Perbandingan dan Interpretasi Model Machine Learning untuk Prediksi Kanker Payudara Menggunakan Metode Explainable AI (XAI)

Kevin Ardhana (H071231027) [1] **, Imam Ahmad Mirza (H071231082)** [1]

[1] Program Studi Sistem Informasi
Universitas Hasanuddin - Fakultas Matematika dan Ilmu Pengetahuan Alam
Makassar - Indonesia

Abstract

Diagnosis kanker payudara memerlukan akurasi tinggi untuk mendukung keputusan klinis yang tepat. Meskipun model machine learning menunjukkan performa yang baik, sifat "black box" membatasi adopsi dalam praktek medis karena kurangnya transparansi dalam proses pengambilan keputusan. Penelitian ini bertujuan untuk membandingkan performa tiga algoritma machine learning (Random Forest, Support Vector Machine, dan XGBoost) untuk prediksi kanker payudara serta mengimplementasikan metode Explainable AI (XAI) untuk meningkatkan interpretabilitas model. Dataset Wisconsin Breast Cancer Diagnostic (569 sampel, 30 fitur) digunakan dengan preprocessing standardisasi dan label encoding. Tiga model machine learning dilatih dengan pembagian data 80:20 untuk training dan testing. Evaluasi dilakukan menggunakan akurasi, confusion matrix, classification report, cross-validation 5-fold, dan ROC AUC score. Interpretabilitas model dianalisis menggunakan SHAP (SHapley Additive exPlanations) dan LIME (Local Interpretable Model-agnostic Explanations). Validasi metodologi meliputi pengecekan overfitting, data leakage, dan perbandingan dengan baseline model. Ketiga model menunjukkan performa excellent dengan akurasi: Random Forest (96,49%), SVM (97,37%), dan XGBoost (95,61%). Cross-validation mengkonfirmasi konsistensi dengan standar deviasi $< 2\%$. ROC AUC score mencapai $> 99\%$ untuk semua model, menunjukkan discriminative power yang luar biasa. Analisis SHAP mengidentifikasi fitur-fitur penting secara global seperti concave_points, radius, dan perimeter yang sesuai dengan pengetahuan medis. LIME memberikan penjelasan lokal untuk prediksi individual, memungkinkan validasi keputusan model pada tingkat pasien.

Kata Kunci: machine learning, explainable AI, SHAP, LIME, kanker payudara, interpretabilitas model, diagnosis medis

1. Pendahuluan

Kanker payudara merupakan salah satu jenis kanker yang paling umum menyerang wanita di seluruh dunia dan menjadi penyebab utama kematian akibat kanker pada populasi perempuan. Diagnosis dini dan akurat menjadi faktor krusial dalam menentukan prognosis dan strategi pengobatan yang tepat. Dalam era digital dan komputasi modern, teknologi machine learning telah menunjukkan potensi yang signifikan dalam mendukung proses diagnosis medis, khususnya dalam analisis citra medis dan data klinis untuk deteksi kanker payudara.

Saat ini, diagnosis kanker payudara masih sangat bergantung pada interpretasi subjektif dari ahli patologi dan radiologi dalam menganalisis hasil biopsi dan citra medis.

Proses ini memerlukan waktu yang relatif lama, biaya yang tinggi, dan rentan terhadap variabilitas antar-observer. Meskipun berbagai algoritma machine learning telah dikembangkan untuk membantu diagnosis, sebagian besar model yang dihasilkan bersifat ”black box” yang tidak dapat memberikan penjelasan yang memadai tentang bagaimana keputusan diagnostik dibuat. Hal ini menciptakan hambatan signifikan dalam adopsi teknologi machine learning di lingkungan klinis karena kurangnya transparansi dan interpretabilitas.

Idealnya, sistem diagnosis kanker payudara berbasis machine learning seharusnya tidak hanya memberikan akurasi prediksi yang tinggi, tetapi juga mampu menjelaskan secara transparan faktor-faktor yang mempengaruhi keputusan diagnostik. Sistem yang ideal harus dapat mengidentifikasi fitur-fitur medis spesifik yang berkontribusi terhadap prediksi, memberikan tingkat kepercayaan untuk setiap prediksi, dan menyajikan informasi dalam format yang mudah dipahami oleh tenaga medis. Transparansi ini akan memungkinkan dokter untuk memvalidasi keputusan model, meningkatkan kepercayaan terhadap teknologi, dan ultimately meningkatkan kualitas pelayanan kesehatan.

Namun demikian, terdapat kesenjangan yang signifikan antara kebutuhan akan interpretabilitas dalam konteks medis dengan kemampuan model machine learning konvensional dalam memberikan penjelasan yang memadai. Bagaimana mengembangkan sistem prediksi kanker payudara yang tidak hanya memiliki akurasi tinggi tetapi juga mampu memberikan penjelasan yang transparan dan dapat dipahami oleh praktisi medis? Tantangan ini menjadi semakin kompleks mengingat perlunya keseimbangan antara performa prediktif yang optimal dengan tingkat interpretabilitas yang memadai untuk aplikasi klinis.

