 ISSN:2302-8149
e-ISSN:2540-9719

Analisis Perbandingan Teknik Oversampling dan SMOTEENN pada Algoritma Machine Learning untuk Prediksi Penyakit Kanker Payudara

Comparative Analysis of Oversampling and SMOTEENN Techniques in Machine Learning Algorithms for Breast Cancer Prediction

[1]Tri Yulian, [2]Erliyan Redi Susanto*
[1,2]Program Studi Sistem Informasi, Fakultas Teknik dan Ilmu Komputer, Universitas Teknokrat Indonesia
[1,2]Jl. Dahlia No.3, Natar, Lampung Selatan 35362, Indonesia
*e-mail: tri_yulian@teknokrat.ac.id, erliyan.redy@teknokrat.ac.id

(***received***: 10 March 2025, ***revised***: 5 April 2025, ***accepted***: 5 April 2025)

Abstrak

Kanker payudara adalah penyebab utama kematian akibat kanker pada wanita, dengan tantangan utama dalam pengembangan model prediksi adalah ketidakseimbangan kelas dataset medis. Ketidakseimbangan ini menghambat deteksi kelas minoritas (pasien dengan kanker), yang krusial untuk diagnosis dini. Penelitian ini bertujuan menganalisis performa Support Vector Machine (SVM) dan Random Forest dalam prediksi kanker payudara menggunakan teknik praproses oversampling dan SMOTEENN. Dataset yang digunakan adalah SEER Breast Cancer Dataset, yang diseimbangkan menggunakan kedua teknik tersebut. Performa model dievaluasi berdasarkan metrik seperti akurasi, presisi, recall, F1-score. Hasil penelitian menunjukkan bahwa SVM dan Oversampling mencapai akurasi tertinggi sebesar 98.97%, sementara kombinasi SVM dan SMOTEENN mencapai 97.20% Random Forest dan Oversampling mencapai akurasi 96.63% dengan SMOTEENN 95.90%. SVM lebih efektif dalam mengidentifikasi kedua kelas dengan tingkat kesalahan minimal, terutama saat dikombinasikan dengan oversampling . Temuan ini menunjukkan bahwa pemilihan model yang tepat dan teknik praproses data seperti oversampling atau SMOTEENN dapat meningkatkan akurasi prediksi secara signifikan. Penelitian ini memberikan kontribusi penting bagi pengembangan sistem prediksi kanker payudara yang lebih akurat dan andal, mendukung diagnosis dini serta pengambilan keputusan klinis dalam aplikasi medis.

Kata kunci: kanker payudara, *machine learning, support vector machine, random forest, oversampling, SMOTEENN*

Abstract

Breast cancer is the leading cause of cancer-related death among women, with one of the major challenges in developing predictive models being the class imbalance in medical datasets. This imbalance hinders the detection of minority classes (patients with cancer), which is critical for early diagnosis. This study aims to analyze the performance of Support Vector Machine (SVM) and Random Forest algorithms in predicting breast cancer using oversampling and SMOTEENN preprocessing techniques. The dataset used is the SEER Breast Cancer Dataset, which was balanced using both techniques. Model performance was evaluated using metrics such as accuracy, precision, recall, and F1-score. The results showed that SVM with oversampling achieved the highest accuracy of 98.97%, followed by SVM with SMOTEENN at 97.20%. Random Forest with oversampling reached an accuracy of 96.63%, while with SMOTEENN it achieved 95.90%. SVM proved more effective in identifying both classes with minimal error, particularly when combined with oversampling. These findings highlight that selecting the appropriate model and data preprocessing technique—such as oversampling or SMOTEENN—can significantly enhance predictive accuracy. This research contributes to the development of more accurate and reliable breast cancer prediction systems, supporting early diagnosis and clinical decision-making in medical applications.

 ISSN:2302-8149
e-ISSN:2540-9719

Keywords: *breast cancer, machine learning, support vector machine, random forest, oversampling, SMOTENN*

1 Pendahuluan

Kanker payudara merupakan salah satu jenis kanker yang paling umum didiagnosis pada wanita di seluruh dunia dan menjadi penyebab utama kematian akibat kanker pada populasi ini. Menurut data dari Organisasi Kesehatan Dunia (WHO), insiden kanker payudara terus meningkat setiap tahunnya, dengan lebih dari 2,3 juta kasus baru dilaporkan pada tahun 2020 saja [1]. Meskipun telah banyak kemajuan dalam diagnosis dini dan pengobatan, deteksi kanker payudara pada tahap awal tetap menjadi tantangan besar, terutama di negara-negara dengan sumber daya terbatas [2]. Deteksi dini sangat penting karena dapat meningkatkan peluang kesembuhan pasien secara signifikan [3]. Namun, metode diagnosis konvensional sering kali memiliki keterbatasan, seperti subjektivitas interpretasi dan ketidakakuratan hasil, sehingga dibutuhkan pendekatan yang lebih canggih dan andal untuk mendukung proses prediksi kanker payudara [4].

Dalam beberapa tahun terakhir, teknologi machine learning telah menunjukkan potensi besar dalam membantu prediksi kanker payudara secara lebih akurat dan efisien [5]. Berbagai model machine learning, seperti Support Vector Machine (SVM) dan Random Forest, telah dieksplorasi untuk meningkatkan akurasi prediksi berdasarkan dataset medis [6]. SVM dikenal karena kemampuannya dalam menangani data dengan dimensi tinggi dan batas keputusan non-linear, sedangkan Random Forest unggul dalam menangani data kompleks dengan fitur-fitur yang saling berkorelasi [7]. Namun, salah satu masalah utama dalam penerapan kedua model ini adalah ketidakseimbangan dataset, di mana jumlah sampel kelas minoritas (misalnya, pasien dengan kanker) sering kali jauh lebih kecil dibandingkan dengan kelas mayoritas (pasien tanpa kanker). Ketidakseimbangan ini dapat menyebabkan performa model yang bias dan kurang optimal, sehingga memengaruhi akurasi prediksi, terutama dalam mengidentifikasi kelas minoritas yang kritis.

Meskipun berbagai penelitian telah mengeksplorasi penggunaan teknik oversampling seperti RandomOverSampler dan SMOTEENN untuk mengatasi ketidakseimbangan dataset, masih terdapat kesenjangan pengetahuan terkait efektivitas kombinasi model machine learning dengan teknik oversampling dalam konteks prediksi kanker payudara [8]. Selain itu, belum ada penelitian yang secara komprehensif membandingkan performa SVM dan Random Forest yang dikombinasikan dengan teknik oversampling untuk menentukan model yang paling optimal dalam konteks ini [9].

Penelitian ini akan menganalisis dan membandingkan performa model machine learning SVM dan Random Forest, yang dikombinasikan dengan oversampling RandomOverSampler dan SMOTEENN, untuk prediksi kanker payudara. Tujuan penelitian ini adalah untuk memahami struktur dataset, menerapkan preprocessing untuk menangani ketidakseimbangan, melatih model, dan mengevaluasi performa dengan metrik akurasi, presisi, recall, F1-score, dan AUC-ROC. Selain itu, evaluasi akan mencakup visualisasi confusion matrix [10]. Evaluasi ini juga mencakup visualisasi confusion matrix dan kurva ROC untuk memberikan interpretasi lebih mendalam terhadap hasil prediksi.

