# KLASIFIKASI KANKER PAYUDARA MENGGUNAKAN EXTRA TREE DENGAN SMOTE

**Kartika Handayani, Erni, Badariatul Lailiah, Rabiatus Sa'adah**
Sistem Informasi, Universitas Bina Sarana Informatika
Jalan Abdurrahman Saleh No.18A Pontianak, Indonesia
*kartika.kth@bsi.ac.id*

**ABSTRAK**

Kanker payudara merupakan jenis kanker yang seringkali didiagnosis pada wanita. Di Indonesia, kanker payudara merupakan jenis kanker dengan tingkat kejadian tertinggi yang menempati peringkat kedua setelah kanker serviks. Mengidentifikasi kanker payudara pada tahap awal sangat krusial untuk mencegah perkembangan yang cepat, selain dari perkembangan metode pencegahan. Metode pembelajaran mesin (ML) dapat digunakan untuk mengindentifikasi kanker payudara. Dalam kasus klasifikasi kanker payudara, mayoritas data yang digunakan mengalami permasalahan data ketidakseimbangan kelas antara kanker payudara dan non-kanker payudara. Untuk mengatasi permasalahan ini digunakan teknik *Synthetic Minority Over-sampling Technique* (SMOTE) dengan algoritma *extra tree*. *Extra Tree* dipilih karena kemampuannya dalam menangani kasus-kasus kompleks dan tingkat akurasi yang tinggi. Hasil eksperimen menunjukkan keberhasilan metode *Extra Tree* dengan SMOTE, dengan tingkat akurasi mencapai 96.71%, *Recall* sebesar 95.29%, *Precision* sebesar 96.13%, *F1 Score* sebesar 95.61%, dan *Area Under the Curve* (AUC) sebesar 99.46%. Temuan ini mengindikasikan bahwa penggunaan metode ini memberikan kontribusi positif dalam meningkatkan kemampuan klasifikasi kanker payudara, yang dapat berpotensi meningkatkan keberhasilan deteksi dini dan pengelolaan penyakit ini dalam praktik klinis.

***Kata kunci:*** *kanker payudara, extra tree, smote*

## 1. PENDAHULUAN

Kanker payudara merupakan jenis kanker yang seringkali didiagnosis pada wanita, dengan lebih dari 1 dari 10 kasus kanker baru terdeteksi setiap tahun. Ini merupakan penyebab kematian kedua paling umum akibat kanker pada perempuan di seluruh dunia. Pertumbuhan kanker payudara seringkali tidak menimbulkan gejala yang jelas, dan sebagian besar kasus ditemukan melalui pemeriksaan secara berkala [1].

Di Indonesia, kanker payudara merupakan jenis kanker dengan tingkat kejadian tertinggi yang menempati peringkat kedua setelah kanker serviks. Terdapat kecenderungan peningkatan insiden kanker payudara dari tahun ke tahun. Sebagian besar kasus kanker payudara umumnya terdeteksi pada stadium lanjut. Jumlah kasus kanker payudara yang baru terjadi di Indonesia diperkirakan sekitar 65.858 kasus setiap tahun, dengan memperhatikan populasi sekitar 273.523.621 orang. Berdasarkan data Riskesdas 2018, prevalensi kanker secara umum di Indonesia mencapai 1.017.290 kasus, dan di wilayah Sulawesi Selatan saja terdapat sekitar 33.693 kasus [2].

Mengidentifikasi kanker payudara pada tahap awal sangat krusial untuk mencegah perkembangan yang cepat, selain dari perkembangan metode pencegahan [3]. Memberikan dukungan untuk pemulihan sedini mungkin dan mengurangi risiko kematian [4]. Meskipun dokter dapat melakukan diagnosis kanker payudara secara manual, proses ini memerlukan waktu yang lebih lama dan dapat menjadi rumit bagi dokter untuk menerapkan klasifikasi [5].