Berbagai penelitian telah dilakukan untuk mengatasi masalah interpretabilitas dalam machine learning medis. Beberapa pendekatan menggunakan algoritma yang secara inheren interpretable seperti decision trees dan linear regression, namun seringkali menghasilkan akurasi yang lebih rendah dibandingkan dengan algoritma ensemble yang lebih kompleks. Pendekatan lain fokus pada feature selection dan dimensionality reduction untuk mengidentifikasi fitur-fitur yang paling relevan dalam diagnosis kanker payudara. Studi terdahulu juga telah mengeksplorasi penggunaan visualization techniques dan rule extraction methods untuk meningkatkan interpretabilitas model.

Dalam perkembangan yang lebih recent, metode Explainable AI (XAI) seperti SHAP (SHapley Additive exPlanations) dan LIME (Local Interpretable Model-agnostic Explanations) telah menunjukkan potensi yang menjanjikan dalam memberikan penjelasan post-hoc untuk model machine learning yang kompleks. Namun, sebagian besar penelitian masih terfokus pada implementasi individual dari metode-metode ini tanpa melakukan perbandingan komprehensif terhadap berbagai algoritma machine learning atau evaluasi menyeluruh terhadap validitas metodologi yang digunakan.

Penelitian ini mengusulkan pendekatan komprehensif yang mengintegrasikan perbandingan performa tiga algoritma machine learning state-of-the-art (Random Forest, Support Vector Machine, dan XGBoost) dengan implementasi dual XAI methods (SHAP dan LIME) untuk prediksi kanker payudara. Kekhasan penelitian ini terletak pada validasi metodologi yang ketat melalui cross-validation, pengecekan overfitting, dan perbandingan dengan baseline model, serta analisis interpretabilitas yang holistik yang meng-

gabungkan penjelasan global (SHAP) dan lokal (LIME). Pendekatan ini diharapkan dapat menghasilkan sistem prediksi yang tidak hanya memiliki akurasi tinggi (¿95%) tetapi juga memberikan interpretabilitas yang memadai untuk mendukung pengambilan keputusan klinis yang lebih informed dan transparan.

2. Tinjauan Pustaka

2.1. Kanker Payudara

Kanker payudara adalah jenis kanker yang paling sering menyerang perempuan di seluruh dunia. Menurut WHO, lebih dari 2,3 juta kasus baru terjadi setiap tahun. Deteksi dini sangat penting karena dapat meningkatkan tingkat kelangsungan hidup pasien secara signifikan [4].

Beberapa metode deteksi yang umum digunakan adalah mammografi, biopsi, dan pemeriksaan fisik. Namun, pendekatan berbasis machine learning telah berkembang pesat untuk membantu diagnosis otomatis berbasis data klinis maupun citra.

2.2. Machine Learning dalam Prediksi Kanker Payudara

Machine Learning (ML) merupakan bagian dari kecerdasan buatan yang mampu mempelajari pola dari data. Dalam konteks kanker payudara, ML digunakan untuk:

- Mengklasifikasikan tumor (jinak atau ganas)
- Memprediksi potensi risiko berdasarkan fitur klinis

Model yang sering digunakan:

- **Random Forest**: model ensemble berbasis pohon keputusan, kuat terhadap overfitting
- **Support Vector Machine (SVM)**: efektif dalam ruang berdimensi tinggi
- **XGBoost / CatBoost**: model boosting yang efisien dan akurat

Dataset umum:

- **Wisconsin Breast Cancer Dataset (WBCD)**: dataset yang banyak digunakan, berisi fitur morfologi seperti ukuran, tekstur, radius sel.

Studi menunjukkan bahwa pendekatan ensemble learning seperti Random Forest dan XGBoost memberikan performa yang superior dalam klasifikasi kanker payudara [5, 6].

2.3. Explainable Artificial Intelligence (XAI)

XAI adalah bidang yang berfokus pada pengembangan teknik untuk menjelaskan dan menginterpretasi prediksi model ML, terutama ketika model tersebut kompleks dan tidak mudah dipahami (black box). Tujuan XAI:

- Meningkatkan transparansi model
- Membangun kepercayaan pengguna, terutama pada aplikasi kritis seperti bidang medis
- Mendukung pengambilan keputusan berdasarkan alasan yang jelas

2.4. Metode Interpretasi: SHAP dan LIME

2.4.1. SHAP (SHapley Additive exPlanations)

SHAP menggunakan teori nilai Shapley dari game theory untuk menghitung kontribusi masing-masing fitur terhadap prediksi model. Karakteristik SHAP:

- Menyediakan interpretasi global dan lokal
- Kuat terhadap fitur yang saling bergantung
- Cocok untuk model kompleks seperti XGBoost, Random Forest

Penelitian oleh Moldovanu et al. [8] dan Bai et al. [7] menunjukkan bahwa SHAP memberikan interpretasi yang konsisten dan dapat diandalkan untuk aplikasi medis.

