning; SVM; Random Forest; XGBoost; Classification

1. PENDAHULUAN

Penyakit batu empedu (*gallstone disease*) merupakan salah satu penyakit pada sistem hepatobilier yang ditandai dengan terbentuknya endapan padat berupa kolesterol, pigmen bilirubin, maupun campuran keduanya di dalam kantung empedu atau saluran empedu. Penyakit ini merupakan salah satu gangguan saluran cerna yang banyak ditemukan di dunia dan masih menjadi permasalahan kesehatan global karena prevalensinya yang cukup tinggi serta berpotensi menimbulkan berbagai komplikasi serius [1]. Berdasarkan hasil *systematic review* dan *meta-analysis* yang melibatkan 115 penelitian dengan 32.610.568 responden, prevalensi penyakit batu empedu di dunia mencapai sekitar 6,1% dari populasi global, dengan prevalensi lebih tinggi pada perempuan (7,6%) dibandingkan laki-laki (5,4%). Selain itu, prevalensi penyakit ini cenderung meningkat pada kelompok usia lanjut serta individu dengan gangguan

metabolik seperti obesitas, *diabetes mellitus*, dan dislipidemia. Meskipun sebagian besar penderita tidak menunjukkan gejala pada fase awal, penyakit batu empedu dapat berkembang menjadi komplikasi serius seperti kolesistitis akut, kolangitis, pankreatitis, hingga obstruksi saluran empedu apabila tidak didiagnosis sejak dini. Kondisi tersebut menunjukkan bahwa diagnosis dini menjadi faktor penting dalam meningkatkan keberhasilan terapi, mengurangi morbiditas, serta mencegah terjadinya komplikasi pada pasien. Oleh karena itu, diperlukan pendekatan yang mampu membantu proses diagnosis secara cepat, objektif, dan akurat sebagai pendukung pengambilan keputusan klinis [2].

Perkembangan teknologi pemeriksaan kesehatan menghasilkan berbagai data klinis dan laboratorium yang dapat dimanfaatkan sebagai sumber informasi dalam proses diagnosis. Pada kasus batu empedu, diagnosis umumnya dilakukan melalui pemeriksaan *Ultrasonography* (USG), *Computed Tomography* (CT Scan), maupun *Magnetic Resonance Cholangiopancreatography* (MRCP). Meskipun metode tersebut memiliki kemampuan diagnostik yang baik, penggunaannya memerlukan fasilitas kesehatan yang memadai dan interpretasi oleh tenaga medis yang berpengalaman. Selain itu, gejala klinis penyakit batu empedu dapat menyerupai gangguan saluran cerna lainnya sehingga diagnosis berbasis gejala saja dapat mengalami keterbatasan. Di sisi lain, data klinis pasien seperti usia, *Body Mass Index* (BMI), kadar glukosa, kolesterol, trigliserida, fungsi hati, serta riwayat penyakit dapat digunakan sebagai sumber informasi untuk membangun model klasifikasi berbasis *Machine Learning*. Salah satu sumber data yang dapat digunakan adalah *Gallstone Dataset* dari UCI Machine Learning Repository yang terdiri atas 319 data pasien dengan 38 atribut klinis [3]. Karakteristik dataset tersebut menunjukkan adanya tantangan metodologis karena jumlah sampel relatif terbatas dibandingkan dengan jumlah atribut yang tersedia. Kondisi tersebut dapat meningkatkan risiko model mempelajari pola yang terlalu spesifik terhadap data pelatihan sehingga kemampuan generalisasi pada data baru perlu menjadi perhatian dalam proses pembangunan dan evaluasi model.

Seiring berkembangnya penerapan *Artificial Intelligence* (AI) di bidang kesehatan, *Machine Learning* menjadi salah satu pendekatan yang digunakan untuk mengidentifikasi pola pada data klinis dan membantu proses klasifikasi penyakit. Pada data tabular kesehatan, pemilihan algoritma perlu mempertimbangkan karakteristik data, jumlah sampel, hubungan antarfitur, serta potensi ketidakseimbangan kelas. *Support Vector Machine* (SVM) sesuai untuk dipertimbangkan karena mampu membentuk batas keputusan yang efektif pada data dengan jumlah sampel terbatas dan bekerja baik pada ruang fitur berdimensi relatif tinggi melalui penggunaan fungsi kernel. *Random Forest* memiliki karakteristik berbeda karena membangun sejumlah pohon keputusan dan melakukan agregasi hasil prediksi, sehingga mampu menangkap hubungan nonlinier serta relatif robust terhadap variasi fitur. Sementara itu, *XGBoost* merupakan algoritma *gradient boosting* yang membangun model secara bertahap dengan memperbaiki kesalahan prediksi model sebelumnya dan memiliki kemampuan menangkap hubungan kompleks pada data tabular. Perbedaan karakteristik ketiga algoritma tersebut menjadi dasar teoritis bahwa perbandingan SVM, *Random Forest*, dan *XGBoost* relevan dilakukan pada dataset klinis dengan jumlah sampel terbatas dan banyak atribut, karena ketiganya memiliki mekanisme pembelajaran yang berbeda dalam membentuk batas keputusan dan menangkap hubungan antarfitur.

Penelitian sebelumnya pada *Gallstone Dataset* membandingkan algoritma *Logistic Regression*, *Decision Tree*, *Random Forest*, *Support Vector Machine* (SVM), dan *K-Nearest Neighbor* (KNN), dengan hasil bahwa *Logistic Regression* dan SVM memperoleh akurasi sebesar 85,9% serta nilai *Area Under Curve* (AUC) di atas 0,90 [4]. Penelitian tersebut menunjukkan bahwa model berbasis batas keputusan seperti SVM dapat memberikan performa tinggi pada dataset batu empedu, tetapi fokus perbandingannya belum mencakup *XGBoost*. Penelitian lain pada klasifikasi penyakit *liver* menunjukkan bahwa *Random Forest* mampu menangani hubungan nonlinier antarfitur, sedangkan penelitian mengenai algoritma *boosting* menunjukkan bahwa *XGBoost* dapat meningkatkan performa klasifikasi melalui proses pembelajaran bertahap untuk meminimalkan kesalahan prediksi [5]. Selain itu, penelitian mengenai perbandingan CatBoost dan XGBoost pada klasifikasi penyakit jantung menunjukkan bahwa pendekatan *gradient boosting* dapat diterapkan pada data kesehatan dan mampu menangani pola hubungan yang kompleks [6]. Meskipun penelitian sebelumnya pada dataset yang sama melaporkan akurasi hingga 85,9%, nilai tersebut tidak dapat secara langsung dijadikan tolok ukur bahwa model dengan akurasi lebih rendah selalu memiliki kinerja yang lebih buruk. Perbedaan performa dapat dipengaruhi oleh tahapan preprocessing, metode penanganan ketidakseimbangan kelas, pemilihan fitur, pembagian data, metode validasi, konfigurasi parameter, serta metrik evaluasi yang digunakan. Penelitian sebelumnya yang memperoleh akurasi 85,9% menggunakan pendekatan dan konfigurasi eksperimen yang berbeda dengan penelitian ini [7], [8]. Oleh karena itu, perbedaan hasil menjadi bagian penting yang perlu dianalisis untuk mengetahui bagaimana perubahan tahapan pengolahan data dan pemilihan algoritma memengaruhi performa model pada dataset yang sama. Dalam penelitian ini, evaluasi tidak hanya menggunakan accuracy, tetapi juga precision, recall, F1-score, dan AUC sehingga kemampuan model dapat dinilai dari beberapa aspek, terutama ketika jumlah data relatif terbatas dan karakteristik kelas perlu diperhatikan.

Berdasarkan sedikitnya empat arah penelitian terdahulu tersebut, terdapat beberapa kesenjangan yang perlu diisi. Pertama, penelitian pada *Gallstone Dataset* yang membandingkan *Logistic Regression*, *Decision Tree*, *Random Forest*, SVM, dan KNN telah menunjukkan performa tinggi pada SVM dan *Logistic Regression*, tetapi belum memberikan perbandingan langsung antara SVM, *Random Forest*, dan *XGBoost* dalam satu rancangan eksperimen yang seragam [4]. Kedua, penelitian *Random Forest* pada penyakit *liver* dan penelitian berbasis *boosting* pada data kesehatan menunjukkan kemampuan algoritma dalam menangkap hubungan nonlinier dan pola kompleks, tetapi objek penyakit dan karakteristik datanya berbeda sehingga hasilnya belum dapat langsung digunakan untuk menjelaskan performa ketiga algoritma pada penyakit batu empedu [5]. Ketiga, penelitian CatBoost dan XGBoost pada penyakit jantung menunjukkan potensi *gradient boosting* pada data kesehatan, tetapi belum menguji perbandingan tersebut

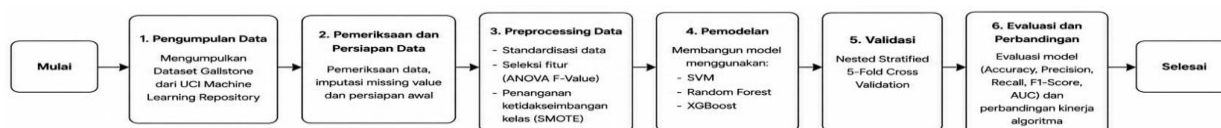
pada *Gallstone Dataset* dengan jumlah 319 sampel dan 38 atribut [6]. Keempat, perbedaan prosedur *preprocessing*, pemilihan fitur, penanganan ketidakseimbangan kelas, dan strategi validasi pada penelitian terdahulu menyebabkan hasil performa antarpelelitian tidak sepenuhnya dapat dibandingkan secara langsung [7], [8].