Walaupun kemajuan dalam bidang pengobatan kanker payudara telah mengakibatkan penurunan angka kematian akibat kanker payudara di seluruh kelompok umur, risiko tinggi dan tingkat kelangsungan hidup yang rendah masih menjadi permasalahan pada kelompok usia muda [6]. Dikarenakan tingkat keparahan penyakit, diperlukan sistem deteksi berbantuan komputer (CAD) yang menerapkan metode pembelajaran mesin (ML) untuk mendiagnosis kanker payudara [7]

Metode machine learning memiliki potensi untuk mengatasi batasan tersebut. Pendekatan ini telah diaplikasikan dalam beberapa studi sebelumnya dengan melakukan perbandingan. beberapa metode seperti neural network, decision tree, naïve bayes, k-nearest neighbor, logistic regresion, random forest, dansupport vector machines[8] [9]. Penelitian tersebut menghasilkan nilai akurasi paling tinggi yaitu sebesar 95,00 %. Pada penelitian tersebut tidak menampilkan hasil dari matriks evaluasi lainnya.

Berdasarkan penjelasan di atas, penelitian ini mengusulkan penerapan *machine learning* dalam klasifikasi kanker payudara. Data yang digunakan dalam penelitian memiliki masalah dengan ketidakseimbangan data. Sehingga dilakukan pengujian dengan *Synthetic Minority Over Sampling Technique* (SMOTE). Penggunaan pendekatan ketidakseimbangan kelas sampling untuk menguji pada dataset yang digunakan dapat memberikan hasil performa yang lebih baik atau tidak.

Penelitian ini mengusulkan model prediktif *Extra Tree* dan diharapakan dapat menghasilkan

*accuracy, recall, precision, f-score* dan *AUC* yang lebih baik dari penelitian sebelumnya dan menambah kontribusi penelitian terkait data yang digunakan.

## 2. TINJAUAN PUSTAKA

Untuk mendukung penelitian ini, tinjauan pustaka atau kerangka pemikiran yang relevan dibutuhkan agar penelitian ini berdasarkan pada sumber yang dihasilkan terstruktur dan sesuai pada intinya, khususnya pada masalah penelitian terkait.

### 2.1. Data Mining

Data *mining* merupakan proses eksplorasi untuk mendapatkan informasi atau pengetahuan dari kumpulan data yang besar. Saat menerapkan teknik penambangan data pada sejumlah besar data, tujuannya adalah mengidentifikasi pola menarik dan hubungan tersembunyi. Selain itu, ada beberapa istilah alternatif yang digunakan untuk menggambarkan konsep ini, seperti ekstraksi pengetahuan, penemuan pengetahuan, analisis pola data, analisis data dalam database, dan lain sebagainya. [10]

Menurut [11]serangkaian proses tahapan memiliki tujuh (7) tahapan yaitu :

1. Pembersihan Data (*data cleaning*)
   Proses pembersihan data melibatkan penghapusan informasi yang tidak relevan, seringkali setelah data tersebut dianalisis dengan membandingkannya dengan hipotesis yang telah disusun. Hal ini bertujuan untuk mempermudah identifikasi hasil yang diinginkan pada tahap analisis berikutnya.
2. Integrasi data (*data integration*)
   Integrasi data adalah proses menggabungkan informasi dari berbagai sumber database ke dalam satu database baru. Seringkali, data yang diperlukan diambil dari berbagai database atau file teks.
3. Seleksi data (*data selection*)
   Data dalam database seringkali tidak semuanya relevan atau diperlukan, oleh karena itu, diperlukan suatu proses pemilihan data untuk menyaring informasi yang benar-benar diperlukan dalam langkah-langkah berikutnya.
4. Transformasi data (*data transformation*)
   Data disatukan atau dimodifikasi sesuai dengan metode yang digunakan dalam data mining. Hal ini disebabkan oleh kebutuhan beberapa format khusus data mining dalam pengolahannya.
5. Proses *mining*
   Merupakan suatu proses penggalian informasi yang tersembunyi dari sebuah basis data atau kumpulan data dengan tujuan memperoleh pemahaman yang lebih dalam dari data yang telah diolah.
6. Evaluasi Pola (*pattern evaluation*)
   Dalam tahap ini, output dari teknik data mining berupa pola-pola akan diuji terhadap hipotesis yang telah dibuat sebelumnya. Tujuannya adalah memperoleh kesimpulan yang mendekati hasil atau hipotesa untuk digunakan dalam langkah-langkah selanjutnya. Ini melibatkan presentasi pengetahuan (knowledge presentation).