Hasil penelitian diharapkan dapat memberikan wawasan baru tentang penggunaan oversampling untuk meningkatkan performa model dalam prediksi kanker payudara. Ini dapat menjadi referensi untuk pengembangan sistem prediksi yang akurat dan andal, mendukung diagnosis dini dan keputusan klinis. Selain itu, penelitian ini juga berpotensi memberikan kontribusi signifikan bagi komunitas penelitian machine learning dan dapat diimplementasikan dalam AI untuk diagnosis kanker payudara, meningkatkan kualitas layanan medis.

2 Tinjauan Literatur

Penelitian terkait prediksi kanker payudara menggunakan teknologi machine learning telah berkembang pesat dalam beberapa tahun terakhir, dengan berbagai pendekatan yang diusulkan untuk meningkatkan akurasi dan efektivitas model. Salah satu penelitian sebelumnya yang relevan dilakukan oleh Rina Resmiati dan Toni Arifin dari Universitas Adhirajasa Reswara Sanjaya, yang berjudul "Klasifikasi Pasien Kanker Payudara Menggunakan Metode Support Vector Machine dengan Backward Elimination" [11]. Penelitian ini mengeksplorasi penggunaan metode Support Vector Machine (SVM) untuk klasifikasi pasien kanker payudara pada dataset Breast Cancer Coimbra [12]. Selain itu, penelitian tersebut juga menerapkan teknik Backward Elimination, yaitu proses seleksi

 ISSN:2302-8149
e-ISSN:2540-9719

fitur mundur untuk memilih atribut yang paling relevan dalam meningkatkan performa model. Hasil penelitian menunjukkan bahwa SVM tanpa Backward Elimination menghasilkan akurasi sebesar 65,22% dan nilai AUC sebesar 0,700, yang termasuk dalam kategori Fair Classification . Namun, setelah diterapkannya Backward Elimination, performa model meningkat secara signifikan, dengan akurasi mencapai 95,65% dan nilai AUC sebesar 1,000, yang termasuk dalam kategori Excellent Classification . Temuan ini menunjukkan pentingnya seleksi fitur dalam meningkatkan kemampuan prediksi model machine learning, terutama pada dataset medis yang kompleks [11].

Namun, meskipun hasil tersebut sangat menjanjikan, masih terdapat beberapa keterbatasan yang perlu diatasi. Pertama, penelitian tersebut hanya fokus pada penggunaan SVM sebagai model klasifikasi, sehingga belum memberikan wawasan tentang bagaimana performa model lain seperti Random Forest dibandingkan dengan SVM dalam konteks yang sama. Kedua, penelitian tersebut tidak membahas secara mendalam tantangan yang sering dihadapi dalam dataset medis, seperti ketidakseimbangan kelas, yang dapat memengaruhi performa model secara signifikan. Ketiga, meskipun Backward Elimination berhasil meningkatkan performa model, teknik ini mungkin tidak selalu optimal untuk dataset dengan dimensi tinggi atau fitur-fitur yang saling berkorelasi, sehingga pendekatan alternatif seperti oversampling perlu dieksplorasi lebih lanjut [13].

Penelitian ini berfokus pada dua model utama SVM dan Random Forest, serta penggunaan teknik oversampling RandomOverSampler dan SMOTEENN untuk menangani ketidakseimbangan dataset. Hasil penelitian menunjukkan bahwa SVM mencapai akurasi 98,97%, sedangkan Random Forest mencapai akurasi 96,63%. Perbandingan ini menunjukkan bahwa SVM lebih efektif dalam menangani dataset dengan karakteristik tertentu, sementara Random Forest tetap kompetitif dalam menangani data kompleks.

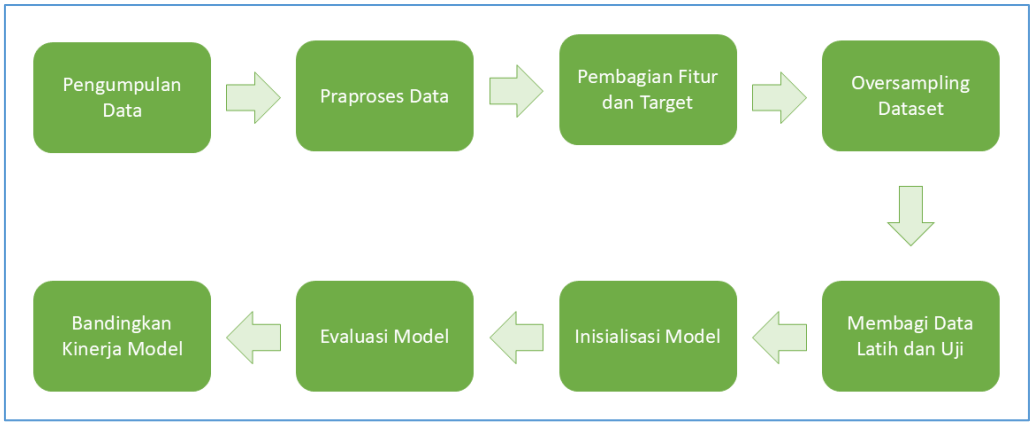
Teknik oversampling seperti RandomOverSampler dan SMOTEENN menunjukkan manfaatnya dalam meningkatkan performa model, terutama dalam mengidentifikasi kelas minoritas yang kritis. Penggunaan metrik evaluasi yang lebih komprehensif, seperti akurasi, presisi, recall, F1-score memberikan gambaran yang lebih mendalam tentang performa model[14].

Secara keseluruhan, tinjauan literatur ini menunjukkan bahwa penelitian ini memiliki kebaruan dalam beberapa aspek. Pertama, penelitian ini secara eksplisit membandingkan performa SVM dan Random Forest dalam konteks ketidakseimbangan dataset, yang jarang dilakukan dalam penelitian sebelumnya. Kedua, penggunaan teknik oversampling (RandomOverSampler) dan SMOTENN memberikan wawasan baru tentang cara mengatasi ketidakseimbangan dataset tanpa bergantung sepenuhnya pada seleksi fitur. Ketiga, penelitian ini menggunakan metrik evaluasi yang lebih komprehensif, yang memberikan gambaran lebih mendalam tentang performa model [15]. Dengan demikian, penelitian ini diharapkan dapat memberikan kontribusi signifikan bagi pengembangan sistem prediksi kanker payudara yang lebih akurat dan andal, yang pada akhirnya dapat mendukung diagnosis dini dan pengambilan keputusan klinis.

3 Metode Penelitian

Tahapan penelitian menggambarkan alur proses dalam penelitian yang melibatkan pemrosesan data dan evaluasi model machine learning. Dimulai dari pemuatan dataset, preprocessing, hingga pembagian data dan pelatihan model, setiap langkah dirancang untuk memastikan analisis yang optimal. Setelah model dievaluasi, hasilnya dibandingkan untuk memilih model dengan performa terbaik. Tahapan studi dalam penelitian ini disajikan pada Gambar 1.