Kesenjangan tersebut menjadi semakin penting karena dataset hanya terdiri atas 319 sampel dengan 38 atribut, sehingga rasio antara jumlah fitur dan jumlah sampel perlu dipertimbangkan dalam pembangunan model. Penggunaan seluruh fitur tanpa seleksi berpotensi meningkatkan kompleksitas model dan risiko *overfitting*. Oleh sebab itu, diperlukan penelitian yang tidak hanya membandingkan algoritma, tetapi juga menerapkan tahapan pengolahan data yang konsisten, yaitu seleksi fitur menggunakan ANOVA F-Value, penanganan ketidakseimbangan kelas menggunakan SMOTE, standarisasi fitur, dan validasi *Stratified 5-Fold Cross Validation*. Kombinasi tersebut diharapkan menghasilkan evaluasi yang lebih terstruktur dan memungkinkan perbandingan SVM, *Random Forest*, dan *XGBoost* dilakukan secara lebih objektif pada dataset yang sama. Berdasarkan permasalahan tersebut, penelitian ini bertujuan membangun dan membandingkan model klasifikasi penyakit batu empedu menggunakan algoritma Support Vector Machine (SVM), Random Forest, dan XGBoost berdasarkan data klinis pasien yang diperoleh dari UCI Machine Learning Repository [9] [10]. Tahapan penelitian meliputi seleksi fitur menggunakan ANOVA F-Value, penanganan ketidakseimbangan kelas menggunakan *Synthetic Minority Over-sampling Technique* (SMOTE), standarisasi fitur, pembangunan model, dan validasi menggunakan *Stratified 5-Fold Cross Validation*. Setiap model dievaluasi menggunakan *accuracy*, *precision*, *recall*, *F1-score*, dan *Area Under Curve* (AUC) [11].

Kontribusi penelitian ini terletak pada penyusunan evaluasi komparatif SVM, *Random Forest*, dan *XGBoost* pada *Gallstone Dataset* dengan mengintegrasikan seleksi fitur ANOVA F-Value, penanganan ketidakseimbangan kelas menggunakan SMOTE, standarisasi fitur, dan *Stratified 5-Fold Cross Validation* dalam satu alur eksperimen yang konsisten. Pendekatan tersebut memberikan evaluasi yang lebih komprehensif terhadap kemampuan ketiga algoritma pada dataset klinis dengan jumlah sampel terbatas, sekaligus memperjelas pengaruh pemilihan algoritma dan tahapan pengolahan data terhadap performa klasifikasi. Dari sisi akademis, penelitian ini memberikan bukti komparatif yang dapat digunakan sebagai referensi untuk pengembangan model klasifikasi berbasis data klinis, sedangkan dari sisi praktis hasil penelitian dapat menjadi dasar bagi pengembangan sistem pendukung keputusan klinis yang membantu proses diagnosis awal secara lebih cepat dan objektif. Penelitian ini diharapkan memberikan manfaat praktis maupun akademis. Dari sisi praktis, hasil penelitian dapat menjadi referensi dalam pengembangan sistem pendukung keputusan klinis berbasis *Machine Learning* untuk membantu tenaga medis melakukan diagnosis awal penyakit batu empedu secara lebih cepat dan objektif. Namun, mengingat dataset yang digunakan hanya terdiri atas 319 sampel, hasil penelitian ini belum dapat digeneralisasikan secara luas dan masih memerlukan pengujian pada dataset yang lebih besar, beragam, dan berasal dari populasi yang berbeda. Dari sisi akademis, penelitian ini memberikan kontribusi berupa analisis komparatif terhadap SVM, Random Forest, dan XGBoost dengan karakteristik pembelajaran yang berbeda pada dataset klinis batu empedu serta evaluasi menggunakan beberapa metrik performa. Hasil penelitian diharapkan dapat menjadi referensi bagi penelitian selanjutnya dalam penerapan *Machine Learning* untuk klasifikasi penyakit berbasis data klinis.

2. METODOLOGI PENELITIAN

Penelitian ini menggunakan pendekatan *machine learning* untuk membangun dan membandingkan model klasifikasi penyakit batu empedu berdasarkan data klinis pasien. Tahapan penelitian meliputi pengumpulan data, persiapan data, *preprocessing*, seleksi fitur, standarisasi, penanganan ketidakseimbangan kelas, pemodelan, validasi, dan evaluasi. Algoritma yang dibandingkan adalah *Support Vector Machine* (SVM), *Random Forest*, dan *Extreme Gradient Boosting* (XGBoost). Untuk memperoleh estimasi performa yang lebih objektif, digunakan *Nested Stratified 5-Fold Cross Validation*. Pada setiap *outer fold*, empat bagian digunakan sebagai data pelatihan dan satu bagian sebagai data validasi akhir, sedangkan *inner Stratified 5-Fold Cross Validation* digunakan untuk memilih konfigurasi parameter terbaik. Seluruh tahapan yang mempelajari karakteristik data, yaitu imputasi, standarisasi, seleksi fitur, dan SMOTE, ditempatkan dalam *pipeline* dan hanya diterapkan pada data pelatihan setiap *fold*. Dengan demikian, data validasi tidak digunakan dalam proses *preprocessing*, pemilihan fitur, maupun pembentukan data sintetis. Alur penelitian ditunjukkan pada Gambar 1.



Gambar 1. Tahapan Penelitian

Berdasarkan Gambar 1, tahapan penelitian dimulai dari pengumpulan *Dataset Gallstone*, dilanjutkan dengan pemeriksaan dan persiapan data. Selanjutnya, data diproses menggunakan *outer Stratified 5-Fold Cross Validation*. Pada setiap *outer training fold*, dilakukan *inner Stratified 5-Fold Cross Validation* untuk memilih konfigurasi model terbaik. Proses tersebut meliputi imputasi, standarisasi, seleksi 15 fitur terbaik menggunakan ANOVA F-Value, SMOTE, dan pembentukan model SVM, Random Forest, atau XGBoost. Model dengan konfigurasi terbaik kemudian

digunakan untuk melakukan prediksi pada *outer validation fold*. Hasil prediksi dari seluruh *outer validation fold* selanjutnya digunakan untuk menghitung *accuracy*, *precision*, *recall*, *F1-score*, dan *Area Under Curve* (AUC).

2.1 Pengumpulan Data

Data yang digunakan pada penelitian ini adalah *Dataset Gallstone* yang diperoleh dari *UCI Machine Learning Repository* dengan alamat <https://archive.ics.uci.edu/dataset/1150/gallstone-1>. Dataset terdiri atas 319 data pasien dengan 38 atribut klinis yang meliputi karakteristik demografi, *Body Mass Index* (BMI), hasil pemeriksaan laboratorium, komorbiditas, serta variabel target yang menunjukkan status penyakit batu empedu. Dataset dimuat menggunakan pustaka *Pandas* pada Python, kemudian dilakukan pemeriksaan terhadap struktur data, tipe atribut, serta distribusi kelas sebelum memasuki tahap *preprocessing* [12]. Penggunaan dataset terbuka dari *UCI Machine Learning Repository* memungkinkan penelitian dilakukan secara objektif dan dapat direplikasi oleh peneliti lain.

2.2 Objek Penelitian

Objek penelitian adalah klasifikasi penyakit batu empedu berdasarkan data klinis pasien menggunakan *machine learning*. Variabel target yang digunakan adalah *Gallstone Status*, sedangkan atribut klinis lainnya digunakan sebagai prediktor. Tiga algoritma, yaitu SVM, *Random Forest*, dan XGBoost, digunakan untuk membangun model klasifikasi. Ketiga algoritma dipilih karena memiliki mekanisme pembelajaran yang berbeda, sehingga memungkinkan dilakukan perbandingan performa dan kemampuan generalisasi pada dataset dengan jumlah sampel terbatas.