Langkah ini merupakan tahap akhir dalam proses data mining, di mana hasil analisis yang telah dilakukan disajikan melalui implementasi. Dengan demikian, kesimpulan yang diperoleh menjadi lebih nyata dan dapat dipresentasikan dengan jelas.

### 2.2. Kanker Payudara

Kanker payudara, atau carcinoma mammae, adalah jenis keganasan yang berasal dari jaringan payudara, termasuk baik epitel duktus maupun lobulusnya. Terjadinya kanker payudara disebabkan oleh kondisi sel yang kehilangan kontrol dan mekanisme normalnya, sehingga menyebabkan pertumbuhan yang tidak normal, cepat, dan tidak terkendali [12].

### 2.3. *Extra Tree*

*Extra tree* membuat sekelompok pohon keputusan yang tidak dipangkas sesuai dengan metode standar top-down. Prediksi seluruh pohon digabungkan untuk menentukan prediksi akhir, melalui alternatif mayoritas[13]. Pengklasifikasi *extra tree* menghasilkan kelipatan acak dari pohon pilihan dengan sub-sampel yang sangat berbeda tanpa melakukan bootstrap. Hal ini menghindari masalah pemasangan yang berlebihan dan meningkatkan akurasi. Efisiensi juga menjadi kekuatan utama algoritma ini[14].

### 2.4. *Synthetic Minority Over Sampling Technique* (SMOTE)

Teknik Synthetic Minority Over Sampling Technique (SMOTE) menghasilkan sampel buatan dari kelas minoritas dengan melakukan interpolasi pada instance yang ada yang sangat dekat satu sama lain [15]. Untuk kategori minoritas dalam kumpulan informasi, SMOTE secara awal memilih instance data dari kelas minoritas secara acak. Setelah itu, k tetangga terdekat Ma, yang terkait dengan kelas minoritas, diidentifikasi. Sebuah contoh data kedua, Mb, kemudian dipilih dari k set tetangga terdekat. Selama proses ini, Ma dan Mb terhubung membentuk garis di ruang fitur. Data buatan baru kemudian dihasilkan sebagai kombinasi dari Ma dan Mb. Proses ini berulang hingga dataset mencapai keseimbangan antara kelas minoritas dan mayoritas. Karena keefektifan SMOTE, teknik pengambilan sampel ini telah berkembang luas sebagai perluasan dari berbagai metode pengambilan sampel. [16]. SMOTE berjalan dengan proses berikut:

$$x_{new} = x_i + (\hat{x}_k - x_i)\delta$$

$x_{new}$= Data sintetis baru
$x_i$ = Data ke-i kelas minoritas
$\hat{x}_k$ = *Data k tetangga terdekat dari* $x_i$
$\delta$ = Bilangan acak (0 dan 1)

### 2.5. Evaluasi Model

Model Evaluasi yang bisa digunakan untuk mengukur nilai optimal dari berbagai parameter dari setiap metode, prediksi yang berasal dari semua kombinasi parameter yang mungkin dievaluasi pada dataset klasifikasi digunakan

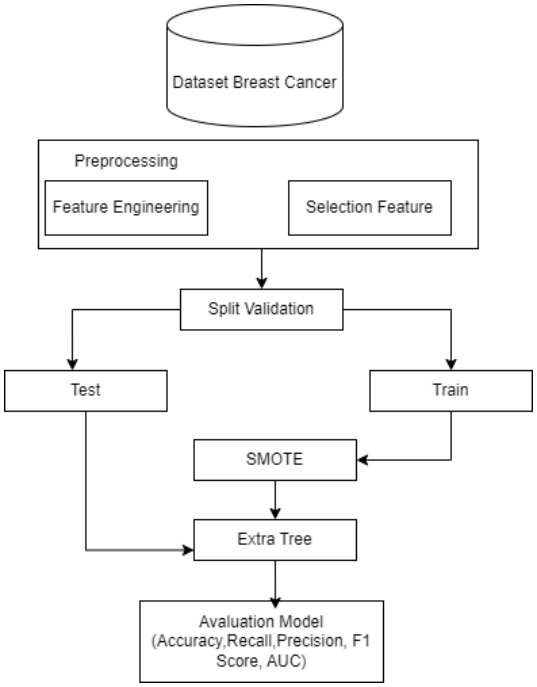
5 nilai yaitu: *accuracy, precision, recall, f1-score* dan *AUC/ROC.* Adapun penjelasan dan rumus-rumus dalam perhitungannya sebagai berikut [17].