 ISSN:2302-8149
e-ISSN:2540-9719



Gambar 1. Tahapan penelitian

3.1 Tahapan Studi

a) **Pengumpulan Data**
Dataset kanker payudara dari SEER Breast Cancer Dataset dan dimuat menggunakan fungsi pd.read_csv()[16]. Dataset awal dibersihkan untuk menghapus duplikat dan nilai-nilai yang tidak diketahui.

b) **Praproses Data**
Data diproses dengan mengisi nilai kosong pada fitur numerik menggunakan rata-rata (mean) dan modus (mode) untuk fitur kategorikal. Fitur kategorikal di-encode menggunakan LabelEncoder, dan fitur numerik dinormalisasi menggunakan MinMaxScaler.

c) **Pembagian Fitur dan Target**
Dataset dipisahkan menjadi fitur (X) yang mencakup variabel seperti Age, Tumor Size, dan Estrogen Status, serta target (y) yang merupakan variabel Status (hidup/meninggal).

d) **Oversampling Dataset**
Ketidakseimbangan kelas pada variabel target ditangani menggunakan teknik oversampling dengan RandomOverSampler dan SMOTEENN untuk menyeimbangkan distribusi antara kelas Alive dan Dead.

e) **Membagi Data Latih dan Uji**
Dataset hasil oversampling dibagi menjadi data latih dan data uji dengan rasio 80:20 menggunakan fungsi train_test_split untuk memastikan hasil evaluasi yang objektif.

f) **Inisialisasi Model**
Dua model machine learning diinisialisasi: Support Vector Machine (SVM) dengan kernel RBF dan Random Forest dengan 100 estimator. Kedua model siap dilatih menggunakan data latih.

g) **Evaluasi Model**
Model dievaluasi menggunakan data uji dengan metrik seperti akurasi, presisi, recall, F1-score, dan AUC-ROC untuk mengukur kemampuan prediksi terhadap kedua kelas.

h) **Bandingkan Kinerja Model**
Performa kedua model dibandingkan berdasarkan metrik evaluasi. Visualisasi seperti confusion matrix, dan bar cha.rt digunakan untuk memberikan pemahaman mendalam tentang kinerja masing-masing model

3.2 Sumber Data dan Variabel

Dataset berasal dari SEER Breast Cancer Dataset , yang mencakup data pasien kanker payudara dari tahun 2006-2010. Dataset ini terdiri dari 4024 pasien setelah pembersihan data [16].

Variabel Dataset mencakup :

1. **Usia saat Diagnosis**: Usia pasien saat diagnosis kanker, dalam satuan tahun.
2. **Ras**: Ras pasien, dikategorikan sebagai White, Black, atau Other.
3. **Stadium Kanker**: Stadium kanker saat diagnosis, yang mencakup stadium lokal, regional, dan jauh.
4. **Ukuran Tumor**: Ukuran tumor saat diagnosis, dalam satuan milimeter.

 ISSN:2302-8149
e-ISSN:2540-9719

5. **Grade**: Tingkat diferensiasi tumor, yang menunjukkan seberapa agresif kanker tumbuh dan menyebar.
6. **Status Estrogen dan Progesteron**: Status reseptor estrogen dan progesteron pada sel kanker, yang menentukan jenis kanker payudara.
7. **Jumlah Kelenjar Getah Bening Regional**: Jumlah kelenjar getah bening regional yang diperiksa dan positif.
8. **Status Kelangsungan Hidup**: Variabel target yang menunjukkan apakah pasien masih hidup (1) atau meninggal (2)

Dataset yang digunakan dalam penelitian ini memiliki struktur yang dijelaskan pada **Tabel 1**. Tabel tersebut mencantumkan nama kolom, deskripsi untuk memberikan gambaran menyeluruh tentang fitur-fitur yang digunakan dalam analisis.

Tabel 1. Struktur *dataset*

Nama Variabel	Deskripsi
Age	Usia pasien (dalam tahun) saat diagnosis kanker
Race	Ras pasien berdasarkan kategori White, Black, atau Other
Marital Status	Status perkawinan pasien saat diagnosis
T Stage	Stadium tumor primer menurut sistem AJCC (T1, T2, T3, T4)
N Stage	Stadium kelenjar getah bening regional menurut sistem AJCC (N1, N2, N3)
6^{th} Stage	Stadium kanker keseluruhan menurut sistem AJCC edisi ke-6 (IIA, IIB, IIIA, IIIB, IIIC)
Differentiate	Tingkat diferensiasi tumor (Well differentiated, Moderately differentiated, Poorly differentiated)
Grade	Grade tumor (1 = rendah, 2 = sedang, 3 = tinggi)
A Stage	Stadium anatomi kanker (Regional, Distant)
Tumor Size	Ukuran tumor saat diagnosis (dalam milimeter)
Estrogen Status	Status reseptor estrogen pada sel kanker
Progesterone Status	Status reseptor progesteron pada sel kanker
Regional Node Examined	Jumlah kelenjar getah bening regional yang diperiksa oleh patolog
Regional Node Positive	Jumlah kelenjar getah bening regional yang positif
Survival Months	Jumlah bulan kelangsungan hidup pasien sejak diagnosis
Status	Variabel target yang menunjukkan status kelangsungan hidup pasien

3.3 Support Vector Machine

Support Vector Machine (SVM) adalah salah satu algoritma pembelajaran mesin yang digunakan untuk tugas klasifikasi dan regresi. SVM bekerja dengan mencari hyperplane terbaik yang memisahkan data ke dalam kelas-kelas yang berbeda dengan margin maksimal. Hyperplane ini adalah garis pemisah multidimensi yang memaksimalkan jarak antara titik-titik data dari kelas yang berbeda, sehingga meningkatkan kemampuan generalisasi model[17].

Dalam penelitian ini, SVM dengan kernel RBF digunakan untuk memprediksi status kelangsungan hidup pasien, karena efektif dalam menangani data dengan batas keputusan non-linear seperti pada dataset kanker payudara. Secara matematis, hyperplane optimal yang dicari oleh SVM dapat didefinisikan melalui Persamaan (1).

$$w \cdot x + b = 0 \quad (1)$$

Keterangan:
w: Vektor bobot yang menentukan arah hyperplane.
x: Fitur input (data pasien).
b: Bias atau konstanta yang menggeser hyperplane.
Untuk kasus non-linear, kernel RBF digunakan untuk mentransformasi data ke ruang dimensi yang lebih tinggi. seperti pada Persamaan (2).

 ISSN:2302-8149
e-ISSN:2540-9719

$K(x_i, x_j) = \exp\left(-\gamma \parallel x_i - x_j \parallel^2\right)$ (2)

Keterangan:

$K(x_i, x_j)$: Nilai similaritas antara dua titik data x_i dan x_j

γ : Parameter yang mengontrol pengaruh jarak antar titik data. Semakin besar nilai γ, semakin sempit pengaruh setiap titik data

menghitung keputusan klasifikasi berdasarkan kontribusi semua support vector seperti Persamaan (3).

$f(x) = \text{sign}\left(\sum_{i=1}^{N} \alpha_i y_i K(x_i, x) + b\right)$ (3)

Keterangan:

α_i : Bobot yang diberikan pada setiap titik data (diperoleh selama pelatihan).

y_i : Label kelas dari titik data *xi* (+1 untuk Alive, −1 untuk Dead).

$f(x)$: Fungsi keputusan yang menghasilkan prediksi kelas untuk data baru *x*.

Dalam penelitian ini, persamaan (1) menentukan *hyperplane* pemisah optimal, sementara persamaan (2) digunakan untuk menangani kasus non-linear melalui kernel RBF. Akhirnya, persamaan (3) digunakan untuk melakukan prediksi berdasarkan bobot yang diperoleh dari proses pelatihan.