2.3 Preprocessing Data

Tahap *preprocessing* bertujuan mempersiapkan data sebelum proses pelatihan model. Berdasarkan implementasi penelitian, *preprocessing* terdiri atas imputasi nilai yang hilang, seleksi fitur menggunakan ANOVA F-Value, standarisasi menggunakan *StandardScaler*, dan penanganan ketidakseimbangan kelas menggunakan SMOTE. Berbeda dengan metode sebelumnya yang hanya menjelaskan bahwa *preprocessing* dilakukan pada *training data*, pada penelitian ini setiap tahapan *preprocessing* yang mempelajari karakteristik data ditempatkan dalam *pipeline*. Pipeline tersebut digunakan dalam proses *inner cross-validation* sehingga parameter yang diperoleh dari proses *preprocessing* hanya berasal dari data pelatihan pada *fold* yang bersangkutan. Pendekatan ini digunakan untuk mencegah terjadinya *data leakage*.

a. Seleksi Fitur

Seleksi fitur dilakukan menggunakan *SelectKBest* dengan fungsi *f_classif* untuk memilih 15 atribut terbaik berdasarkan ANOVA *F-Value* [13]. Nilai *F-score* menunjukkan perbandingan variasi antar kelas dengan variasi di dalam kelas; semakin besar nilainya, semakin kuat perbedaan atribut antar kelas. Perhitungan ANOVA menggunakan komponen *Mean Square Between Groups* (MS_B) dan *Mean Square Within Groups* (MS_W), dengan jumlah kelas (C), jumlah data (N), jumlah data pada kelas ke- i (n_i), rata-rata atribut pada kelas ke- i ($\bar{x}_i$), dan rata-rata keseluruhan atribut ($\bar{x}$). Secara ringkas, nilai tersebut digunakan untuk memperoleh *F-score* sebagai dasar pemeringkatan fitur.

$$F = \frac{MS_B}{MS_W} \quad (1)$$

Pada penelitian ini, pemilihan fitur dilakukan kembali pada setiap *training fold* sehingga fitur yang dipilih tidak dipengaruhi oleh data validasi. Dengan cara tersebut, 15 fitur terbaik yang dihasilkan benar-benar berasal dari informasi data pelatihan pada *fold* yang bersangkutan.

b. Standardisasi Data

Standardisasi dilakukan menggunakan *StandardScaler* untuk menyamakan skala atribut numerik. Tahapan ini penting terutama bagi SVM karena perbedaan rentang antarfitur dapat memengaruhi pembentukan batas keputusan. Nilai setiap atribut ditransformasikan berdasarkan rata-rata (μ) dan simpangan baku (σ) yang dihitung dari data pelatihan pada *fold* tersebut. Proses standardisasi dihitung menggunakan persamaan berikut.

$$z = \frac{x - \mu}{\sigma} \quad (4)$$

Parameter μ dan σ tidak dihitung dari data validasi. Setelah parameter diperoleh dari *training fold*, parameter yang sama digunakan untuk mentransformasikan data validasi pada *fold* tersebut sehingga informasi validasi tetap terpisah dari proses pembelajaran.

c. Penanganan Ketidakseimbangan Kelas

Ketidakseimbangan kelas ditangani menggunakan *Synthetic Minority Over-sampling Technique* (SMOTE). SMOTE membentuk sampel sintetis pada kelas minoritas dengan melakukan interpolasi antara suatu sampel dan *nearest neighbor*-nya. Secara umum, sampel sintetis dibentuk menggunakan nilai sampel minoritas (X_i), tetangga terdekatnya (x_{zi}), dan bilangan acak λ pada interval 0 sampai 1. data sintetis dihitung menggunakan persamaan:

$$x_{baru} = x_i + \lambda(x_{zi} - x_i) \quad (5)$$

Pada penelitian ini, SMOTE tidak diterapkan pada keseluruhan dataset sebelum *cross-validation*, tetapi ditempatkan dalam *pipeline* bersama imputasi, *StandardScaler*, dan *SelectKBest*. Pipeline dengan urutan Imputasi

→ StandardScaler → SelectKBest (ANOVA *F-Value*) → SMOTE → Model diterapkan pada *inner Stratified 5-Fold Cross Validation*. Pada setiap *inner fold*, SMOTE hanya diterapkan pada *training fold*, sedangkan *validation fold* tetap menggunakan data asli. *Outer validation fold* juga tidak digunakan dalam proses SMOTE dan hanya digunakan untuk evaluasi akhir model. Dengan demikian, informasi dari data validasi tidak digunakan dalam proses *oversampling*, sehingga risiko *data leakage* dapat diminimalkan.

2.4 Pemodelan

Tahap pemodelan menggunakan SVM, *Random Forest*, dan XGBoost. Ketiga algoritma dimasukkan ke dalam *pipeline* bersama tahapan *preprocessing* sehingga seluruh proses pembelajaran mengikuti pembagian *fold*. Konfigurasi parameter setiap algoritma dipilih menggunakan *GridSearchCV* berdasarkan ROC-AUC pada *inner Stratified 5-Fold Cross Validation* [14]. Konfigurasi terbaik kemudian digunakan untuk melatih model pada *outer training fold* dan menghasilkan prediksi pada *outer validation fold*.

a. Support Vector Machine (SVM)

Support Vector Machine (SVM) merupakan algoritma *supervised learning* yang digunakan untuk melakukan klasifikasi dengan membentuk *hyperplane* optimal sebagai pemisah antar kelas. Algoritma ini mencari batas keputusan (*decision boundary*) dengan margin terbesar sehingga mampu memisahkan data secara optimal [15]. Kemampuan tersebut menjadikan SVM efektif dalam menangani data berdimensi tinggi dan menghasilkan kemampuan generalisasi yang baik. Pada penelitian ini, SVM digunakan untuk membangun model diagnosis penyakit batu empedu berdasarkan atribut klinis pasien. Data numerik terlebih dahulu melalui proses standarisasi agar skala setiap atribut seimbang. SVM dipilih karena mampu mengidentifikasi pola hubungan antarvariabel klinis dan banyak diterapkan pada klasifikasi penyakit berbasis *machine learning* [16]. Konfigurasi parameter SVM ditentukan menggunakan *GridSearchCV* pada *inner cross-validation*, kemudian model terbaik digunakan pada *outer validation fold* untuk memperoleh estimasi performa.

b. Random Forest

Random Forest merupakan algoritma *ensemble learning* yang dibangun dari sekumpulan *decision tree*. Setiap pohon dibentuk menggunakan sampel data melalui teknik *bootstrap sampling*, dengan pemilihan atribut secara acak. Hasil klasifikasi akhir ditentukan berdasarkan *majority voting* dari seluruh pohon [17].

Pada penelitian ini, *Random Forest* digunakan untuk membangun model diagnosis penyakit batu empedu. Algoritma ini dipilih karena mampu menangani hubungan nonlinier antarfitur, relatif tahan terhadap *noise*, dan dapat mengurangi risiko *overfitting* dibandingkan *decision tree* tunggal. Selain itu, *Random Forest* mampu memberikan performa yang stabil pada data medis dengan jumlah atribut yang cukup banyak [18]. Konfigurasi parameter ditentukan melalui *GridSearchCV* pada *inner Stratified 5-Fold Cross Validation*, kemudian model terbaik dilatih pada *outer training fold* dan dievaluasi pada *outer validation fold*.

c. Extreme Gradient Boosting (XGBoost)

Extreme Gradient Boosting (XGBoost) merupakan pengembangan dari *Gradient Boosting Decision Tree* (GBDT) yang membangun model secara bertahap dengan memperbaiki kesalahan prediksi pada iterasi sebelumnya. Algoritma ini menggunakan fungsi kerugian (*loss function*) dan regularisasi untuk meningkatkan akurasi sekaligus mengurangi risiko *overfitting*. XGBoost juga mampu menangani hubungan nonlinier antarvariabel dengan baik [19]. Pada penelitian ini, XGBoost digunakan untuk membandingkan performanya dengan SVM dan *Random Forest* dalam diagnosis penyakit batu empedu. Penggunaan XGBoost diharapkan mampu menghasilkan model klasifikasi yang baik melalui proses pembelajaran bertahap [20]. Konfigurasi parameter ditentukan menggunakan *GridSearchCV* pada *inner Stratified 5-Fold Cross Validation*, kemudian model terbaik digunakan untuk prediksi pada *outer validation fold*. *outer validation fold* tidak digunakan dalam proses pemilihan parameter.