1. *Accuracy*
   *Accuracy* adalah metrik evaluasi yang paling umum digunakan untuk klasifikasi. Namun, untuk masalah klasifikasi data yang tidak seimbang, akurasi mungkin bukan pilihan yang baik karena akurasi sering kali memiliki bias terhadap kelas mayoritas [18].
   $$Accuracy = \frac{(TP+TN)}{Totalsample}\ x\ 100 \text{..... (1)}$$
2. *Precision*
   *Precision* adalah bagian dari data yang diambil sesuai dengan informasi yang diperlukan [19].
   $$Precision = \frac{(TP)}{(TP+FP)}\ x\ 100\% \text{ ..... (2)}$$
3. *Recall*
   *Recall* adalah pengumpulan data yang berhasil diambil dari bagian data yang relevan dengan queri [19].
   $$Recall = \frac{(TP)}{(TP+FN)}\ x\ 100\% \text{ ..... (3)}$$
4. *AUC/ROC*
   Kurva ROC (*Receiver Operating Characteristics*) diakui sebagai pilihan paling rasional untuk data yang tidak seimbang, yang menggambarkan pertukaran relatif antara manfaat dan biaya (Fawcett). Kurva ROC dihasilkan berdasarkan matriks dasar dalam *machine learning* yang disebut matriks kebingungan adalah matriks kebingungan dari masalah klasifikasi biner [18].
   Rumus menghitung AUC/ROC adalah:
   $$AUC = (\frac{1+sensitifity-FPR}{2}.) \text{.....(4)}$$
5. *F-1 Score*
   Menggabungkan *precision* dan *recall* menjadi satu ukuran. Secara matematis *F-1 score* merupakan rata-rata harmonik dari *precision* dan *recall*. Untuk menghitung *F-1 score binary classification* [20] .
   Rumus menghitung *F-1 score* adalah:
   $$F-1 = 2\frac{(Precision.Recall)}{(Precision+Recall)} \text{ ..... ..(5)}$$

Keterangan :
TN = *True Negative*
FP = *False Positive*
FN = *False Negative*
TP = *True Positive*
FPR = *False Positive Rate*

## 3. METODE PENELITIAN

Berikut gambaran tahapan penelitian yang dilakukan




Gambar 1. Tahapan Penelitian

### 3.1. Dataset

Dataset yang digunakan adalah datase yang terdapat pada situs kaggle dengan link : **https://www.kaggle.com/datasets/yasserh/breast-cancer-dataset/data.** Data terdiri dari 569 data rekam medis, dengan 32 atribut. Atribut ditunjukkan pada Tabel 1.

Tabel 1. Data Atribut

| No | Atribut | Keterangan |
|---|---|---|
| 1 | *id* | Tidak Digunakan |
| 2 | *diagnosis* | Digunakan |
| 3 | *radius_mean* | Digunakan |
| 4 | *texture_mean* | Digunakan |
| 5 | *perimeter_mean* | Digunakan |
| 6 | *area_mean* | Digunakan |
| 7 | *smoothness_mean* | Digunakan |
| 8 | *compactness_mean* | Digunakan |
| 9 | *concavity_mean* | Digunakan |
| 10 | *concave points_mean* | Digunakan |
| 11 | *symmetry_mean* | Digunakan |
| 12 | *fractal_dimension_mean* | Digunakan |
| 13 | *radius_se* | Digunakan |
| 14 | *texture_se* | Digunakan |
| 15 | *perimeter_se* | Digunakan |
| 16 | *area_se* | Digunakan |
| 17 | *smoothness_se* | Digunakan |
| 18 | *compactness_se* | Digunakan |
| 19 | *concavity_se* | Digunakan |
| 20 | *concave points_se* | Digunakan |
| 21 | *symmetry_se* | Digunakan |