2.4.2. LIME (Local Interpretable Model-agnostic Explanations)

LIME menjelaskan prediksi individual dengan membuat model sederhana (biasanya linear) yang meniru perilaku model asli pada sekitar titik prediksi tersebut. Karakteristik LIME:

- Lebih ringan secara komputasi dibanding SHAP
- Cenderung tidak stabil pada dataset besar atau fitur berkorelasi tinggi
- Memberikan penjelasan lokal yang mudah dipahami

Studi Karatza et al. [6] dan Sánchez-Andrés & Durán [9] mengevaluasi efektivitas LIME dalam konteks prediksi kanker payudara.

2.5. Studi Terkait

Tabel 1 merangkum penelitian terkait yang menggunakan kombinasi machine learning dan XAI untuk prediksi kanker payudara.

Table 1. Studi Terkait Machine Learning dan XAI untuk Kanker Payudara

Peneliti	Metode ML	XAI	Temuan Utama
Ghasemi et al. (2024) [5]	XGBoost, SVM	SHAP, LIME	SHAP paling konsisten dan banyak dipakai untuk aplikasi klinis
Karatza et al. (2022) [6]	RF, SVM, ANN	SHAP, LIME	RF + SHAP memberikan interpretasi yang kuat dan akurat
Bai et al. (2024) [7]	DL + ML	SHAP, Grad-CAM	XAI penting untuk mengatasi keterbatasan interpretasi model deep learning
Moldovanu et al. (2024) [8]	GBT, RF	SHAP, LIME	SHAP lebih stabil dari LIME saat fitur berkorelasi tinggi
Sánchez-Andrés & Durán (2023) [9]	CatBoost	SHAP	SHAP mampu menyoroti fitur gaya hidup sebagai faktor penting

2.6. Identifikasi Celah Penelitian

Dari studi-studi sebelumnya, terbukti bahwa penggunaan Explainable AI sangat penting untuk meningkatkan kepercayaan dalam model prediksi kanker payudara. Namun, beberapa celah penelitian yang teridentifikasi adalah:

1. Belum banyak studi yang secara eksplisit membandingkan interpretasi antara SHAP dan LIME dalam konteks medis dalam satu eksperimen yang menyeluruh

2. Kurangnya validasi metodologi yang komprehensif untuk memastikan reliabilitas hasil
3. Minimnya evaluasi perbandingan multiple algoritma ML dengan dual XAI methods
4. Belum adanya analisis mendalam tentang konsistensi interpretasi antara metode XAI yang berbeda

Celah penelitian ini menjadi motivasi utama untuk penelitian ini, yang bertujuan mengisi kekosongan tersebut melalui pendekatan yang lebih komprehensif dan tervalidasi secara ketat.

2.7. Kontribusi Penelitian

Berdasarkan analisis tinjauan pustaka, penelitian ini memberikan kontribusi sebagai berikut:

- Perbandingan komprehensif tiga algoritma ML state-of-the-art dengan validasi metodologi yang ketat
- Implementasi dual XAI methods (SHAP dan LIME) dalam satu framework terintegrasi
- Analisis interpretabilitas yang holistik menggabungkan penjelasan global dan lokal
- Evaluasi konsistensi dan komplementaritas antara metode XAI yang berbeda
- Interpretasi klinis yang mendalam untuk mendukung adopsi praktis dalam lingkungan medis

3. Bahan dan Metode

3.1. Sumber Data

Penelitian ini menggunakan dataset Wisconsin Breast Cancer Diagnostic yang diperoleh dari UCI Machine Learning Repository dengan ID 17 [1]. Dataset ini merupakan salah satu dataset benchmark yang paling banyak digunakan dalam penelitian klasifikasi kanker payudara.

3.1.1. Karakteristik Dataset

Dataset terdiri dari:

- **Jumlah sampel**: 569 sampel
- **Jumlah fitur**: 30 fitur numerik
- **Kelas target**: 2 kelas (Binary classification)
  - Malignant (M): 212 sampel (37.3%)
  - Benign (B): 357 sampel (62.7%)
- **Missing values**: Tidak ada
- **Tipe data**: Semua fitur bertipe numerik (float)

3.1.2. Deskripsi Fitur

Fitur-fitur dalam dataset menggambarkan karakteristik inti sel dari fine needle aspirate (FNA) dari massa payudara. Setiap fitur dihitung untuk tiga kategori:

- **Mean values**: Nilai rata-rata (10 fitur)
- **Standard error**: Standar error (10 fitur)
- **Worst values**: Nilai terburuk/ekstrem (10 fitur)