2.5 Validasi Model

Validasi model menggunakan *Nested Stratified 5-Fold Cross Validation*. Pada tahap *outer cross-validation*, dataset dibagi menjadi lima *fold* dengan mempertahankan proporsi kelas. Empat *fold* digunakan sebagai *outer training data* dan satu *fold* sebagai *outer validation data*. Pada setiap *outer training data*, dilakukan *inner Stratified 5-Fold Cross Validation* menggunakan *GridSearchCV* untuk memilih konfigurasi parameter terbaik. Seluruh *preprocessing* dan SMOTE dilakukan di dalam *pipeline* pada *inner fold*, sehingga *outer validation fold* tetap disimpan untuk evaluasi yang benar-benar independen. Setelah konfigurasi terbaik diperoleh, *pipeline* dengan konfigurasi tersebut dilatih kembali menggunakan seluruh *outer training fold*. Model kemudian memprediksi *outer validation fold*. Prosedur ini diulang sebanyak lima kali sehingga setiap data menjadi bagian validasi tepat satu kali. Hasil lima iterasi tersebut menjadi dasar pengukuran performa dan perbandingan antaralgoritma [21]. Pendekatan ini juga menjawab risiko *data leakage* yang dapat terjadi apabila SMOTE, imputasi, standarisasi, atau seleksi fitur dilakukan sekali sebelum *cross-validation*.

2.6 Evaluasi Model

Tahap evaluasi dilakukan untuk mengetahui performa masing-masing algoritma dalam mendiagnosis penyakit batu empedu berdasarkan data klinis pasien. Evaluasi dilakukan menggunakan *confusion matrix* yang membandingkan hasil prediksi model dengan label sebenarnya pada data uji. Berdasarkan *confusion matrix*, performa model diukur

menggunakan beberapa metrik evaluasi, yaitu *accuracy*, *precision*, *recall*, *F1-score*, dan *Area Under Curve* (AUC). Penggunaan beberapa metrik evaluasi bertujuan untuk memberikan gambaran yang lebih komprehensif mengenai kemampuan model dalam melakukan klasifikasi dibandingkan hanya menggunakan nilai akurasi [22]. Perhitungan metrik menggunakan komponen *True Positive* (TP), *True Negative* (TN), *False Positive* (FP), dan *False Negative* (FN). TP adalah data positif yang diprediksi benar, TN adalah data negatif yang diprediksi benar, FP adalah data negatif yang diprediksi sebagai positif, dan FN adalah data positif yang diprediksi sebagai negatif [23]. Rumus metrik dapat dinyatakan sebagai berikut.

$$\text{Akurasi} = \frac{TP+TN}{TP+TN+FP+FN} \quad (6)$$

$$\text{Precision} = \frac{TP}{TP+FP} \quad (7)$$

$$\text{Recall} = \frac{TP}{TP+FN} \quad (8)$$

$$\text{F1 - Score} = \frac{2 \times \text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \quad (9)$$

Nilai *accuracy*, *precision*, *recall*, *F1-score*, dan *AUC* yang diperoleh dari masing-masing algoritma selanjutnya dibandingkan untuk menentukan model klasifikasi yang memiliki performa terbaik dalam mendiagnosis penyakit batu empedu. Nilai *AUC* diperoleh dari luas area di bawah kurva *Receiver Operating Characteristic* (ROC), yang menunjukkan kemampuan model dalam membedakan kelas positif dan kelas negatif. Semakin tinggi nilai *AUC*, semakin baik kemampuan model dalam melakukan klasifikasi.

2.7 Penentuan Model Terbaik

Penentuan model terbaik dilakukan dengan membandingkan performa SVM, Random Forest, dan XGBoost berdasarkan nilai *accuracy*, *precision*, *recall*, *F1-score*, dan *AUC* yang diperoleh dari *outer validation fold*. Model tidak ditentukan hanya berdasarkan nilai *accuracy*, tetapi juga mempertimbangkan *precision*, *recall*, *F1-score*, dan *AUC*. Penggunaan beberapa metrik tersebut bertujuan memberikan penilaian yang lebih komprehensif terhadap kemampuan model dalam mengklasifikasikan pasien dengan status penyakit batu empedu. Nilai rata-rata dan standar deviasi dari lima *outer fold* digunakan sebagai dasar perbandingan performa. Model dengan performa yang lebih baik secara konsisten pada sebagian besar metrik dipertimbangkan sebagai model terbaik.

3. HASIL DAN PEMBAHASAN

3.1 Hasil Klasifikasi Menggunakan Support Vector Machine

Model *Support Vector Machine* (SVM) digunakan sebagai salah satu model klasifikasi untuk mengidentifikasi penyakit batu empedu berdasarkan atribut klinis pasien. Model SVM menggunakan *kernel Radial Basis Function* (RBF) dan dilatih menggunakan konfigurasi terbaik yang diperoleh melalui proses *inner Stratified 5-Fold Cross Validation*. Evaluasi performa model dilakukan berdasarkan hasil prediksi pada *outer validation fold*. Nilai evaluasi kemudian dihitung berdasarkan keseluruhan prediksi dari lima *outer fold* sehingga diperoleh gambaran performa model yang lebih objektif. Hasil evaluasi model SVM ditunjukkan pada Tabel 1.

Tabel 1. Hasil Evaluasi SVM

Metrik	Mean	Std
Accuracy	73,70%	7,69%
Precision	77,10%	7,96%
Recall	67,22%	10,16%
F1-Score	71,57%	8,33%
ROC-AUC	83,01%	5,91%

Berdasarkan Tabel 1, SVM memperoleh *accuracy* rata-rata sebesar 73,70% dengan standar deviasi 7,69%. Nilai *precision* mencapai 77,10%, sedangkan *recall* sebesar 67,22%. F1-Score yang diperoleh sebesar 71,57%, sementara ROC-AUC mencapai 83,01%. Nilai *recall* yang lebih rendah dibandingkan *precision* menunjukkan bahwa kemampuan SVM dalam mengenali seluruh data positif masih lebih terbatas. Variasi nilai antar-*fold* juga terlihat dari standar deviasi, terutama pada *recall* yang mencapai 10,16%. Secara keseluruhan, SVM mampu memberikan kemampuan diskriminasi yang cukup baik berdasarkan nilai ROC-AUC. Namun, nilai *accuracy*, *recall*, dan F1-Score masih lebih rendah dibandingkan dua algoritma lainnya.

3.2 Hasil Klasifikasi Menggunakan Random Forest

Hasil pengujian Random Forest memperlihatkan peningkatan performa dibandingkan SVM pada beberapa metrik utama. Model ini memberikan nilai *accuracy*, *recall*, F1-Score, dan ROC-AUC yang lebih tinggi dibandingkan hasil SVM. Hasil pengujian *Random Forest* ditampilkan pada Tabel 2.

Tabel 2. Hasil Evaluasi Random Forest

Metrik	Mean	Std
Accuracy	77,43%	4,06%
Precision	78,11%	4,19%
Recall	75,99%	7,71%
F1-Score	76,83%	4,78%
ROC-AUC	84,71%	5,16%

Berdasarkan Tabel 2, algoritma Random Forest memperoleh *accuracy* rata-rata sebesar 77,43%, dengan standar deviasi sebesar 4,06%. Nilai *precision* sebesar 78,11%, *recall* sebesar 75,99%, dan F1-Score sebesar 76,83%. Sementara itu, ROC-AUC mencapai 84,71%. Jika dibandingkan dengan SVM, Random Forest menunjukkan peningkatan *accuracy* sebesar sekitar 3,73 poin persentase. Peningkatan juga terlihat pada *recall* dan F1-Score. Nilai *recall* yang lebih tinggi menunjukkan bahwa Random Forest mampu mengenali kelas positif dengan lebih baik dibandingkan SVM. Standar deviasi *accuracy* sebesar 4,06% menunjukkan variasi performa antar-*fold* yang relatif lebih kecil dibandingkan SVM. Hal ini menunjukkan bahwa performa Random Forest pada pembagian data yang berbeda cenderung lebih stabil.

3.3 Hasil Klasifikasi Menggunakan XGBoost

Pada pengujian XGBoost, diperoleh hasil yang relatif dekat dengan Random Forest. Meskipun XGBoost menghasilkan nilai *precision* yang paling tinggi, beberapa metrik lainnya masih berada sedikit di bawah Random Forest. Hasil pengujian XGBoost ditampilkan pada Tabel 3.