| No | Atribut | Keterangan |
|---|---|---|
| 22 | *fractal_dimension_se* | Digunakan |
| 23 | *radius_worst* | Digunakan |
| 24 | *texture_worst* | Digunakan |
| 25 | *perimeter_worst* | Digunakan |
| 26 | *area_worst* | Digunakan |
| 27 | *smoothness_worst* | Digunakan |
| 28 | *compactness_worst* | Digunakan |
| 29 | *concavity_worst* | Digunakan |
| 30 | *concave points_worst* | Digunakan |
| 31 | *symmetry_worst* | Digunakan |
| 32 | *fractal_dimension_worst* | Digunakan |

Data yang terdiri dari 569 data memiliki dua label *class* yaitu yaitu *benign (B) atau* jinak dan atau *malignant (M) atau ganas*. 357 data yang jinak dan 212 data yang ganas. Berikut adalah statistik distribusi kelas dapat dilihat pada Tabel 2.

Tabel 2. Distribusi Kelas *Dataset*

| Nama Kelas | Jumlah Data |
|---|---|
| *Benign* | 357 |
| *Malignant* | 212 |

### 3.2. *Preprocessing*

Dilakukan *feature engineering* Untuk mengubah nilai pada variabel yang berbentuk kategorikal menjadi bentuk numerik, digunakan teknik *Label Encoding* dengan bantuan *Label Encoder* dari *library Scikit Learn* pada atribut target dataset tersebut. Proses ini melibatkan transformasi label klasifikasi dari "*category*" menjadi "*numeric*". Pada dataset asli label target "*malignant*" diubah menjadi "1" dan "*benign*" diubah menjadi "0". Dilakukan *selection feature* dengan menghapus atribut ID.

### 3.3. *Split Validation*

Langkah berikutnya adalah melakukan validasi split, yang bertujuan untuk membagi data menjadi dua bagian, yaitu data pelatihan dan data pengujian. Data pelatihan digunakan untuk melatih model, sementara data pengujian digunakan untuk menguji dan mengevaluasi model. Dalam penelitian ini, pembagian dilakukan dengan menggunakan persentase split sebesar 80:20, di mana 80% data digunakan sebagai data pelatihan dan 20% data digunakan sebagai data pengujian.

### 3.4. *SMOTE*

Setelah melihat statistik dari distribusi kelas data yang tidak seimbang, penulis selanjutnya melakukan eksperimen dengan menambahkan SMOTE. Secara singkat cara kerja SMOTE sebagai berikut

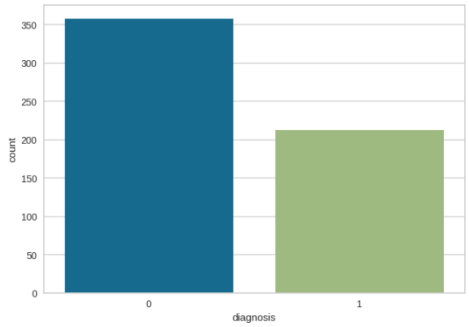
1) Pilih satu sampel Data ke-i kelas minoritas
2) Temukan k-tetangga dari Data ke-i kelas minorita dalam kelas minoritas secara acak.
3) Hitung perbedaan antara Data ke-i kelas minoritas acak dan data k tetangga terdekat dari $x_i$
4) Kalikan perbedaan dengan bilangan acak
5) Tambahkan hasil perkalian ke Data ke-i kelas minoritas untuk mendapatkan sampel sintetis

### 3.5. Evaluasi Model

Evaluasi model digunakan untuk mengetahui kinerja dari hasil uji model algoritma LightGBM dengan *resampling.* Digunakan *Accuracy*, *Precision*, *Recall*, AUC dam F1-Score untuk mengevaluasi kinerja model.

