

**ANALYSIS OF GRADIENT BOOSTED TREES ALGORITHM
IN BREAST CANCER CLASSIFICATION**

**LAPORAN NON-REGULER
SCIENTIST**

Diajukan untuk memenuhi salah satu syarat mencapai derajat Sarjana
Program Studi Informatika



Disusun oleh :

CANTIKA OKZEN SURYAPUTRI

22.11.4734

**FAKULTAS ILMU KOMPUTER
UNIVERSITAS AMIKOM YOGYAKARTA
YOGYAKARTA**

2026

ANALYSIS OF GRADIENT BOOSTED TREES ALGORITHM IN BREAST CANCER CLASSIFICATION

LAPORAN NON-REGULER

Diajukan untuk memenuhi salah satu syarat mencapai derajat Sarjana
Program Studi Informatika



Disusun oleh :

CANTIKA OKZEN SURYAPUTRI

22.11.4734

**FAKULTAS ILMU KOMPUTER
UNIVERSITAS AMIKOM YOGYAKARTA
YOGYAKARTA**

2026

HALAMAN PERSETUJUAN

JALUR NON-REGULER

**ANALYSIS OF GRADIENT BOOSTED TREES ALGORITHM IN BREAST
CANCER CLASSIFICATION**

yang disusun dan diajukan oleh
Cantika Okzen Suryaputri
22.11.4734

telah disetujui oleh Dosen Pembimbing
pada tanggal 24 Februari 2026

Dosen Pembimbing,


Mujid Rohardi, S.Kom., M.Eng.
NIK. 190302393

HALAMAN PENGESAHAN

JALUR NON-REGULER

**ANALYSIS OF GRADIENT BOOSTED TREES ALGORITHM IN BREAST
CANCER CLASSIFICATION**

yang disusun dan diajukan oleh

Cantika Okzen Suryaputri
22.11.4734

Telah dipertuhankan di depan Dewan Penguji
pada tanggal 24 Februari 2026

Susunan Dewan Penguji

Nama Penguji

Tanda Tangan

Subektiingsih, S.Kom., M.Kom.
NIK. 190302413

Bayu Setiaji, M.Kom.
NIK. 190302216

Majid Rahardi, S.Kom., M.Eng.
NIK. 190302193

Laporan ini telah diterima sebagai salah satu persyaratan
untuk memperoleh gelar Sarjana Komputer
Tanggal 24 Februari 2026

DEKAN FAKULTAS ILMU KOMPUTER



Prof. Dr. Kusriadi, M.Kom.
NIK. 190302106

HALAMAN PERNYATAAN KEASLIAN KARYA

Yang bertandatangan di bawah ini,

Nama mahasiswa : Cantika Okren Suryaputri

NIM : 22.31.4734

Menyatakan bahwa Laporan dengan judul berikut

Analysis of Gradient Boosted Tree Algorithm in Breast Cancer Classification

Dosen Pembimbing : Majid Rifa'idi, S.Kom., M.Eng

1. Karya tulis ini adalah benar-benar ASLI dan BELUM PERNAH dipaparkan untuk mendapatkan gelar akademik, baik di Universitas AMIKOM Yogyakarta maupun di Perguruan Tinggi lainnya.
2. Karya tulis ini merupakan gagasan, pemikiran dan kegiatan SAYA sendiri, tanpa bantuan pihak lain kecuali arahan dari Dosen Pembimbing.
3. Dalam karya tulis ini tidak terdapat karya atau pendapat orang lain, kecuali secara tertulis dengan jelas dicantumkan sebagai acuan dalam naskah dengan disebutkan nama pengarang dan disebutkan dalam Daftar Pustaka pada karya tulis ini.
4. Perangkat lunak yang digunakan dalam penelitian ini sepenuhnya menjadi tanggung jawab SAYA, bukan tanggung jawab Universitas AMIKOM Yogyakarta.
5. Pernyataan ini SAYA buat dengan sesungguhnya, apabila di kemudian hari terdapat penyimpangan dan ketidak-hormatan dalam pernyataan ini, maka SAYA bersedia menerima SANKSI AKADEMIK dengan pembatasan nilai yang sudah diperoleh, serta sanksi lainnya sesuai dengan norma yang berlaku di Perguruan Tinggi.

Yogyakarta, 24 Februari 2026

Yang Menyatakan,



Cantika Okren Suryaputri

HALAMAN PERSEMBAHAN

Tiada lembar skripsi yang paling indah dalam laporan skripsi ini kecuali lembar persembahan. Skripsi ini penulis persembahkan segenap ketulusan hati untuk :

1. Sembah sujud serta syukur kepada Allah SWT, sang pemilik semesta, atas segala nikmat, kekuatan, dan petunjuk-Nya yang tak tertingga, sehingga penulis mampu melampaui setiap kesulitan dalam menyelesaikan skripsi ini dengan lancar dan baik.
2. Ibu dan Ayah saya, nafas dan sumber kekuatan utama penulis. Terima kasih atas doa-doa tulus yang dilantikan di setiap sujud, kasih sayang yang tak bertepi, serta pengorbanan yang menjadi fondasi bagi penulis untuk berdiri tegak hingga hari ini.
3. Kepada adik tercinta, yang dengan kehadirannya senantiasa memberikan warna, semangat, dan pengingat bagi penulis untuk terus melangkah maju.
4. Kepada seseorang yang telah kebersamaan penulis sejak semester kedua hingga saat ini, sosok yang tetap setia berdiri di sisi penulis dalam suka maupun duka. Terima kasih telah menjadi pendengar terbaik, pemberi semangat paling tulus, dan bagian terindah dalam perjalanan pendewasaan ini.
5. Kepada sahabat saya, Nanda Annisa dan Sabrina Febriani, atas kehangatan persahabatan, dukungan tanpa pamrih, dan tawa yang meringankan beban selama masa-masa perkuliahan yang penuh perjuangan.
6. Kepada diri sendiri, terima kasih telah bertahan, telah berjuang dengan hebat, dan tidak memilih untuk menyerah meskipun jalan yang dilalui tidaklah mudah.