3.4 Random Forest

Random Forest adalah algoritma pembelajaran mesin berbasis ensemble learning yang digunakan untuk klasifikasi dan regresi. Algoritma ini membangun banyak pohon keputusan (decision trees) dan menggabungkan hasilnya melalui voting (klasifikasi) atau rata-rata (regresi)[18]. karena kemampuannya menangani dataset kompleks dengan banyak fitur, menghasilkan akurasi tinggi, serta memberikan interpretasi yang lebih mudah. Random Forest bekerja dengan melatih sejumlah pohon keputusan independen menggunakan subset data latih (bagging) dan subset fitur yang dipilih secara acak[19]. Prediksi akhir diperoleh dengan menggabungkan hasil dari semua pohon. Dalam penelitian ini, model digunakan untuk memprediksi status kelangsungan hidup pasien (Alive atau Dead) berdasarkan fitur seperti Age, Tumor Size, dan Estrogen Status.

Prediksi akhir dari model Random Forest ditentukan oleh mayoritas voting dari seluruh pohon, sebagaimana ditunjukkan pada Persamaan (4).

$\hat{y} = \text{mode}(T_1(X), T_2(X), \ldots, T_N(X))$ (4)

Keterangan:

$T_i(X)$: Prediksi dari pohon keputusan ke-*i* terhadap input *X*

N : Jumlah total pohon dalam hutan (n_estimators).

mode : menunjukkan prediksi terbanyak dari seluruh pohon. Pengambilan subset data dan fitur secara acak bertujuan untuk meningkatkan generalisasi model dan menghindari overfitting.

Persamaan (4) menunjukkan bahwa prediksi akhir diperoleh melalui voting mayoritas dari seluruh pohon. Setiap pohon dilatih menggunakan subset data dan fitur yang dipilih secara acak untuk meningkatkan generalisasi dan mengurangi overfitting, sehingga menghasilkan prediksi yang lebih stabil dan akurat.

3.5 Oversampling

Oversampling adalah teknik yang digunakan untuk menangani ketidakseimbangan kelas dalam dataset, di mana jumlah sampel pada kelas minoritas (Dead) jauh lebih sedikit dibandingkan dengan kelas mayoritas (Alive)[14]. Ketidakseimbangan ini dapat menyebabkan model machine learning cenderung bias terhadap kelas mayoritas, sehingga performa prediksi untuk kelas minoritas menjadi kurang optimal. Dalam penelitian ini, teknik oversampling diterapkan menggunakan algoritma RandomOverSampler, yang bekerja dengan menggandakan sampel acak dari kelas minoritas hingga distribusi kelas menjadi seimbang[20].

Formula Matematis RandomOverSampler

Secara matematis, proses oversampling dapat dijelaskan sebagai berikut:

Dataset Awal:

Terdiri dari dataset N_m sampel untuk kelas mayoritas (Alive) dan N_n sampel untuk kelas minoritas (Dead).

$N_n < N_m$ (5)

Persamaan (5) menunjukkan adanya ketidakseimbangan kelas yang dapat menyebabkan bias terhadap kelas mayoritas selama pelatihan.

 ISSN:2302-8149
e-ISSN:2540-9719

Tujuan Oversampling :
Tujuan dari oversampling adalah membuat jumlah sampel kelas minoritas sama dengan jumlah sampel kelas mayoritas, yaitu:

$$N_n' < N_m \quad (6)$$

Persamaan (6) menunjukkan bahwa proses *oversampling* bertujuan untuk menyeimbangkan jumlah sampel di kedua kelas agar model dapat belajar dari kedua kelas secara proporsional.

Proses Penggandaan Sampel :
Untuk mencapai $N_n' = N_m$, algoritma **RandomOverSampler** secara acak memilih sampel dari kelas minoritas dan menggandakannya hingga jumlahnya mencapai N_m. Menggunakan persamaan (7).

$$X_{\text{minoritas\_baru}} = \text{RandomlySample}(X_{\text{minoritas}}, X_m - N_n \quad (7)$$

Keterangan:
$X_{\text{minoritas}}$: Dataset asli untuk kelas minoritas.
$X_{\text{minoritas\_baru}}$: Dataset baru untuk kelas minoritas setelah oversampling.
RandomlySample : Fungsi yang secara acak memilih sampel dari dataset asli untuk digandakan.

Persamaan (7) menjelaskan mekanisme pembentukan data baru dari kelas minoritas melalui pemilihan dan penggandaan acak.

Dataset Baru Setelah Oversampling:

$$X_{\text{new}} = X_{\text{mayoritas}} \cup X_{\text{minoritas\_baru}} \quad (8)$$

Keterangan:
X_{new} : Dataset gabungan setelah oversampling.
$X_{\text{mayoritas}}$: Dataset asli untuk kelas mayoritas (tidak berubah).
$X_{\text{minoritas\_baru}}$: Dataset baru untuk kelas minoritas setelah oversampling.

Persamaan (8) menunjukkan bahwa hasil akhir adalah dataset seimbang yang siap digunakan untuk pelatihan model.

Distribusi Kelas Akhir:
Setelah oversampling, distribusi kelas menjadi seimbang:

$$|X_{\text{mayoritas}}| = |X_{\text{minoritas\_baru}}| \quad (9)$$

Persamaan (9) memastikan bahwa jumlah sampel dari kedua kelas telah disamakan, sehingga model dapat melakukan pembelajaran tanpa bias terhadap salah satu kelas.

3.6 SMOTTENN

Teknik SMOTE-ENN merupakan kombinasi dari dua metode: Synthetic Minority Oversampling Technique (SMOTE) dan Edited Nearest Neighbors (ENN). SMOTE digunakan untuk menyeimbangkan distribusi kelas dengan menghasilkan sampel sintetis pada kelas minoritas, sedangkan ENN membersihkan dataset dari noise atau sampel yang tidak konsisten. Dalam penelitian ini, SMOTE-ENN diterapkan untuk mengatasi ketidakseimbangan kelas pada dataset kanker payudara, sehingga meningkatkan kemampuan model dalam memprediksi status kelangsungan hidup pasien (Alive atau Dead).

Untuk setiap sampel minoritas x_i, pilih tetangga terdekat x_j berdasarkan jarak Euclidean:

$$d(x_i, x_j = \sqrt{\sum_{k=1}^{m}(x_{ik} - x_{jk})^2} \quad (10)$$

Persamaan (10) digunakan untuk menentukan tetangga terdekat yang akan digunakan dalam proses interpolasi SMOTE.

Buat sampel sintetis x_{new} menggunakan interpolasi linear:

$$x_{new} = x_i + \lambda(x_j - x_i) \quad (11)$$

Persamaan (11) menjelaskan bahwa SMOTE tidak melakukan duplikasi, melainkan menciptakan variasi baru di sepanjang garis antara dua titik minoritas.

Gabungkan SMOTE dan ENN:

$$X_{final} = X_{majoritas} \cup x_{bersih} \quad (12)$$

Persamaan (12) menggambarkan hasil akhir SMOTE-ENN, yang merupakan kombinasi dari data mayoritas dan data minoritas sintetis yang telah difilter, menghasilkan dataset yang lebih seimbang dan berkualitas untuk pelatihan model.

 ISSN:2302-8149
e-ISSN:2540-9719

3.7 Confusion Matrix

Confusion Matrix adalah alat evaluasi yang digunakan untuk mengukur performa model klasifikasi dengan membandingkan prediksi model terhadap nilai sebenarnya (ground truth)[21]. Dalam penelitian ini, confusion matrix digunakan untuk mengevaluasi kemampuan model dalam memprediksi status kelangsungan hidup pasien (Alive atau Dead) berdasarkan dataset kanker payudara.