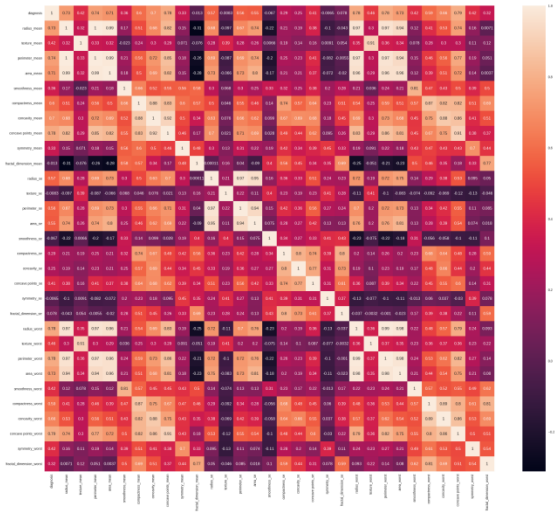
## 4. HASIL DAN PEMBAHASAN

### 4.1. Analisis Statistika Dataset




Gambar 2. Distribusi Kelas Dataset

Berdasarkan distribusi kelas pada Gambar 2 dapat dilihat bahwa terjadinya *imbalance* pada kelas jumlah klasifikasi kanker payudara (0) jinak sebanyak 357 dan ganas (1) sebanyak 212.



Gambar 3. *Feature Correlation*

### 4.2. *Preprocessing*

*Preprocessing* yang dilakukan dalam penelitian ada beberapa tahapan diantaranya menghapus atribut yang tidak dibutuhkan, penambahan atribut yang baru, *feature engineering dan* penghapusan fitur ID. *feature engineering* dilakukan pada fitur *diagnosis.* Berikut data contoh data sebelum dilakukan transformasi dengan gambaran data pada Tabel 3 dan pada Tabel 4 setelah dilakukan trasformasi.

Tabel 3. Fitur *diagnosis* Sebelum Transformasi

| ***diagnosis*** |
|---|
| M |
| M |
| B |
| B |
| B |

Tabel 4. Fitur *Gender* Setelah Transformasi

| ***diagnosis*** |
|---|
| 1 |
| 1 |
| 0 |
| 0 |
| 0 |

### 4.3. Hasil Eksperimen Dan Analisis

Setelah dilakukan tahapan penelitian, berikut hasil dari pengujian dengan *Extra Tree*:

Tabel 5. Hasil Pengujian

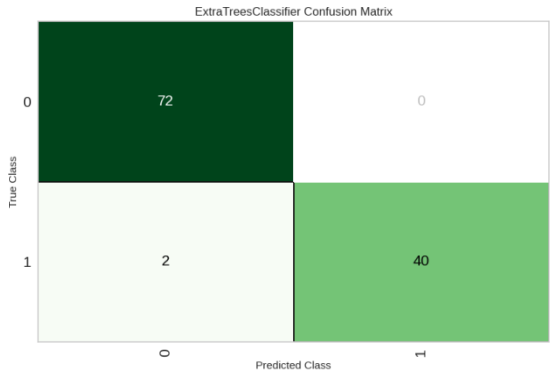
| Metode | Accuracy | Recall | Precision | F1 |
|---|---|---|---|---|
| **Sebelum SMOTE** | 96,71% | 94,12% | 97,05% | 95,47% |
| **Setelah SMOTE** | 96,71% | 95,29% | 96,13% | 95,61% |

Berdasarkan hasil pengujian hasil tanpa menggunakan teknik SMOTE diperoleh accuracy 96,71%, recall 95,29% dimana terjadi peningkatan sebelum menggunakan SMOTE, precision 96,13% dan F1-Score 95,61% dimana terjadi peningkatan sebelum menggunakan SMOTE. Sedangkan hasil AUC pada tabel 6.