Tabel 3. Hasil Evaluasi XGBoost

Metrik	Mean	Std
Accuracy	77,11%	3,26%
Precision	79,68%	4,06%
Recall	72,84%	8,55%
F1-Score	75,77%	4,24%
ROC-AUC	83,23%	5,08%

Berdasarkan Tabel 3, XGBoost memperoleh *accuracy* rata-rata sebesar 77,11%, dengan standar deviasi 3,26%. Nilai *precision* sebesar 79,68% merupakan yang tertinggi di antara ketiga algoritma. Sementara itu, *recall* sebesar 72,84%, F1-Score sebesar 75,77%, dan ROC-AUC sebesar 83,23%. Nilai standar deviasi *accuracy* sebesar 3,26% merupakan yang paling rendah dibandingkan SVM dan Random Forest. Hal ini menunjukkan bahwa hasil XGBoost memiliki variasi performa antar-*fold* yang relatif lebih kecil. Namun, dari sisi *accuracy*, XGBoost memperoleh nilai 77,11%, sedikit lebih rendah dibandingkan Random Forest yang memperoleh 77,43%. Perbedaan tersebut hanya sebesar 0,32 poin persentase. Dengan demikian, XGBoost tidak dapat dinyatakan sebagai algoritma dengan performa keseluruhan tertinggi. XGBoost memiliki nilai *precision* tertinggi, sedangkan Random Forest memperoleh nilai *accuracy*, *recall*, F1-Score, dan ROC-AUC yang lebih tinggi.

3.4 Perbandingan Performa Model

Untuk memperoleh gambaran menyeluruh mengenai performa setiap algoritma, dilakukan perbandingan berdasarkan metrik *accuracy*, *precision*, *recall*, *F1-score*, dan *Area Under Curve (AUC)*. Hasil perbandingan ketiga algoritma ditunjukkan pada Tabel 4.

Tabel 4. Perbandingan Performa Model

Model	Accuracy	Precision	Recall	F1-Score	ROC-AUC
SVM	73,70%	77,10%	67,22%	71,57%	83,01%
Random Forest	77,43%	78,11%	75,99%	76,83%	84,71%
XGBoost	77,11%	79,68%	72,84%	75,77%	83,23%

Berdasarkan Tabel 4, Random Forest memperoleh nilai *accuracy* tertinggi sebesar 77,43%, diikuti XGBoost sebesar 77,11% dan SVM sebesar 73,70%. Pada metrik *precision*, XGBoost memperoleh nilai tertinggi sebesar 79,68%, sedangkan Random Forest sebesar 78,11% dan SVM sebesar 77,10%. Sementara itu, Random Forest memperoleh nilai *recall* tertinggi sebesar 75,99%, diikuti XGBoost 72,84% dan SVM 67,22%. F1-Score tertinggi juga diperoleh Random Forest sebesar 76,83%, sedangkan XGBoost memperoleh 75,77% dan SVM 71,57%. Pada ROC-AUC, Random Forest kembali memperoleh nilai tertinggi sebesar 84,71%. Dengan demikian, berdasarkan nilai rata-rata, Random Forest memberikan performa paling tinggi pada empat dari lima metrik, sedangkan XGBoost unggul pada *precision*. Perbedaan *accuracy* antara Random Forest dan XGBoost hanya sebesar 0,32 poin persentase. Selisih tersebut sangat kecil sehingga perbedaan nilai rata-rata belum cukup untuk menyatakan bahwa Random Forest memiliki keunggulan yang signifikan secara statistik. Oleh karena itu, dilakukan pengujian statistik terhadap hasil setiap *fold*.

3.5 Uji Signifikansi Statistik

Nilai rata-rata pada Tabel 4 menunjukkan adanya perbedaan performa antaralgoritma. Namun, perbedaan tersebut belum menunjukkan apakah model memang memiliki performa yang berbeda secara statistik. Hal ini terutama terlihat pada Random Forest dan XGBoost yang memiliki selisih *accuracy* hanya 0,32 poin persentase. Untuk mengetahui apakah perbedaan performa tersebut signifikan, dilakukan *Paired t-test* dan *Wilcoxon signed-rank test* menggunakan hasil evaluasi pada setiap *outer fold*. Tingkat signifikansi yang digunakan adalah $\alpha = 0,05$. Hasil Uji Signifikansi Statistik dapat dilihat pada Tabel 5.

Tabel 5. Hasil Uji Signifikansi Statistik

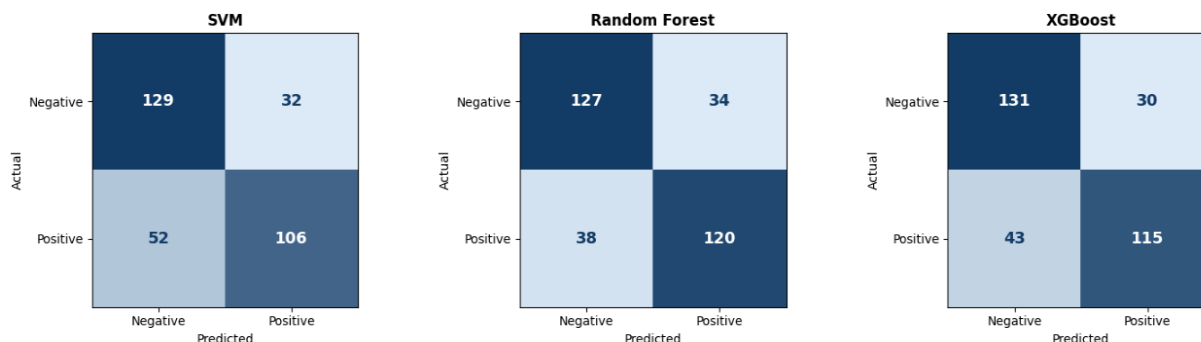
Metrik	Perbandingan	Paired t-test	Wilcoxon	Keputusan
Accuracy	XGBoost vs Random Forest	0,8584	1,0000	Tidak signifikan
F1-Score	XGBoost vs Random Forest	0,6566	1,0000	Tidak signifikan
ROC-AUC	XGBoost vs Random Forest	0,1320	0,1875	Tidak signifikan
Accuracy	XGBoost vs SVM	0,2999	0,4375	Tidak signifikan
F1-Score	XGBoost vs SVM	0,2446	0,3125	Tidak signifikan
ROC-AUC	XGBoost vs SVM	0,8733	0,8125	Tidak signifikan
Accuracy	Random Forest vs SVM	0,2696	0,3750	Tidak signifikan
F1-Score	Random Forest vs SVM	0,2075	0,3125	Tidak signifikan
ROC-AUC	Random Forest vs SVM	0,1480	0,0625	Tidak signifikan

Berdasarkan Tabel 5, seluruh *p-value* lebih besar dari 0,05. Dengan demikian, tidak terdapat perbedaan performa yang signifikan secara statistik pada ketiga pasangan algoritma yang diuji. Pada perbandingan Random Forest dan XGBoost, *Paired t-test* untuk *accuracy* menghasilkan *p-value* 0,8584, sedangkan *Wilcoxon signed-rank test* menghasilkan 1,0000. Kedua nilai tersebut jauh di atas 0,05. Hasil ini menunjukkan bahwa selisih *accuracy* Random Forest sebesar 77,43% dan XGBoost sebesar 77,11% tidak signifikan secara statistik. Dengan demikian, Random Forest tidak dapat dinyatakan secara statistik lebih unggul daripada XGBoost meskipun memiliki nilai rata-rata *accuracy* yang sedikit lebih tinggi. Hal yang sama terlihat pada F1-Score dan ROC-AUC. Perbandingan Random Forest dan XGBoost menghasilkan *p-value* masing-masing sebesar 0,6566 dan 0,1320 pada *Paired t-test*, sedangkan *Wilcoxon* menghasilkan 1,0000 dan 0,1875. Berdasarkan hasil pengujian, tidak terdapat bukti yang cukup untuk menyatakan bahwa salah satu dari ketiga algoritma memiliki performa yang secara statistik lebih unggul pada tingkat signifikansi 5%. Hasil ini secara langsung menjawab kritik reviewer bahwa perbedaan performa perlu diuji untuk mengetahui apakah perbedaan tersebut hanya merupakan variasi akibat pembagian *fold*.

3.6 Confusion Matrix

Selain menggunakan nilai *accuracy*, *precision*, *recall*, F1-Score, dan ROC-AUC, evaluasi model juga dilakukan menggunakan *confusion matrix*. Analisis ini digunakan untuk melihat lebih rinci hasil klasifikasi yang dilakukan oleh masing-masing algoritma, terutama dalam membedakan prediksi yang benar dan prediksi yang masih mengalami kesalahan. *Confusion matrix* menampilkan jumlah data yang diklasifikasikan dengan benar maupun salah pada masing-masing kelas, sehingga dapat memberikan gambaran mengenai pola kesalahan yang dihasilkan oleh model. Pada penelitian ini, *confusion matrix* diperoleh berdasarkan hasil prediksi pada seluruh *outer validation fold*. Dengan demikian, hasil yang ditampilkan tidak berasal dari satu pembagian data saja, tetapi merupakan akumulasi prediksi selama proses *outer cross-validation*. Hasil *confusion matrix* untuk SVM, Random Forest, dan XGBoost ditampilkan pada Gambar 2.