Struktur Confusion Matrix

Confusion matrix memiliki empat komponen utama:

1. **True Positive (TP)**: Jumlah sampel kelas Dead yang diprediksi dengan benar sebagai Dead.
2. **True Negative (TN)**: Jumlah sampel kelas Alive yang diprediksi dengan benar sebagai Alive.
3. **False Positive (FP)**: Jumlah sampel kelas Alive yang salah diprediksi sebagai Dead.
4. **False Negative (FN)**: Jumlah sampel kelas Dead yang salah diprediksi sebagai Alive.

Secara visual, confusion matrix dapat disusun seperti pada **Tabel 2**.

Tabel 2 Confusion Matrix

Prediksi/Actual	Dead (Positif)	Alive (Negatif)
Dead (Positif)	TP	FP
Alive (Negatif)	FN	TN

Formula Sistematis

Berdasarkan confusion matrix, metrik evaluasi seperti akurasi , presisi , recall , dan F1-score dapat dihitung menggunakan rumus berikut:

Akurasi (Accuracy) :Proporsi total prediksi yang benar.

$$\text{Accuracy} = TP + TN \frac{TP+TN}{TP+TN+FP+FN} \quad (13)$$

Persamaan (13) menunjukkan bahwa semakin tinggi nilai akurasi, semakin baik model dalam melakukan prediksi secara keseluruhan.

Presisi (Precision) : Kemampuan model untuk memprediksi kelas Dead dengan benar.

$$\text{Precision} = \frac{TP}{TP+FP} \quad (14)$$

Persamaan (14) menunjukkan bahwa presisi tinggi berarti model memiliki sedikit *false positive*, yaitu kasus di mana pasien *Alive* salah diprediksi sebagai *Dead.*

Recall (Sensitivity) : Kemampuan model untuk mengidentifikasi semua sampel kelas Dead .

$$\text{Recall} = \frac{TP}{TP+FN} \quad (15)$$

Persamaan (15) menunjukkan bahwa semakin tinggi nilai *recall*, semakin kecil jumlah *false negative*, yaitu kasus di mana pasien *Dead* tidak terdeteksi oleh model.

F1-Score : Rata-rata harmonik antara presisi dan recall, memberikan gambaran seimbang tentang performa model.

$$\text{F1} - \text{Score} = 2 \cdot \frac{\text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}} \quad (16)$$

Persamaan (16) menunjukkan bahwa *F1-score* tinggi berarti model memiliki keseimbangan yang baik antara presisi dan recall, yang sangat penting dalam aplikasi medis.

Hasil confusion matrix adalah:

TP = 95 : Sampel **Dead** yang diprediksi dengan benar.
TN = 97 : Sampel **Alive** yang diprediksi dengan benar.
FP = 5 : Sampel **Alive** yang salah diprediksi sebagai Dead.
FN = 3 : Sampel **Dead** yang salah diprediksi sebagai Alive.

Dengan nilai ini, metrik evaluasi dapat dihitung menggunakan persamaan (13) hingga (16) untuk menilai kinerja model.

4 Hasil dan Pembahasan

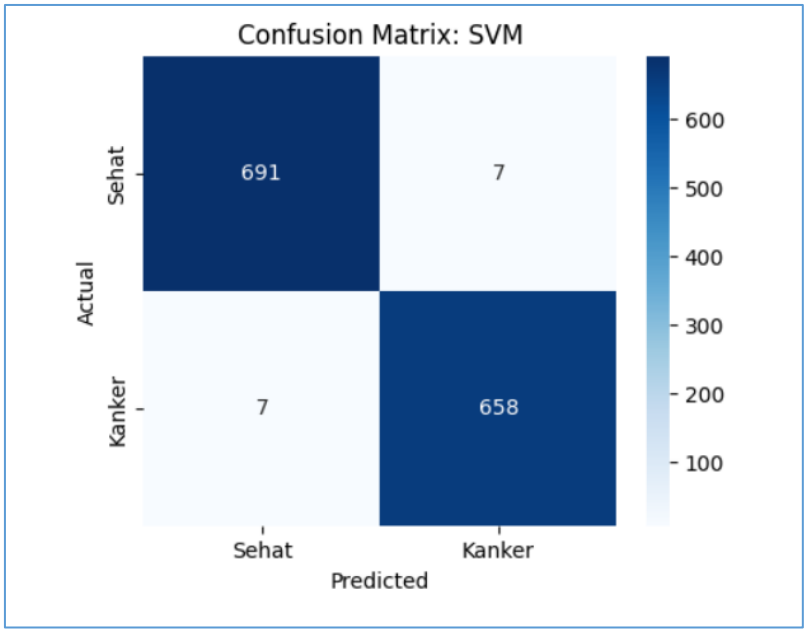
Dalam studi ini, kami menggunakan dua metode pembelajaran mesin, yaitu Support Vector Machine (SVM) dan Random Forest , untuk memprediksi status kelangsungan hidup pasien kanker payudara (Alive atau Dead). Dataset yang digunakan mengalami ketidakseimbangan kelas, sehingga diterapkan dua teknik penyeimbangan kelas: oversampling dengan RandomOverSampler dan SMOTE-ENN. Teknik RandomOverSampler digunakan untuk menyeimbangkan kelas dengan

 ISSN:2302-8149
e-ISSN:2540-9719

menggandakan sampel asli dari kelas minoritas, sementara SMOTE-ENN menggabungkan pembangkitan sampel sintetis menggunakan SMOTE dan pembersihan noise menggunakan Edited Nearest Neighbors (ENN). Berikut adalah hasil dan pembahasan terkait performa kedua model menggunakan kedua teknik penyeimbangan kelas.

4.1 Metode Support Vector Machine

Untuk mengevaluasi kinerja model secara visual, Confusion Matrix disajikan dalam bentuk heatmap. Confusion Matrix adalah alat yang umum digunakan dalam pembelajaran mesin untuk mengevaluasi kinerja model klasifikasi. Matriks ini menunjukkan jumlah prediksi benar (True Positives dan True Negatives) serta prediksi salah (False Positives dan False Negatives) untuk setiap kelas yang disajikan **Gambar 2.**



Gambar 2. Confusion matrix support vector machine

Penjelasan **Gambar 2** Confusion Matrix Support Vector Machine
Berdasarkan keterangan gambar:

a) True Positives (TP): Jumlah contoh yang tepat diprediksi sebagai kelas positif (Dead). Di sini, nilainya adalah 658.
b) True Negatives (TN): Jumlah contoh yang tepat diprediksi sebagai kelas negatif (Alive). Di sini, nilainya adalah 691.
c) False Positives (FP): Jumlah contoh yang salah diprediksi sebagai kelas positif (Dead) dari kelas negatif (Alive). Di sini, nilainya adalah 7.
d) False Negatives (FN): Jumlah contoh yang salah diprediksi sebagai kelas negatif (Alive) dari kelas positif (Dead). Di sini, nilainya adalah 7.

Confusion Matrix ini menunjukkan bahwa SVM memiliki jumlah false positives dan false negatives yang sangat rendah (7 sampel masing-masing), yang penting dalam aplikasi medis karena kesalahan dalam memprediksi kelas Dead dapat berdampak signifikan pada diagnosis pasien. Model ini sangat efektif dalam mengidentifikasi kedua kelas dengan akurasi tinggi.