Tabel 6. Hasil AUC

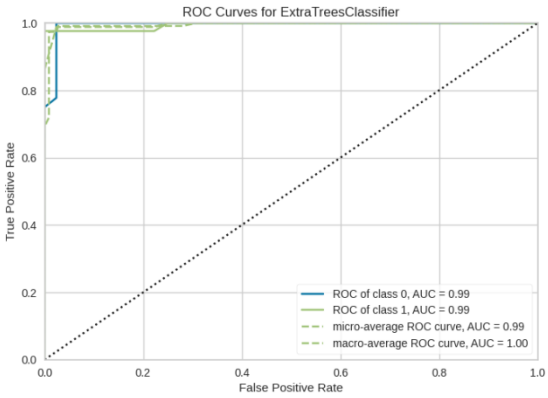
| Metode | AUC |
|---|---|
| **Sebelum SMOTE** | 0.9937 |
| **Setelah SMOTE** | 0.9946 |

Berdasarkan hasil pengujian, AUC menggunakan teknik SMOTE mengalami hasil peningkatan walau tidak seignifikan Hasil ini termasuk dalam *excellent classification.* AUC *excellent classification* menunjukkan kemampuan model untuk memisahkan antara kelas normal dan kelas anomali dengan nilai AUC diantara 0,90 – 1,00.




Gambar 4. *confusion matrix*

Pada Gambar diatas dapat disimpulkan berdasarkan hasil evaluasi menggunakan confusion matriks pasien kanker payudara yang diprediksi pasien tersebut dan benar menderita kanker payudara sebanyak 40, sedangkan pasien tidak menderita kanker payudara dan diprediksi benar tidak menderita kanker sebanyak 72. Pada pasien tidak menderita kanker payudara namun diprediksi kanker payudara sebanyak 2. Kemudian pada pasien menderita kanker payudara yang diprediksi tidak menderita kanker payudara sebanyak *0.*




Gambar 5. *AUC*

Pada gambar diatas dijelaskan bahwa kurva sangat mendekati angka 1.0, dengan hasil 0.99 sehingga dapat disimpulkan bahwa model klasifikasi ini tergolong *excellent classification* yang menunjukkan kemampuan model untuk memisahkan antara kelas normal dan kelas anomali dengan nilai AUC diantara 0,90 – 1,00.

## 5. KESIMPULAN DAN SARAN

Berdasarkan hasil penelitian, dapat disimpulkan bahwa dalam klasfikasi kanker payudata menggunakan *extra tree* dengan SMOTE, memperoleh hasil accuracy 96,71%, recall 95,29% dimana terjadi peningkatan sebelum menggunakan SMOTE, precision 96,13% dan F1-Score 95,61% dimana terjadi peningkatan sebelum menggunakan SMOTE dan AUC 0.9946 dimana terjadi peningkatan sebelum menggunakan SMOTE. Penelitian selanjutnya dapat dilakukan dengan perbandingan metode lain dan melakukan *prerocessing* data yang belum dilakukan pada penelitian ini.

## DAFTAR PUSTAKA

[1] A. Rizka, M. K. Akbar, and N. A. Putri, "CARCINOMA MAMMAE SINISTRA T4bN2M1 METASTASIS PLEURA," *AVERROUS J. Kedokt. dan Kesehat. Malikussaleh*, vol. 8, no. 1, p. 23, 2022, doi: 10.29103/averrous.v8i1.7006.

[2] A. Herawati, S. Rijal, A. St Fahira Arsal, R. Purnamasari, D. Amelia Abdi, and S. Wahid, "Karakteristik Kanker Payudara," *Fakumi Med. J. J. Mhs. Kedokt.*, vol. 2, no. 5, pp. 359–367,

2022.

[3] Y. S. Sun *et al.*, "Risk factors and preventions of breast cancer," *Int. J. Biol. Sci.*, vol. 13, no. 11, pp. 1387–1397, 2017, doi: 10.7150/ijbs.21635.

[4] E. M. Badr, M. A. Salam, and H. Ahmed, "Optimizing Support Vector Machine using Gray Wolf Optimizer Algorithm for Breast Cancer Detection," vol. 3, no. November, 2019.

[5] N. Khuriwal and N. Mishra, "Breast Cancer Diagnosis Using Deep Learning Algorithm," *Proc. - IEEE 2018 Int. Conf. Adv. Comput. Commun. Control Networking, ICACCCN 2018*, pp. 98–103, 2018, doi: 10.1109/ICACCCN.2018.8748777.