Berdasarkan Gambar 2, Confusion matrix memberikan gambaran mengenai kemampuan setiap model dalam membedakan kelas negatif dan positif. Hasil tersebut terdiri atas *True Negative* (TN), *True Positive* (TP), *False Positive* (FP), dan *False Negative* (FN). Berdasarkan hasil evaluasi, masih terdapat data yang mengalami kesalahan klasifikasi pada ketiga algoritma. Kesalahan tersebut menunjukkan bahwa model belum mampu mengidentifikasi seluruh data dengan sempurna.

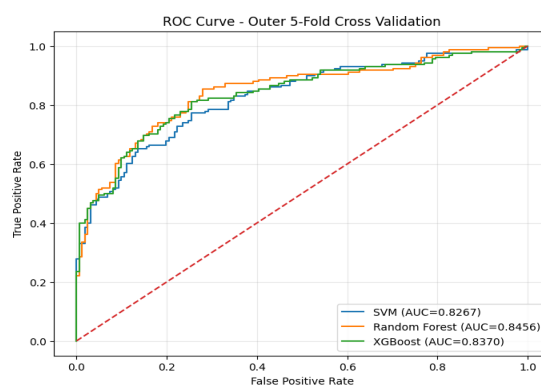


Gambar 2. Confusion Matrix SVM, Random Forest, dan XGBoost

Nilai *recall* yang lebih rendah pada SVM dibandingkan Random Forest dan XGBoost juga menunjukkan adanya keterbatasan dalam mengenali kelas positif. Sebaliknya, Random Forest memiliki *recall* tertinggi sebesar 75,99%, sehingga model tersebut mampu mengenali proporsi data positif yang lebih besar dibandingkan dua algoritma lainnya. Dalam konteks klasifikasi penyakit, nilai *recall* perlu diperhatikan karena kesalahan berupa *false negative* berarti data yang sebenarnya termasuk kelas positif diprediksi sebagai kelas negatif. Oleh karena itu, evaluasi tidak hanya didasarkan pada *accuracy*, tetapi juga mempertimbangkan *recall* dan F1-Score.

3.7 Kurva ROC

Evaluasi selanjutnya dilakukan menggunakan *Receiver Operating Characteristic* (ROC) untuk melihat kemampuan masing-masing model dalam membedakan dua kelas pada berbagai nilai *threshold*. Kurva ROC menggambarkan hubungan antara *True Positive Rate* dan *False Positive Rate*. Semakin baik kemampuan model dalam membedakan kedua kelas, semakin dekat kurva ROC dengan bagian kiri atas grafik. Selain bentuk kurva, performa model juga dapat dilihat melalui nilai *Area Under the Curve* (AUC). Nilai AUC digunakan sebagai ukuran kemampuan diskriminasi model secara keseluruhan dan menjadi salah satu metrik yang digunakan dalam penelitian ini. Oleh karena itu, kurva ROC digunakan sebagai pelengkap hasil *accuracy*, *precision*, *recall*, dan F1-Score agar kemampuan model dalam membedakan kedua kelas dapat dilihat dari sudut pandang yang berbeda. Kurva ROC yang dihasilkan dari prediksi *outer validation fold* untuk ketiga algoritma ditampilkan pada Gambar 3.2.

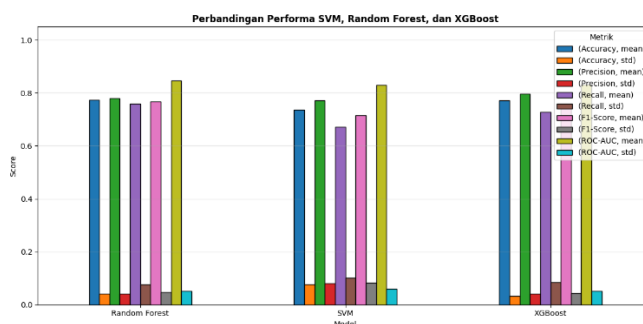


Gambar 3. Kurva ROC SVM, Random Forest, dan XGBoost

Berdasarkan Gambar 3, hasil evaluasi menunjukkan bahwa ketiga algoritma memiliki nilai ROC-AUC di atas 0,80. Random Forest memperoleh nilai tertinggi sebesar 0,8471, XGBoost sebesar 0,8323, dan SVM sebesar 0,8301. Nilai tersebut menunjukkan bahwa ketiga model memiliki kemampuan diskriminasi yang cukup baik dalam membedakan kedua kelas pada dataset yang digunakan. Random Forest memiliki nilai ROC-AUC paling tinggi, tetapi perbedaannya dengan XGBoost hanya sebesar 0,0148. Hasil uji statistik menunjukkan bahwa perbedaan ROC-AUC antara Random Forest dan XGBoost tidak signifikan. Oleh karena itu, nilai AUC yang lebih tinggi pada Random Forest tidak dapat diinterpretasikan sebagai keunggulan statistik yang mutlak.

3.8 Analisis Feature Importance

Selain mengevaluasi performa klasifikasi, penelitian ini juga melakukan analisis terhadap kontribusi atribut yang digunakan dalam pembentukan model. Analisis *feature importance* dilakukan untuk mengetahui atribut mana yang memberikan kontribusi relatif lebih besar terhadap proses prediksi, khususnya pada model berbasis *tree ensemble*, yaitu Random Forest dan XGBoost. Analisis ini penting karena hasil klasifikasi tidak hanya dilihat berdasarkan nilai metrik evaluasi, tetapi juga dapat memberikan informasi mengenai atribut yang paling banyak digunakan oleh model dalam menghasilkan keputusan. Dengan melihat urutan tingkat kepentingan atribut, dapat diketahui pola kontribusi fitur terhadap model tanpa mengubah proses klasifikasi yang telah dilakukan. Hasil *feature importance* dari model Random Forest dan XGBoost ditampilkan pada Gambar 3.3.



Gambar 4 Feature Importance Random Forest dan XGBoost

Berdasarkan Gambar 4, setiap atribut memiliki tingkat kontribusi yang berbeda terhadap model. Atribut dengan nilai *importance* lebih tinggi menunjukkan kontribusi relatif yang lebih besar dalam proses pembentukan prediksi dibandingkan atribut dengan nilai *importance* yang lebih rendah. Namun, nilai *feature importance* tidak dapat diartikan sebagai hubungan sebab-akibat antara suatu atribut dengan penyakit batu empedu. Nilai tersebut hanya menunjukkan tingkat kontribusi relatif atribut terhadap prediksi model. Oleh karena itu, hasil ini digunakan sebagai informasi pendukung untuk memahami model, bukan sebagai dasar untuk menentukan hubungan klinis secara langsung.

3.9 Pembahasan

Hasil eksperimen menunjukkan bahwa ketiga algoritma dapat digunakan untuk melakukan klasifikasi penyakit batu empedu. Namun, performa yang diperoleh masih berada di bawah 80%, dengan *accuracy* sebesar 73,70% pada SVM, 77,43% pada Random Forest, dan 77,11% pada XGBoost. Random Forest menghasilkan *accuracy* tertinggi, yaitu 77,43%, serta memperoleh nilai *recall*, F1-Score, dan ROC-AUC tertinggi. Sementara itu, XGBoost menghasilkan nilai *precision* tertinggi sebesar 79,68%. SVM menghasilkan performa terendah pada sebagian besar metrik. Perbedaan performa tersebut dapat dikaitkan dengan karakteristik masing-masing algoritma. SVM membentuk *hyperplane* untuk memisahkan kelas sehingga performanya dipengaruhi oleh struktur ruang fitur. Random Forest menggunakan beberapa *decision tree* melalui pendekatan *ensemble* sehingga dapat menangani variasi data dan hubungan nonlinier. Sementara itu, XGBoost menggunakan pendekatan *gradient boosting* yang membangun model secara bertahap dengan memperbaiki kesalahan prediksi pada proses berikutnya. Hasil Random Forest yang lebih tinggi pada *accuracy*, *recall*, F1-Score, dan ROC-AUC menunjukkan bahwa pendekatan *ensemble* berbasis *bagging* cukup sesuai untuk pola data klinis yang digunakan dalam penelitian ini. Sementara itu, tingginya *precision* XGBoost menunjukkan bahwa model tersebut relatif baik dalam menghasilkan prediksi positif yang tepat. Hasil penelitian ini berbeda dengan penelitian terdahulu pada Dataset Gallstone yang melaporkan *accuracy* sekitar 85,9% dan AUC di atas 0,90 pada beberapa algoritma klasifikasi [4], [7], [8]. Perbedaan tersebut dapat dipengaruhi oleh ukuran dan karakteristik dataset, preprocessing, seleksi fitur, konfigurasi parameter, serta metode validasi yang digunakan. Oleh karena itu, hasil penelitian ini tidak dibandingkan secara langsung berdasarkan nilai *accuracy* saja.