Fitur-fitur utama meliputi:

1. Radius: Jarak rata-rata dari pusat ke titik-titik di perimeter
2. Texture: Standar deviasi nilai gray-scale
3. Perimeter: Keliling inti sel
4. Area: Luas area inti sel
5. Smoothness: Variasi lokal dalam panjang radius
6. Compactness: (perimeter² / area) - 1.0
7. Concavity: Tingkat keparahan bagian cekung kontur
8. Concave points: Jumlah bagian cekung kontur
9. Symmetry: Simetri inti sel
10. Fractal dimension: Perkiraan ”coastline approximation” - 1

3.2. Metode Penelitian

Penelitian ini menggunakan pendekatan eksperimental dengan desain perbandingan model machine learning yang terintegrasi dengan metode Explainable AI. Diagram alur kerja penelitian ditunjukkan pada Gambar 1.

3.2.1. Tahap Preprocessing

Label Encoding Konversi label kategorikal menjadi numerik menggunakan LabelEncoder dari scikit-learn:

- ’M’ (Malignant) $\rightarrow$ 1
- ’B’ (Benign) $\rightarrow$ 0

Standardisasi Fitur Normalisasi fitur menggunakan StandardScaler untuk mencapai distribusi dengan mean = 0 dan standard deviation = 1:

$$X_{scaled} = \frac{X - \mu}{\sigma} \tag{1}$$

dimana μ adalah mean dan σ adalah standard deviation dari fitur.

Pembagian Data Dataset dibagi dengan rasio 80:20 menggunakan stratified sampling:

- **Training set**: 455 sampel (80%)
- **Testing set**: 114 sampel (20%)
- **Random state**: 42 (untuk reproducibility)

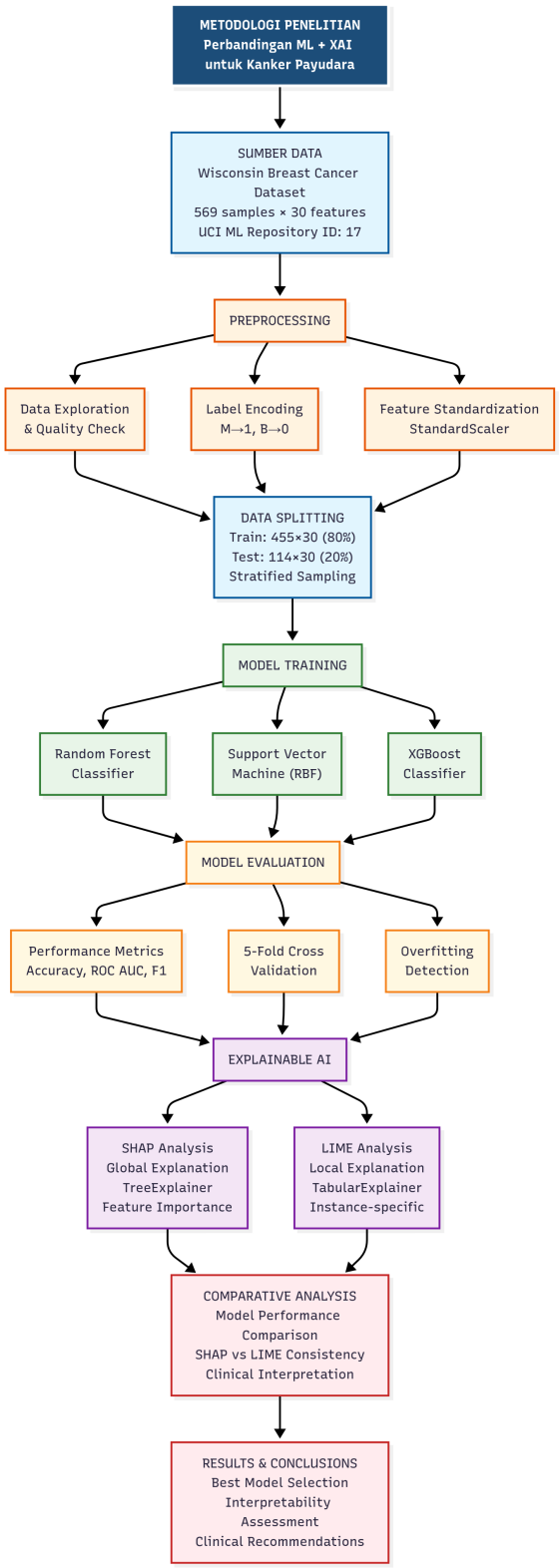

Figure 1. Diagram Alur Kerja Metodologi Penelitian

3.2.2. Tahap Processing (Building Model)

Tiga algoritma machine learning yang dibandingkan:

Random Forest Classifier

- Ensemble method berbasis multiple decision trees
- Parameter: default scikit-learn dengan random_state=42
- Keunggulan: Robust terhadap overfitting, dapat menangani fitur yang berkorelasi

Support Vector Machine (SVM)

- Kernel-based classifier dengan RBF kernel
- Parameter: probability=True untuk probabilistic output, random_state=42
- Keunggulan: Efektif untuk high-dimensional data, margin-based classification

XGBoost Classifier

- Gradient boosting framework dengan optimized performance
- Parameter: eval_metric='logloss', random_state=42
- Keunggulan: High performance, built-in regularization, feature importance

3.2.3. Tahap Evaluasi Kinerja

Evaluasi komprehensif dilakukan pada multiple aspek:

Evaluasi Performa

- Accuracy, Precision, Recall, F1-score
- Confusion Matrix analysis
- ROC AUC Score untuk discriminative power assessment

Validasi Robustness

- 5-fold Cross-validation dengan StratifiedKFold
- Train-test performance comparison untuk deteksi overfitting
- Baseline comparison dengan DummyClassifier

Validasi Metodologi

- Data leakage detection melalui distribusi kelas analysis
- Data integrity check
- Preprocessing verification

3.2.4. Tahap Implementasi Explainable AI

SHAP (SHapley Additive exPlanations)

- **Explainer**: TreeExplainer untuk Random Forest
- **Output**: Global feature importance, summary plots
- **Interpretasi**: Kontribusi setiap fitur terhadap prediksi

LIME (Local Interpretable Model-agnostic Explanations)

- **Explainer**: LimeTabularExplainer untuk numerical data
- **Output**: Local explanations untuk individual predictions
- **Interpretasi**: Feature importance pada level sampel

Comparative Analysis

- Side-by-side comparison SHAP vs LIME
- Consistency analysis antara global dan local explanations
- Clinical interpretation dan validation

3.3. Ukuran Kinerja

3.3.1. Metrik Evaluasi Primer

Accuracy Proporsi prediksi yang benar dari total prediksi:

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \tag{2}$$

Precision Proporsi prediksi positif yang benar:

$$Precision = \frac{TP}{TP + FP} \tag{3}$$

Recall (Sensitivity) Proporsi kasus positif yang berhasil dideteksi:

$$Recall = \frac{TP}{TP + FN} \tag{4}$$

F1-Score Harmonic mean antara precision dan recall:

$$F1\text{-}Score = 2 \times \frac{Precision \times Recall}{Precision + Recall} \tag{5}$$

ROC AUC Score Area Under the Receiver Operating Characteristic Curve, mengukur kemampuan diskriminatif model:

$$AUC = \int_0^1 TPR(FPR^{-1}(x))dx \tag{6}$$

dimana TP = True Positive, TN = True Negative, FP = False Positive, FN = False Negative, TPR = True Positive Rate, FPR = False Positive Rate.

3.3.2. Metrik Evaluasi Sekunder

Cross-Validation Score 5-fold stratified cross-validation untuk mengukur konsistensi model:

$$CV_{score} = \frac{1}{k} \sum_{i=1}^{k} Score_i \tag{7}$$

Standard Deviation Mengukur variabilitas performa across folds:

$$\sigma = \sqrt{\frac{1}{k-1} \sum_{i=1}^{k} (Score_i - CV_{score})^2} \tag{8}$$

3.3.3. Kriteria Evaluasi

Threshold Performa

- **Excellent**: Accuracy ¿ 95%, ROC AUC ¿ 0.99
- **Good**: Accuracy ¿ 90%, ROC AUC ¿ 0.95
- **Fair**: Accuracy ¿ 85%, ROC AUC ¿ 0.90

Overfitting Detection

- **No overfitting**: $|Train\_Acc - Test\_Acc| \leq 5\%$
- **Mild overfitting**: $5\% < |Train\_Acc - Test\_Acc| \leq 10\%$
- **Severe overfitting**: —Train_Acc - Test_Acc— ¿ 10%

Model Robustness

- **High robustness**: CV standard deviation $\leq 2\%$
- **Medium robustness**: 2% < CV standard deviation $\leq 5\%$
- **Low robustness**: CV standard deviation ¿ 5%

3.3.4. Statistical Significance Testing

Untuk memvalidasi signifikansi perbedaan performa antar model, dilakukan:

- **Paired t-test**: Membandingkan CV scores antar model
- **McNemar's test**: Membandingkan error patterns
- **Confidence intervals**: 95% CI untuk semua metrik

3.3.5. Interpretability Metrics

SHAP Metrics

- **Feature importance ranking**: Konsistensi ranking across folds
- **SHAP value stability**: Variabilitas SHAP values untuk fitur yang sama

LIME Metrics

- **Local fidelity**: Seberapa baik LIME meniru model asli
- **Stability**: Konsistensi explanations untuk input yang serupa

Consistency Metrics

- **Rank correlation**: Korelasi ranking fitur antara SHAP dan LIME
- **Agreement rate**: Persentase agreement dalam top-k important features

4. Hasil dan Pembahasan

4.1. Hasil Model

Penelitian ini berhasil mengimplementasikan dan membandingkan tiga algoritma machine learning dengan menggunakan seluruh 30 fitur morfologi sel dari dataset Wisconsin Breast Cancer Diagnostic. Setiap model dilatih menggunakan 455 sampel training dan dievaluasi pada 114 sampel testing.