[6] H. B. Lee and W. Han, "Unique features of young age breast cancer and its management," *J. Breast Cancer*, vol. 17, no. 4, pp. 301–307, 2014, doi: 10.4048/jbc.2014.17.4.301.

[7] D. A. Omondiagbe, S. Veeramani, and A. S. Sidhu, "Machine Learning Classification Techniques for Breast Cancer Diagnosis," *IOP Conf. Ser. Mater. Sci. Eng.*, vol. 495, no. 1, 2019, doi: 10.1088/1757-899X/495/1/012033.

[8] C. Chazar and B. Erawan, "Machine Learning Diagnosis Kanker Payudara Menggunakan Algoritma Support Vector Machine," *Inf. (Jurnal Inform. dan Sist. Informasi)*, vol. 12, no. 1, pp. 67–80, 2020, doi: 10.37424/informasi.v12i1.48.

[9] N. R. Muntiari and K. H. Hanif, "Klasifikasi Penyakit Kanker Payudara Menggunakan Perbandingan Algoritma Machine Learning," *J. Ilmu Komput. dan Teknol.*, vol. 3, no. 1, pp. 1–6, 2022, doi: 10.35960/ikomti.v3i1.766.

[10] K. K. Pandey and D. Shukla, "Challenges of big data to big data mining with their processing framework," *Proc. - 2018 8th Int. Conf. Commun. Syst. Netw. Technol. CSNT 2018*, pp. 89–94, 2018, doi: 10.1109/CSNT.2018.8820282.

[11] Y. Mardi, "Data Mining : Klasifikasi Menggunakan Algoritma C4.5," *Edik Inform.*, vol. 2, no. 2, pp. 213–219, 2017, doi: 10.22202/ei.2016.v2i2.1465.

[12] Z. A. H. Nurhayati, *Analisis Faktor-Faktor Yang Berhubungan Dengan Kejadian Kanker Payudara*. 2019.

[13] E. K. Ampomah, Z. Qin, and G. Nyame, "Evaluation of tree-based ensemble machine learning models in predicting stock price direction of movement," *Inf.*, vol. 11, no. 6, 2020, doi: 10.3390/info11060332.

[14] A. Zafari, R. Zurita-Milla, and E. Izquierdo-Verdiguier, "Land Cover Classification Using Extremely Randomized Trees: A Kernel Perspective," *IEEE Geosci. Remote Sens. Lett.*, vol. 17, no. 10, pp. 1702–1706, 2020, doi: 10.1109/LGRS.2019.2953778.

[15] E. Rendón, R. Alejo, C. Castorena, F. J. Isidro-Ortega, and E. E. Granda-Gutiérrez, "Data sampling methods to dealwith the big data multi-class imbalance problem," *Appl. Sci.*, vol. 10, no. 4, 2020, doi: 10.3390/app10041276.

[16] M. Dorn *et al.*, "Comparison of machine learning techniques to handle imbalanced COVID-19 CBC datasets," *PeerJ Comput. Sci.*, vol. 7, p. e670, 2021, doi: 10.7717/peerj-cs.670.

[17] K. Handayani and E. Erni, "Penerapan Light Gradient Boosting Dalam Prediksi Rasio Klik Tayang," *JATI (Jurnal Mhs. Tek. Inform.*, vol. 7, no. 1, pp. 13–18, 2023, doi: 10.36040/jati.v7i1.6010.

[18] G. Haixiang, L. Yijing, L. Yanan, L. Xiao, and L. Jinling, "BPSO-Adaboost-KNN ensemble learning algorithm for multi-class imbalanced data classification," *Eng. Appl. Artif. Intell.*, vol. 49, pp. 176–193, 2016, doi: 10.1016/j.engappai.2015.09.011.

[19] H. M. Nawawi, S. Rahayu, J. J. Purnama, and S. I. Komputer, "Algoritma c4.5 untuk memprediksi pengambilan keputusan memilih deposito berjangka," vol. 16, no. 1, pp. 65–72, 2019.

[20] J. Mohajon, "Confusion Matrix for Your Multi-Class Machine Learning Model," *Towar. data Sci.*, 2020.