Selain itu, ukuran dataset yang digunakan dalam penelitian ini relatif terbatas, yaitu 319 data pasien dengan 38 atribut. Kondisi tersebut dapat menyebabkan variasi performa pada setiap *fold* dan membatasi kemampuan generalisasi model terhadap populasi yang lebih luas. Hal yang penting dari penelitian ini adalah penggunaan *nested cross-validation*. Proses pemilihan fitur dan preprocessing dilakukan pada data training sehingga validation fold tidak digunakan dalam proses pembelajaran. Pada setiap *outer fold*, data dibagi menjadi training dan validation, kemudian proses pemilihan konfigurasi dilakukan pada bagian training. SMOTE juga diterapkan hanya pada data training. Notebook secara eksplisit menerapkan urutan training fold → preprocessing → SMOTE → model → validation fold → evaluasi untuk mencegah informasi validation fold masuk ke proses pembelajaran. Pendekatan tersebut penting untuk mengatasi permasalahan *data leakage* yang menjadi perhatian reviewer. Dengan menempatkan proses preprocessing dan SMOTE pada data training, estimasi performa model pada validation fold menjadi lebih terkontrol.

Dari sisi statistik, Random Forest dan XGBoost memiliki performa yang sangat berdekatan. Selisih *accuracy* keduanya hanya 0,32 poin persentase. Hasil *Paired t-test* dan *Wilcoxon signed-rank test* juga menunjukkan bahwa perbedaan tersebut tidak signifikan. Dengan demikian, penelitian ini tidak menyatakan bahwa Random Forest merupakan algoritma yang secara mutlak paling unggul. Random Forest menghasilkan performa rata-rata tertinggi pada sebagian besar metrik, sedangkan XGBoost unggul pada *precision*. Namun, berdasarkan uji statistik, belum terdapat bukti bahwa perbedaan performa ketiga algoritma signifikan secara statistik. Temuan tersebut menjadi penting karena menjawab kritik terhadap klaim sebelumnya mengenai XGBoost sebagai algoritma terbaik. Hasil eksperimen terbaru menunjukkan bahwa klaim tersebut perlu diperbaiki berdasarkan hasil evaluasi yang sebenarnya.

Meskipun memberikan hasil yang cukup baik, penelitian ini masih memiliki beberapa keterbatasan. Dataset yang digunakan terdiri atas 319 data pasien dengan 38 atribut sehingga jumlah sampel yang relatif terbatas dapat memengaruhi kemampuan generalisasi model. Selain itu, performa ketiga algoritma masih berada di bawah 80%, dengan Random Forest sebagai model yang memperoleh *accuracy* tertinggi sebesar 77,43%. Hasil uji signifikansi juga menunjukkan bahwa perbedaan performa antaralgoritma tidak signifikan secara statistik. Oleh karena itu, hasil penelitian ini belum dapat dijadikan dasar untuk menyatakan bahwa satu algoritma memiliki keunggulan yang mutlak. Penelitian selanjutnya dapat menggunakan dataset dengan jumlah sampel yang lebih besar dan berasal dari sumber yang lebih beragam serta melakukan validasi eksternal untuk menguji kemampuan generalisasi model.

4. KESIMPULAN

Penelitian ini telah membangun dan membandingkan model klasifikasi penyakit batu empedu menggunakan algoritma *Support Vector Machine* (SVM), *Random Forest*, dan *Extreme Gradient Boosting* (XGBoost) pada *Dataset Gallstone* dari UCI Machine Learning Repository. Proses penelitian dilakukan melalui tahapan preprocessing yang mencakup imputasi, standardisasi, seleksi fitur menggunakan *ANOVA F-Value*, serta penanganan ketidakseimbangan kelas menggunakan *Synthetic Minority Over-sampling Technique* (SMOTE). Seluruh tahapan preprocessing dan SMOTE diterapkan pada data *training fold* dalam skema *nested cross-validation* untuk menghindari *data leakage*.

Berdasarkan hasil evaluasi, Random Forest memperoleh performa rata-rata tertinggi pada sebagian besar metrik, dengan *accuracy* sebesar 77,43%, *precision* 78,11%, *recall* 75,99%, F1-Score 76,83%, dan ROC-AUC 84,71%. XGBoost memperoleh *accuracy* sebesar 77,11%, *precision* 79,68%, *recall* 72,84%, F1-Score 75,77%, dan ROC-AUC 83,23%. Sementara itu, SVM memperoleh *accuracy* sebesar 73,70%, *precision* 77,10%, *recall* 67,22%, F1-Score 71,57%, dan ROC-AUC 83,01%. Hasil tersebut menunjukkan bahwa Random Forest memiliki nilai *accuracy*, *recall*, F1-Score, dan ROC-AUC yang lebih tinggi dibandingkan XGBoost dan SVM, sedangkan XGBoost memperoleh nilai *precision* tertinggi. Namun, perbedaan performa tersebut tidak terbukti signifikan secara statistik. Hasil *Paired t-test* dan *Wilcoxon signed-rank test* menunjukkan nilai *p-value* lebih besar dari 0,05 pada seluruh perbandingan model yang diuji. Dengan demikian, Random Forest tidak dapat dinyatakan secara statistik lebih unggul dibandingkan XGBoost maupun SVM. Hasil penelitian juga menunjukkan bahwa performa ketiga model masih berada di bawah 80%. Kondisi tersebut menunjukkan bahwa model yang dihasilkan masih memiliki keterbatasan dalam melakukan klasifikasi secara konsisten. Selain itu, dataset yang digunakan hanya terdiri atas 319 data pasien, sehingga hasil penelitian memiliki keterbatasan dalam hal generalisasi terhadap populasi pasien yang lebih luas. Kontribusi penelitian ini terletak pada penyediaan evaluasi komparatif SVM, *Random Forest*, dan XGBoost menggunakan *Dataset Gallstone* dengan tahapan preprocessing dan validasi yang diterapkan secara konsisten. Hasil penelitian memberikan gambaran mengenai karakteristik kinerja masing-masing algoritma dan dapat menjadi dasar dalam mempertimbangkan pemilihan metode klasifikasi penyakit batu empedu pada penelitian selanjutnya. Oleh karena itu, model pada penelitian ini belum dapat digunakan sebagai dasar diagnosis klinis secara langsung, tetapi dapat menjadi dasar untuk penelitian lebih lanjut mengenai penerapan metode *machine learning* pada data klinis penyakit batu empedu.

5. PERNYATAAN PENULIS

5.1 Kontribusi Penulis

Konseptualisasi: A. D. E., dan E. R. S.;

Metodologi: A. D. E., dan E. R. S.;

Perangkat Lunak: A. D. E.;

Validasi: A. D. E., dan E. R. S.;

Analisis Formal: A. D. E., dan E. R. S.;

Investigasi: A. D. E., dan E. R. S.;

Sumber Daya: A. D. E., dan E. R. S.;

Kurasi Data: A. D. E.;

Penulisan Draf Awal: A. D. E.;

Penulisan, Peninjauan, dan Penyuntingan: A. D. E., dan E. R. S.;

Visualisasi: A. D. E.;

Seluruh penulis telah membaca dan menyetujui versi manuskrip yang telah diterbitkan.

5.2 Ketersediaan Data

Data yang disajikan dalam penelitian ini tersedia berdasarkan permintaan kepada penulis korespondensi.

5.3 Pendanaan

Penelitian ini tidak menerima pendanaan eksternal.

5.4 Pernyataan Dewan Peninjau Institusional

Tidak Berlaku (Not applicable).

5.5 Pernyataan Persetujuan Berdasarkan Informasi

Tidak Berlaku (Not applicable).

5.6 Pernyataan Konflik Kepentingan

Para penulis menyatakan bahwa mereka tidak memiliki kepentingan keuangan yang bersaing atau hubungan pribadi yang diketahui yang dapat dianggap memengaruhi pekerjaan yang dilaporkan dalam artikel ini.

5.7 Penggunaan AI Generatif dan Teknologi Berbantuan AI dalam Proses Penulisan

Para penulis menggunakan AI Generatif untuk membantu meningkatkan kejelasan dan penyusunan bahasa dalam artikel ini. Para penulis telah meninjau dan menyunting konten yang dihasilkan dengan bantuan AI serta bertanggung jawab penuh atas versi akhir publikasi.