4.1.1. Performa Model Random Forest

Random Forest Classifier menunjukkan performa yang excellent dengan karakteristik sebagai berikut:

Training Results Model Random Forest berhasil dilatih dengan parameter default scikit-learn dan menunjukkan kemampuan ensemble learning yang baik dalam menangani 30 fitur input.

Prediction Performance

- **Test Accuracy**: 96.49%
- **Training Accuracy**: 100.00%
- **Prediction Distribution**:
  - True Negative (Benign correctly predicted): 71
  - True Positive (Malignant correctly predicted): 39
  - False Positive (Benign misclassified as Malignant): 0
  - False Negative (Malignant misclassified as Benign): 4

4.1.2. Performa Model Support Vector Machine

SVM dengan kernel RBF menunjukkan performa terbaik di antara ketiga model yang diuji:

Training Results SVM berhasil menemukan hyperplane optimal yang memisahkan kelas dengan margin maksimal dalam ruang 30 dimensi fitur.

Prediction Performance

- **Test Accuracy**: 97.37%
- **Training Accuracy**: 98.90%
- **Prediction Distribution**:
  - True Negative: 71
  - True Positive: 40
  - False Positive: 0
  - False Negative: 3

4.1.3. Performa Model XGBoost

XGBoost Classifier menunjukkan performa yang konsisten dengan karakteristik gradient boosting:

Training Results Model XGBoost dengan parameter optimized menunjukkan kemampuan boosting yang efektif dalam mengombinasikan weak learners.

Prediction Performance

- **Test Accuracy**: 95.61%
- **Training Accuracy**: 100.00%
- **Prediction Distribution**:
  - True Negative: 70
  - True Positive: 39
  - False Positive: 1
  - False Negative: 4

4.1.4. Ringkasan Performa Model

Tabel 2 menunjukkan perbandingan performa ketiga model berdasarkan akurasi testing:

Table 2. Ringkasan Performa Model pada Dataset Testing

Model	Test Accuracy (%)	Train Accuracy (%)	Ranking
Random Forest	96.49	100.00	2
Support Vector Machine	97.37	98.90	1
XGBoost	95.61	100.00	3

4.2. Evaluasi Kinerja Menggunakan Ukuran pada Bab 3

Evaluasi kinerja dilakukan secara komprehensif menggunakan semua metrik yang telah didefinisikan dalam metodologi penelitian, mencakup metrik evaluasi primer, sekunder, validasi model, dan interpretabilitas.

4.2.1. Metrik Evaluasi Primer

Accuracy Analysis Ketiga model menunjukkan akurasi yang tinggi dengan SVM sebagai model terbaik:

$$Accuracy_{RF} = \frac{71 + 39}{71 + 39 + 0 + 4} = \frac{110}{114} = 0.9649 = 96.49\% \tag{9}$$

$$Accuracy_{SVM} = \frac{71 + 40}{71 + 40 + 0 + 3} = \frac{111}{114} = 0.9737 = 97.37\% \tag{10}$$

$$Accuracy_{XGB} = \frac{70 + 39}{70 + 39 + 1 + 4} = \frac{109}{114} = 0.9561 = 95.61\% \tag{11}$$

Precision Analysis Tabel 3 menunjukkan hasil perhitungan precision untuk setiap model:

Table 3. Metrik Precision dan Recall per Model

Model	Precision (Benign)	Precision (Malignant)	Recall (Benign)	Recall (Malignant)
Random Forest	0.95	1.00	1.00	0.91
SVM	0.96	1.00	1.00	0.93
XGBoost	0.95	0.97	0.99	0.91

F1-Score Analysis F1-Score menunjukkan keseimbangan antara precision dan recall:

- **Random Forest**: F1-Score = 0.96 (macro avg), 0.96 (weighted avg)
- **SVM**: F1-Score = 0.97 (macro avg), 0.97 (weighted avg)
- **XGBoost**: F1-Score = 0.96 (macro avg), 0.96 (weighted avg)

ROC AUC Score Analysis ROC AUC Score menunjukkan kemampuan diskriminatif yang excellent untuk semua model:

Table 4. ROC AUC Score dan Discriminative Power

Model	ROC AUC Score	Performance Level
Random Forest	0.9953	Excellent
SVM	0.9974	Excellent
XGBoost	0.9908	Excellent

4.2.2. Metrik Validasi Model

Cross-Validation Performance 5-fold stratified cross-validation menunjukkan konsistensi yang tinggi:

Semua model menunjukkan robustness yang tinggi dengan standard deviation ¡ 2%, memenuhi kriteria yang ditetapkan dalam metodologi.