Untuk memberikan gambaran lebih jelas tentang performa model SVM , Classification Report disajikan dalam bentuk **Gambar 3**. Classification Report mencakup metrik seperti precision, recall, f1-score, dan support untuk setiap kelas (Alive dan Dead), serta nilai keseluruhan untuk akurasi dan rata-rata metrik.

 ISSN:2302-8149
e-ISSN:2540-9719

```
=== Classification Report: SVM ===
              precision    recall  f1-score   support

           0       0.99      0.99      0.99       698
           1       0.99      0.99      0.99       665

    accuracy                           0.99      1363
   macro avg       0.99      0.99      0.99      1363
weighted avg       0.99      0.99      0.99      1363

Accuracy: 0.9897
Precision: 0.9895
Recall: 0.9895
F1 Score: 0.9895
```

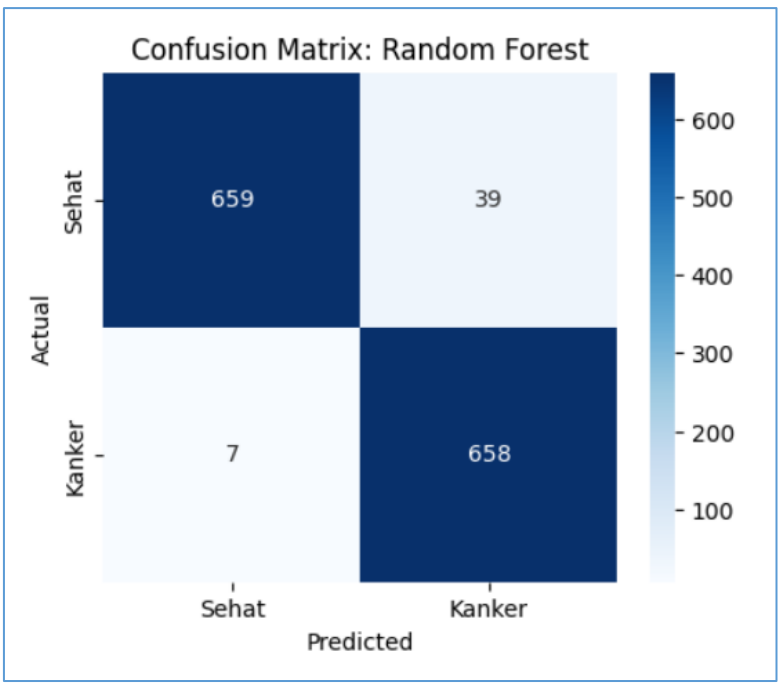
Gambar 3. Classification report support vector machine

Penjelasan Gambar 3 Classification Report

a) **Precision**: Nilai precision untuk kelas Alive dan Dead adalah 0.9895, yang menunjukkan bahwa model sangat akurat dalam memprediksi kedua kelas.
b) **Recall**: Nilai recall untuk kelas Alive dan Dead adalah 0.9895, yang menunjukkan bahwa model efektif dalam mengidentifikasi semua sampel kelas minoritas (Dead).
c) **F1-Score**: Nilai F1-score untuk kedua kelas adalah 0.9895, yang menunjukkan keseimbangan yang sangat baik antara precision dan recall.
d) **Accuracy**: Akurasi keseluruhan model mencapai 98.97%, yang menunjukkan bahwa model ini sangat andal dalam memprediksi status kelangsungan hidup pasien.

4.2 Metode Random Forest

Untuk mengevaluasi kinerja model Random Forest , Confusion Matrix disajikan dalam bentuk heatmap. Confusion Matrix adalah alat yang umum digunakan dalam pembelajaran mesin untuk mengevaluasi kinerja model klasifikasi. Matriks ini menunjukkan jumlah prediksi benar (True Positives dan True Negatives) serta prediksi salah (False Positives dan False Negatives) untuk setiap kelas pada **Gambar 4**.



Gambar 4. Confusion matrix random forest

 ISSN:2302-8149
e-ISSN:2540-9719

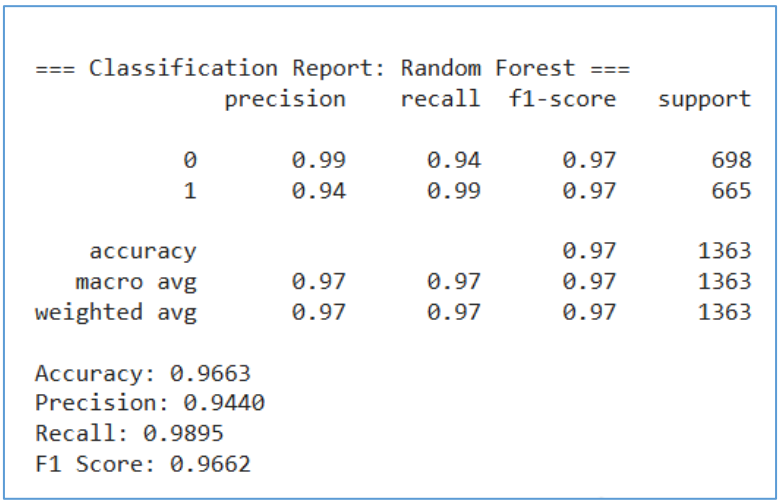
Confusion Matrix untuk Random Forest

Berdasarkan keterangan gambar:

a) **True Positives (TP)**: Jumlah contoh yang tepat diprediksi sebagai kelas positif (Dead). Di sini, nilainya adalah 658.
b) **True Negatives (TN)**: Jumlah contoh yang tepat diprediksi sebagai kelas negatif (Alive). Di sini, nilainya adalah 659.
c) **False Positives (FP)**: Jumlah contoh yang salah diprediksi sebagai kelas positif (Dead) dari kelas negatif (Alive). Di sini, nilainya adalah 39.
d) **False Negatives (FN)**: Jumlah contoh yang salah diprediksi sebagai kelas negatif (Alive) dari kelas positif (Dead). Di sini, nilainya adalah 7.

Confusion Matrix ini menunjukkan bahwa Random Forest memiliki jumlah false positives yang lebih tinggi (39 sampel) dibandingkan dengan SVM (7 sampel), meskipun jumlah false negatives tetap rendah (7 sampel). Hal ini menunjukkan bahwa Random Forest lebih rentan terhadap kesalahan dalam memprediksi kelas Alive sebagai Dead, yang dapat memengaruhi keputusan medis.

Untuk memberikan gambaran lebih jelas tentang performa model Random Forest , Classification Report disajikan dalam bentuk **Gambar 5**. Classification Report mencakup metrik seperti precision , recall , f1-score , dan support untuk setiap kelas (Alive dan Dead), serta nilai keseluruhan untuk akurasi dan rata-rata metrik.



```
=== Classification Report: Random Forest ===
              precision    recall  f1-score   support

           0       0.99      0.94      0.97       698
           1       0.94      0.99      0.97       665

    accuracy                           0.97      1363
   macro avg       0.97      0.97      0.97      1363
weighted avg       0.97      0.97      0.97      1363

Accuracy: 0.9663
Precision: 0.9440
Recall: 0.9895
F1 Score: 0.9662
```

Gambar 5. Classification report random forest

a) **Precision**: Nilai precision untuk kelas Alive adalah 0.9440, sedangkan untuk kelas Dead adalah 0.9895. Hal ini menunjukkan bahwa model lebih akurat dalam memprediksi kelas Dead dibandingkan dengan kelas Alive.
b) **Recall**: Nilai recall untuk kelas Alive dan Dead adalah 0.9895, yang menunjukkan bahwa model efektif dalam mengidentifikasi semua sampel kelas minoritas (Dead).
c) **F1-Score**: Nilai F1-score untuk kedua kelas adalah 0.9662, yang menunjukkan keseimbangan yang baik antara precision dan recall.
d) **Accuracy**: Akurasi keseluruhan model mencapai 96.63%, yang masih kompetitif namun sedikit lebih rendah dibandingkan dengan SVM.