REFERENCES

- [1] A. H. Andini, M. Romdhoni, and F. Oktavrisa, "Karakteristik Pasien Batu Empedu Yang di Rawat di RSUD Waled Periode 2019-2022," *Syntax Literate ; Jurnal Ilmiah Indonesia*, vol. 7, no. 9, pp. 15291–15303, Dec. 2023, doi: 10.36418/syntax-literate.v7i9.14253.

- [2] X. Wang *et al.*, “Global Epidemiology of Gallstones in the 21st Century: A Systematic Review and Meta-Analysis,” *Clinical Gastroenterology and Hepatology*, vol. 22, no. 8, pp. 1586–1595, Aug. 2024, doi: 10.1016/j.cgh.2024.01.051.
- [3] F. Lammert *et al.*, “EASL Clinical Practice Guidelines on the prevention, diagnosis and treatment of gallstones,” *J. Hepatol.*, vol. 65, no. 1, pp. 146–181, Jul. 2022, doi: 10.1016/j.jhep.2016.03.005.
- [4] S. A. Alowais *et al.*, “Revolutionizing healthcare: the role of artificial intelligence in clinical practice,” Dec. 01, 2023, *BioMed Central Ltd.* doi: 10.1186/s12909-023-04698-z.
- [5] A. Rindang, A. Amin, and B. Satria, “Analisis Komparatif Model Logistic Regression, Decision Tree, Random Forest, Support Vector Machine, dan K-Nearest Neighbor untuk Klasifikasi Penyakit Batu Empedu Menggunakan Machine Learning,” *Bulletin of Information System Research (BIOS)*, vol. 4, no. 2, pp. 90–101, Apr. 2026, doi: 10.62866/bios.v4i2.281.
- [6] Yoan, D. Marta Putria, A. Rahmawatia, and B. Wajhi Akramunnas, “Perbandingan Algoritma CatBoost dan XGBoost dalam Klasifikasi Penyakit Jantung,” *Jurnal APTEK (Artikel Ilmiah Aplikasi Teknologi)*, vol. 15, no. 2, pp. 126–133, Jun. 2023, doi: <https://doi.org/10.30606/aptek.v15i2.1930>.
- [7] A. Rindang, A. Amin, and B. Satria, “Analisis Komparatif Model Logistic Regression, Decision Tree, Random Forest, Support Vector Machine, dan K-Nearest Neighbor untuk Klasifikasi Penyakit Batu Empedu Menggunakan Machine Learning,” *Bulletin of Information System Research (BIOS)*, vol. 4, no. 2, 2026, doi: 10.62866/bios.v4i2.281.
- [8] A. Setiawan, D. Mulia Hutabalian, R. Irnanda, H. Fredi, and P. Studi Ilmu Komputer, “Evaluasi Perbandingan Kinerja Model Machine Learning Untuk Prediksi Diabetes: Studi Kasus Xgboost, Random Forest, Dan Svm 1,” *Jurnal Informatika dan Komputer*, vol. 12, no. 2, pp. 105–115, Dec. 2024, doi: <https://doi.org/10.56689/infokom.v12i2.2350>.
- [9] S. A. Hicks *et al.*, “On evaluation metrics for medical applications of artificial intelligence,” *Sci. Rep.*, vol. 12, no. 1, Dec. 2022, doi: 10.1038/s41598-022-09954-8.
- [10] D. Mpouziotas, J. Besharat, I. G. Tsoulos, and C. Stylios, “AliAmvra—Enhancing Customer Experience through the Application of Machine Learning Techniques for Survey Data Assessment and Analysis,” *Information (Switzerland)*, vol. 15, no. 2, Feb. 2024, doi: 10.3390/info15020083.
- [11] Y. Yuliani, “Algoritma Random Forest Untuk Prediksi Kelangsungan Hidup Pasien Gagal Jantung Menggunakan Seleksi Fitur Bestfirst,” *Infotek: Jurnal Informatika dan Teknologi*, vol. 5, no. 2, pp. 298–306, Jul. 2022, doi: 10.29408/jit.v5i2.5896.
- [12] B. H. Garin Jhovany, “BATU EMPEDU,” *Proceeding of The 18th Continuing Medical Education*, pp. 540–550, Mar. 2025, Accessed: Aug. 04, 2026. [Online]. Available: <https://proceedings.ums.ac.id/kedokteran/article/view/6294/5603>
- [13] Ary Prandika Siregar, Dwi Priyadi Purba, Jojo Putri Pasaribu, and Khairul Reza Bakara, “Implementasi Algoritma Random Forest Dalam Klasifikasi Diagnosis Penyakit Stroke,” *Jurnal Penelitian Rumpun Ilmu Teknik*, vol. 2, no. 4, pp. 155–164, Nov. 2023, doi: 10.55606/juprit.v2i4.3039.
- [14] A. G. Zulus Saputra and Fikri Budiman, “Analisis Komparatif Model Random Forest dan XGBoost untuk Klasifikasi Penyakit Jantung Berbasis Data Klinis,” *JURIKOM (Jurnal Riset Komputer)*, vol. 13, no. 1, pp. 254–264, Feb. 2026, doi: 10.30865/jurikom.v13i1.9427.
- [15] F. Widyawati, H. P. Suhito, W. Yassin, and H. Agus Santoso, “Classification Of Toddler Nutritional Status Using Support Vector Machine And Random Forest Techniques With Optimal Feature Selection,” *Jurnal Teknik Informatika (Jutif)*, vol. 5, no. 6, pp. 1893–1904, Jan. 2025, doi: 10.52436/1.jutif.2024.5.6.4162.
- [16] R. S. Wijaya, A. Qur’ania, and I. Anggraeni, “Klasifikasi Penyakit Cacar Monyet Menggunakan Support Vector Machine (SVM),” *MALCOM: Indonesian Journal of Machine Learning and Computer Science*, vol. 4, no. 4, pp. 1253–1260, Jul. 2024, doi: 10.57152/malcom.v4i4.1417.
- [17] D. Muallifah, W. Fadila, and R. Firdaus, “Teknik SMOTE untuk Mengatasi Imbalance Data pada Deteksi Penyakit Stroke Menggunakan Algoritma Random Forest,” *Jurnal CoSciTech (Computer Science and Information Technology)*, vol. 3, no. 2, pp. 107–113, Aug. 2022, doi: 10.37859/coscitech.v3i2.3912.
- [18] H. A. Nabila and Endang Wahyu Pamungkas, “Perbandingan Algoritma Machine Learning: Svm, Random Forest, Dan Xgboost Untuk Prediksi Stroke,” *Rabit : Jurnal Teknologi dan Sistem Informasi Univrab*, vol. 10, no. 2, pp. 1098–1110, Jul. 2025, doi: 10.36341/rabit.v10i2.6444.
- [19] V. W. Mada and D. Damayanti, “Perbandingan Algoritma Random Forest, XGBoost dan SVM Pada Klasifikasi Penyakit Demam Berdarah Dengue (DBD),” *Building of Informatics, Technology and Science (BITS)*, vol. 7, no. 3, pp. 2157–2167, Dec. 2025, doi: 10.47065/bits.v7i3.8591.
- [20] O. Opitarsi, F. Natsir, and E. S. Marsiani, “Klasifikasi Diagnosis untuk Penyakit Kanker Serviks Menggunakan Algoritma Extreme Gradient Boosting (XGBoost),” *Jurnal Ilmiah FIFO*, vol. 16, no. 1, p. 55, May 2024, doi: 10.22441/fifo.2024.v16i1.006.
- [21] P. Septiana Rizky, R. Haiban Hirzi, U. Hidayaturrohmah, U. Hamzanwadi Selong Jl TGKH Muhammad Zainuddin Abdul Madjid Pancor, and L. Timur, “Perbandingan Metode LightGBM dan XGBoost dalam Menangani Data dengan Kelas Tidak Seimbang,” *J Statistika*, vol. 15, no. 2, pp. 228–236, Dec. 2022, doi: <https://doi.org/10.36456/jstat.vol15.no2.a5548>.
- [22] R. Harahap, M. Irpan, M. Azzuhri Dinata, and L. Efrizoni, “Perbandingan Algoritma Random Forest Dan Xgboost Untuk Klasifikasi Penyakit Paru-Paru Berdasarkan Data Demografi Pasien,” *Jurnal Ilmiah Betrik*, vol. 15, no. 02, pp. 130–141, Aug. 2024, doi: <https://doi.org/10.36050/3v3xwn06>.
- [23] J. P. Anggraini, Chaya Gladys Zhafirah A, and A. Desiani, “Perbandingan Algoritma Random Forest dan Extreme Gradient Boosting (XGBoost) dalam Klasifikasi Penyakit Gagal Jantung,” *Komputika : Jurnal Sistem Komputer*, vol. 14, no. 2, pp. 149–157, Nov. 2025, doi: 10.34010/komputika.v14i2.16618.