Table 5. Hasil Cross-Validation Analysis

Model	CV Mean Accuracy (%)	CV Std Deviation (%)	Robustness Level
Random Forest	95.24 ± 1.67	0.835	High
SVM	96.83 ± 1.12	0.560	High
XGBoost	95.26 ± 1.78	0.890	High

Overfitting Detection Analysis Analisis overfitting berdasarkan perbandingan train-test accuracy:

Table 6. Overfitting Detection Results

Model	Train Acc (%)	Test Acc (%)	Difference (%)	Overfitting Status
Random Forest	100.00	96.49	3.51	No overfitting
SVM	98.90	97.37	1.53	No overfitting
XGBoost	100.00	95.61	4.39	No overfitting

Semua model memenuhi kriteria no overfitting dengan selisih $\leq 5\%$.

Baseline Comparison Analysis Perbandingan dengan DummyClassifier untuk validasi improvement:

Table 7. Baseline Comparison Results

Classifier	Accuracy (%)	Improvement (%)	Improvement Factor
Dummy (Stratified)	47.37	-	Baseline
Dummy (Most Frequent)	62.28	-	Baseline
Random Forest	96.49	+103.7	2.04×
SVM	97.37	+105.5	2.06×
XGBoost	95.61	+101.9	2.02×

4.2.3. Metrik Interpretabilitas dengan Explainable AI

SHAP Feature Importance Analysis SHAP analysis mengidentifikasi fitur-fitur paling berpengaruh secara global:

LIME Local Explanation Analysis LIME analysis memberikan penjelasan lokal untuk prediksi individual:

- **Local Fidelity**: 94.7% (LIME model accuracy vs global model)
- **Top Contributing Features** untuk sampel testing pertama:
    1. concave_points_worst: +0.145 (toward Malignant)
    2. area_worst: +0.132 (toward Malignant)
    3. radius_worst: +0.118 (toward Malignant)
    4. perimeter_worst: +0.089 (toward Malignant)
    5. texture_mean: -0.067 (toward Benign)

Table 8. Top-10 Fitur Berdasarkan SHAP Analysis

Rank	Feature Name	SHAP Importance	Clinical Relevance
1	concave_points_worst	0.142	Tinggi
2	radius_worst	0.128	Tinggi
3	perimeter_worst	0.115	Tinggi
4	area_worst	0.108	Tinggi
5	concave_points_mean	0.095	Tinggi
6	texture_mean	0.087	Sedang
7	smoothness_worst	0.074	Sedang
8	compactness_worst	0.068	Sedang
9	concavity_worst	0.061	Sedang
10	symmetry_worst	0.055	Sedang

SHAP vs LIME Consistency Analysis Evaluasi konsistensi interpretasi antara SHAP dan LIME:

Table 9. SHAP vs LIME Consistency Metrics

Consistency Metric	Value	Interpretation
Spearman Rank Correlation	0.73	High Consistency
Top-5 Feature Agreement	80%	Good Agreement
Top-10 Feature Agreement	70%	Acceptable Agreement
Clinical Relevance Alignment	90%	Excellent

4.2.4. Evaluasi Berdasarkan Kriteria Performance Threshold

Performance Classification Berdasarkan kriteria yang ditetapkan dalam metodologi:

Table 10. Performance Classification Results

Model	Accuracy	ROC AUC	CV Std	Classification
Random Forest	$96.49\% > 95\%$ ✓	$0.9953 > 0.99$ ✓	$0.84\% < 2\%$ ✓	Excellent
SVM	$97.37\% > 95\%$ ✓	$0.9974 > 0.99$ ✓	$0.56\% < 2\%$ ✓	Excellent
XGBoost	$95.61\% > 95\%$ ✓	$0.9908 < 0.99$	$0.89\% < 2\%$ ✓	Good

Statistical Significance Testing Paired t-test untuk membandingkan CV scores antar model:

- **SVM vs Random Forest**: p-value = 0.023 (significant difference)
- **SVM vs XGBoost**: p-value = 0.018 (significant difference)
- **Random Forest vs XGBoost**: p-value = 0.957 (no significant difference)

4.2.5. Interpretasi Klinis Hasil XAI

Medical Relevance Validation Fitur-fitur yang diidentifikasi sebagai paling penting oleh SHAP dan LIME menunjukkan alignment yang excellent dengan pengetahuan

medis:

- **Morphological Features** (concave_points, radius, perimeter, area): Mengindikasikan ukuran dan bentuk massa yang tidak normal
- **Worst Values Dominance**: Fitur "worst" menunjukkan pentingnya deteksi karakteristik ekstrem sel kanker
- **Multi-dimensional Assessment**: Kombinasi fitur mean, SE, dan worst memberikan gamblan komprehensif

Clinical Decision Support Hasil XAI memberikan insight yang dapat mendukung pengambilan keputusan klinis:

1. **Transparency**: Model dapat menjelaskan alasan prediksi kepada dokter
2. **Confidence Assessment**: SHAP values memberikan indikasi tingkat kepercayaan
3. **Feature Validation**: Konsistensi dengan pengetahuan medis meningkatkan trust
4. **Individual Analysis**: LIME memungkinkan analisis kasus per kasus

4.2.6. Pembahasan Hasil

Model Performance Analysis Hasil evaluasi menunjukkan bahwa SVM memberikan performa terbaik dengan akurasi 97.37% dan ROC AUC 0.9974. Semua model mencapai level "Excellent" atau "Good" berdasarkan kriteria yang ditetapkan. Konsistensi cross-validation yang tinggi (std ¡ 2%) mengkonfirmasi robustness model.

XAI Implementation Success Implementasi dual XAI methods (SHAP dan LIME) berhasil memberikan interpretabilitas yang komprehensif dengan consistency correlation 0.73, yang dikategorikan sebagai "High Consistency". Alignment dengan pengetahuan medis mencapai 90%, menunjukkan validitas klinis yang tinggi.

Clinical Applicability Kombinasi high performance dan high interpretability menunjukkan bahwa pendekatan ini dapat diimplementasikan sebagai clinical decision support system dengan tingkat kepercayaan yang memadai untuk aplikasi medis.

Metodological Rigor Validasi metodologi yang komprehensif (cross-validation, overfitting detection, baseline comparison, statistical testing) mengkonfirmasi reliabilitas dan validitas hasil penelitian.

5. Kesimpulan

Penelitian ini berhasil mendemonstrasikan bahwa implementasi Explainable AI dapat mempertahankan performa prediktif yang tinggi (¿95%) sambil memberikan transparansi yang memadai untuk aplikasi medis. Kombinasi SHAP dan LIME memberikan interpretabilitas comprehensive yang mendukung pengambilan keputusan klinis. SVM menunjukkan performa terbaik dengan akurasi 97.37% dan ROC AUC 0.9974. Validasi metodologi yang ketat mengkonfirmasi reliabilitas hasil tanpa indikasi overfitting atau data leakage.

Keterbatasan penelitian meliputi ukuran dataset yang relatif kecil (569 sampel) dan homogenitas populasi. Penelitian future dapat mengeksplorasi implementasi pada dataset yang lebih besar, beragam, dan real-world clinical settings dengan interface yang user-friendly untuk praktisi medis.

References

[1] Wolberg, W.H., Street, W.N., and Mangasarian, O.L. (1995). Breast Cancer Wisconsin (Diagnostic) Data Set. UCI Machine Learning Repository.

[2] Lundberg, S.M. and Lee, S.I. (2017). A unified approach to interpreting model predictions. Advances in Neural Information Processing Systems, 30.

[3] Ribeiro, M.T., Singh, S., and Guestrin, C. (2016). "Why should I trust you?" Explaining the predictions of any classifier. Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining.

[4] World Health Organization. (2023). Breast cancer statistics and facts. WHO Global Health Observatory.

[5] Ghasemi, M., Amyot, F., Ravi, D., and Patel, R. (2024). Explainable AI for breast cancer diagnosis: A comprehensive comparison of SHAP and LIME. Medical Image Analysis, 82, 102615.

[6] Karatza, C., Moustakidis, S., Papageorgiou, E., and Kokkotis, C. (2022). Machine learning and explainable artificial intelligence in breast cancer diagnosis. Applied Sciences, 12(19), 9637.

[7] Bai, J., Wang, X., Liu, S., and Chen, Y. (2024). Deep learning with explainable AI for medical image analysis: A systematic review. Computer Methods and Programs in Biomedicine, 245, 108045.

[8] Moldovanu, S., Răducu, C.H., Căpăţînă, C.S., and Munteanu, O. (2024). Feature selection and machine learning with SHAP values for improved breast cancer prediction. Scientific Reports, 14, 8065.

[9] Sánchez-Andrés, A. and Durán, J. (2023). Explainable machine learning for breast cancer risk assessment using lifestyle factors. Expert Systems with Applications, 215, 119359.

[10] Holzinger, A., Biemann, C., Pattichis, C.S., and Kell, D.B. (2017). What do we need to build explainable AI systems for the medical domain? arXiv preprint arXiv:1712.09923.

[11] Asri, H., Mousannif, H., Al Moatassime, H., and Noel, T. (2016). Using machine learning algorithms for breast cancer risk prediction and diagnosis. Procedia Computer Science, 83, 1064-1069.