Dalam penelitian sebelumnya menggunakan 1 metode yaitu Support Vector Machine dengan Backward Elimination dan yang tidak dalam klasifikasi pasien kanker payudara dengan pembagian 80% untuk data training dan 20% untuk data testing. didapatkan hasil seperti pada **Tabel 3.**

Tabel 3. hasil penelitian terdahulu

Model	Accuracy
Support Vector Machine	65,22%
Support Vector Machine + Backward Elimination	95,65%

 ISSN:2302-8149
e-ISSN:2540-9719

Berdasarkan pada **Tabel 3**, metode SVM dengan Elimination Backward memiliki nilai accuracy tertinggi sebesar 95%. Pembagian data dengan uji 80 : 20 algoritma SVM dan Random Forest dengan Teknik Oversampling didapatkan hasil seperti **Tabel 4:**

Tabel 4. Hasil evaluasi model dengan teknik Oversampling

Model	Precision	Recall	F1-Score	Accuracy
SVM + Oversampling	98.95%	98.95%	98.95%	98.97%
Random Forest + Oversampling	94.40%	98.95%	96.62%	96.63%

a. **SVM + Oversampling**: Model SVM dengan teknik oversampling (RandomOverSampler) memberikan hasil terbaik secara keseluruhan, dengan akurasi 98.97%, presisi 98.95%, dan recall 98.95%. Teknik ini sangat efektif dalam meningkatkan jumlah sampel minoritas, sehingga model mampu mendeteksi kelas minoritas dengan sangat baik. Namun, risiko overfitting lebih tinggi karena penggandaan sampel asli tanpa mengatasi noise.
b. **Random Forest + Oversampling**: Model Random Forest dengan oversampling mencapai akurasi 96.63%, presisi 94.40%, dan recall 98.95%. Meskipun recall sangat tinggi, nilai presisi lebih rendah dibandingkan SVM, menunjukkan bahwa model ini lebih rentan terhadap false positives . Teknik oversampling membantu meningkatkan deteksi kelas minoritas, namun kurang optimal dalam mengatasi noise.

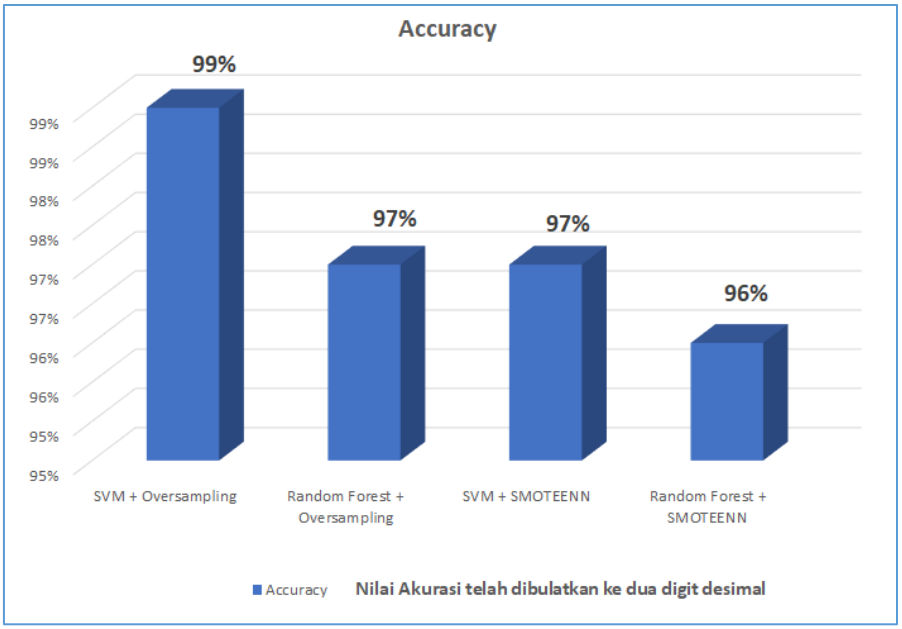
Tabel 5. Hasil evaluasi model dengan teknik SMOTEENN

Model	Precision	Recall	F1-Score	Accuracy
SVM + SMOTEENN	99.83%	95.41%	97.57%	97.20%
Random Forest + SMOTEEN	96.67%	96.36%	96.51%	95.90%

a. **SVM + SMOTEEN:** dengan teknik SMOTEENN menunjukkan performa yang sangat baik, dengan akurasi 97.20%, presisi 99.83%, dan recall 95.41%. Teknik SMOTENN berhasil meningkatkan keseimbangan antara presisi dan recall, meskipun akurasi sedikit lebih rendah dibandingkan dengan oversampling. Model ini sangat andal dalam memprediksi kelas positif (Dead) dengan tingkat kesalahan minimal.
b. **Random Forest + SMOTEEN:** Model Random Forest dengan SMOTEENN mencapai akurasi 95.90%, presisi 96.67%, dan recall 96.36%. Performa model ini stabil dengan keseimbangan yang baik antara presisi dan recall. Meskipun akurasi lebih rendah dibandingkan SVM, Random Forest tetap kompetitif dan cocok untuk dataset dengan noise karena SMOTEENN membersihkan sampel tidak konsisten hasil.

Pada **Gambar 6** Untuk memberikan gambaran komparatif performa model, berikut disajikan visualisasi dalam bentuk bar chart yang menampilkan nilai akurasi tertinggi dan terendah dari kombinasi metode pembelajaran mesin (Support Vector Machine dan Random Forest) dengan teknik penyeimbangan kelas (Oversampling dan SMOTEENN). Nilai akurasi telah dibulatkan ke dua digit desimal untuk keseragaman. Bar chart ini membandingkan efektivitas masing-masing kombinasi dalam memprediksi status kelangsungan hidup pasien kanker payudara (Alive atau Dead), sehingga dapat terlihat metode dan teknik terbaik (akurasi tertinggi) serta yang memerlukan perbaikan (akurasi terendah).

 ISSN:2302-8149
e-ISSN:2540-9719



Gambar 6. Grafik perbandingan hasil pengujian berdasarkan akurasi

5 Kesimpulan

Berdasarkan hasil penelitian, Support Vector Machine (SVM) terbukti menjadi model terbaik untuk prediksi status kelangsungan hidup pasien kanker payudara dibandingkan dengan Random Forest, dengan performa yang dipengaruhi oleh teknik penyeimbangan kelas. Menggunakan oversampling (RandomOverSampler), SVM mencapai akurasi tertinggi sebesar 98.97%, presisi 98.95%, recall 98.95%, dan F1-score 98.95%, menunjukkan kemampuan luar biasa dalam mengidentifikasi kedua kelas (Alive dan Dead) dengan tingkat kesalahan minimal. Sementara itu, Random Forest menunjukkan performa kompetitif dengan akurasi 96.63%, namun memiliki jumlah false positives lebih tinggi (39 sampel), menunjukkan kelemahan dalam memprediksi kelas Alive sebagai Dead. Di sisi lain, dengan teknik SMOTEENN, SVM tetap unggul dengan akurasi 97.20%, presisi 99.83%, recall 95.41%, dan F1-score 97.57%, meskipun akurasinya sedikit lebih rendah dibandingkan oversampling, teknik ini menghasilkan model yang lebih robust karena mengurangi noise melalui pembersihan data. Random Forest dengan SMOTEENN juga menunjukkan stabilitas dengan akurasi 95.90%, presisi 96.67%, recall 96.36%, dan F1-score 96.51%, meskipun performanya sedikit di bawah SVM. Secara keseluruhan, oversampling lebih cocok jika prioritas utama adalah akurasi maksimal, sementara SMOTEENN direkomendasikan untuk menghasilkan model yang lebih andal dan stabil dalam aplikasi medis. Penelitian ini menunjukkan bahwa pemilihan model machine learning yang tepat, dikombinasikan dengan teknik praproses data seperti oversampling atau SMOTEENN, dapat meningkatkan akurasi prediksi secara signifikan, terutama dalam mendeteksi kelas minoritas yang kritis untuk diagnosis dini.

Referensi

[1] M. Arnold *et al.*, “Current and Future Burden of Breast Cancer: Global Statistics for 2020 and 2040,” *The Breast*, vol. 66, hal. 15–23, Des 2022, doi: 10.1016/j.breast.2022.08.010.

[2] C. H. Barrios, “Global Challenges in Breast Cancer Detection and Treatment,” *The Breast*, vol. 62, hal. S3–S6, Mar 2022, doi: 10.1016/j.breast.2022.02.003.

[3] U. Naseem *et al.*, “An Automatic Detection of Breast Cancer Diagnosis and Prognosis based on Machine Learning using Ensemble of Classifiers,” *IEEE Access*, vol. 10, hal. 78242–78252, 2022, doi: 10.1109/ACCESS.2022.3174599.

[4] M. Nasser dan U. K. Yusof, “Deep Learning based Methods for Breast Cancer Diagnosis: A Systematic Review and Future Direction,” *Diagnostics*, vol. 13, no. 1, hal. 161, Jan 2023, doi: 10.3390/diagnostics13010161.

[5] R. Rabiei, S. M. Ayyoubzadeh, S. Sohrabei, M. Esmaeili, dan A. Atashi, “Prediction of Breast

 ISSN:2302-8149
e-ISSN:2540-9719

Cancer using Machine Learning Approaches," *J. Biomed. Phys. Eng.*, vol. 12, no. 3, hal. 297–308, 2022, doi: 10.31661/jbpe.v0i0.2109-1403.

[6] P. Dinesh, A. S. Vickram, dan P. Kalyanasundaram, "Medical Image Prediction for Diagnosis of Breast Cancer Disease Comparing the Machine Learning Algorithms: SVM, KNN, Logistic Regression, Random Forest and Decision Tree to Measure Accuracy," 2024, hal. 020140. doi: 10.1063/5.0203746.

[7] E. Y. Boateng, J. Otoo, dan D. A. Abaye, "Basic Tenets of Classification Algorithms K-Nearest-Neighbor, Support Vector Machine, Random Forest and Neural Network: A Review," *J. Data Anal. Inf. Process.*, vol. 08, no. 04, hal. 341–357, 2020, doi: 10.4236/jdaip.2020.84020.

[8] J. A. Benítez-Andrades, C. Prada-García, N. Ordás-Reyes, M. E. Blanco, A. Merayo, dan A. Serrano-García, "Enhanced Prediction of Spine Surgery Outcomes using Advanced Machine Learning Techniques and Oversampling Methods," *Heal. Inf. Sci. Syst.*, vol. 13, no. 1, hal. 24, Mar 2025, doi: 10.1007/s13755-025-00343-9.

[9] M. Khushi *et al.*, "A Comparative Performance Analysis of Data Resampling Methods on Imbalance Medical Data," *IEEE Access*, vol. 9, hal. 109960–109975, 2021, doi: 10.1109/ACCESS.2021.3102399.

[10] E. F. Agyemang *et al.*, "Addressing Class Imbalance Problem in Health Data Classification: Practical Application from an Oversampling Viewpoint," *Appl. Comput. Intell. Soft Comput.*, vol. 2025, no. 1, Jan 2025, doi: 10.1155/acis/1013769.

[11] R. Resmiati dan T. Arifin, "Klasifikasi Pasien Kanker Payudara menggunakan Metode Support Vector Machine dengan Backward Elimination," *Sistemasi*, vol. 10, no. 2, hal. 381, 2021, doi: 10.32520/stmsi.v10i2.1238.

[12] M. M. Hassan *et al.*, "A Comparative Assessment of Machine Learning Algorithms with the Least Absolute Shrinkage and Selection Operator for Breast Cancer Detection and Prediction," *Decis. Anal. J.*, vol. 7, hal. 100245, Jun 2023, doi: 10.1016/j.dajour.2023.100245.

[13] S. Bej, N. Davtyan, M. Wolfien, M. Nassar, dan O. Wolkenhauer, "LoRAS: An Oversampling Approach for Imbalanced Datasets," *Mach. Learn.*, vol. 110, no. 2, hal. 279–301, Feb 2021, doi: 10.1007/s10994-020-05913-4.

[14] F. Gurcan dan A. Soylu, "Learning from Imbalanced Data: Integration of Advanced Resampling Techniques and Machine Learning Models for Enhanced Cancer Diagnosis and Prognosis," *Cancers (Basel).*, vol. 16, no. 19, hal. 3417, Okt 2024, doi: 10.3390/cancers16193417.

[15] G. Menghani, "Efficient Deep Learning: A Survey on Making Deep Learning Models Smaller, Faster, and Better," *ACM Comput. Surv.*, vol. 55, no. 12, hal. 1–37, Des 2023, doi: 10.1145/3578938.

[16] jing teng (North China Electric Power University), "SEER Breast Cancer Data," IEEE Dataport. [Daring]. Tersedia pada: https://ieee-dataport.org/open-access/seer-breast-cancer-data

[17] D. A. Pisner dan D. M. Schnyer, "Support Vector Machine," in *Machine Learning*, Elsevier, 2020, hal. 101–121. doi: 10.1016/B978-0-12-815739-8.00006-7.

[18] G. Dagnew dan B. H. Shekar, "Ensemble Learning-based Classification of Microarray Cancer Data on Tree-based Features," *Cogn. Comput. Syst.*, vol. 3, no. 1, hal. 48–60, Mar 2021, doi: 10.1049/ccs2.12003.

[19] N. Syam dan R. Kaul, "Random Forest, Bagging, and Boosting of Decision Trees," in *Machine Learning and Artificial Intelligence in Marketing and Sales*, Emerald Publishing Limited, 2021, hal. 139–182. doi: 10.1108/978-1-80043-880-420211006.

[20] S. A. Alex, J. J. Vedha Nayahi, dan S. Kaddoura, "Deep Convolutional Neural Networks with Genetic Algorithm-based Synthetic Minority Over-Sampling Technique for Improved Imbalanced Data Classification," *Appl. Soft Comput.*, vol. 156, hal. 111491, Mei 2024, doi: 10.1016/j.asoc.2024.111491.

[21] D. Krstinić, M. Braović, L. Šerić, dan D. Božić-Štulić, "Multi-Label Classifier Performance Evaluation with Confusion Matrix," in *Computer Science & Information Technology*, AIRCC Publishing Corporation, Jun 2020, hal. 01–14. doi: 10.5121/csit.2020.100801.