KATA PENGANTAR

Puji syukur penulis panjatkan ke hadirat Tuhan Yang Maha Esa, karena atas rahmat dan karunia-Nya laporan non-reguler ini dapat diselesaikan. Laporan ini disusun sebagai salah satu bentuk pemenuhan tugas akademik dengan judul “Analysis of Gradient Boosted Trees Algorithm in Breast Cancer Classification”.

Dalam proses penyusunan laporan ini, penulis memperoleh bantuan dari berbagai pihak. Oleh karena itu, penulis menyampaikan terima kasih kepada:

1. Majid Rahardi, S.Kom., M.Eng. selaku Dosen Pembimbing yang telah memberikan bimbingan, arahan, serta masukan selama penyusunan laporan.
2. Tim Dosen Penguji yang telah memberikan kritik, saran, dan evaluasi yang membangun untuk penyempurnaan laporan.
3. Segenap Dosen dan Staf Akademik yang telah membekali penulis dengan ilmu pengetahuan dan fasilitas selama masa studi.

Penulis menyadari bahwa laporan ini masih memiliki banyak kekurangan dan keterbatasan. Oleh karena itu, penulis mengharapkan kritik dan saran yang membangun demi perbaikan di masa mendatang. Semoga laporan ini dapat memberikan manfaat bagi pembaca dan pihak-pihak yang berkepentingan.

Yogyakarta, 27 Januari 2026

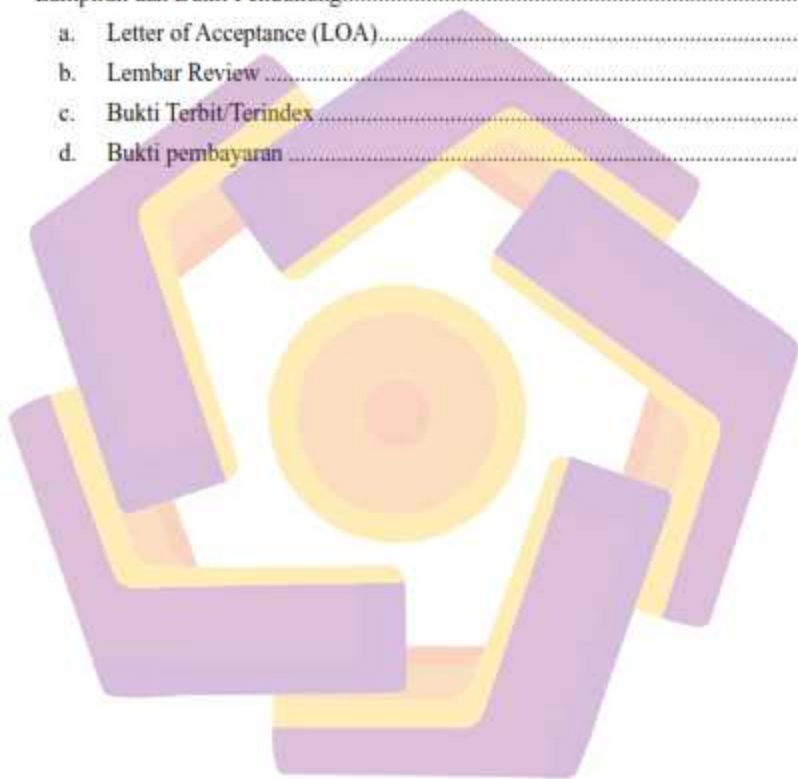
Penulis

DAFTAR ISI

Halaman Judul.....	i
Halaman Persetujuan.....	ii
Halaman Pengesahan.....	iii
Halaman Pernyataan Keaslian Karya.....	iv
Halaman Persembahan.....	v
Kata Pengantar.....	vi
Daftar Isi.....	vii
Daftar Tabel.....	x
Daftar Gambar.....	xi
Daftar Lampiran.....	xii
Daftar Lambang dan Singkatan.....	xiii
Daftar Istilah.....	xiv
Intisari.....	xv
<i>Abstract</i>	xvi
Bab I Pendahuluan.....	1
1.1. Gambaran Umum.....	1
1.2. Rumusan Masalah.....	4
1.3. Batasan Masalah.....	5
1.4. Tujuan.....	5
Bab II Tinjauan Pustaka.....	7
2.1. Studi Literatur.....	7
2.2. Landasan Teori.....	10
2.2.1. Kanker Payudara.....	10
2.2.2. Supervised Learning.....	10
2.2.3. Machine Learning.....	11
2.2.4. Gradient Boosted Trees.....	11
2.2.5. Exploratory Data Analysis (EDA).....	12
2.2.6. Split Data.....	12
2.2.7. Preprocessing.....	12
2.2.8. Standardisasi Fitur.....	12

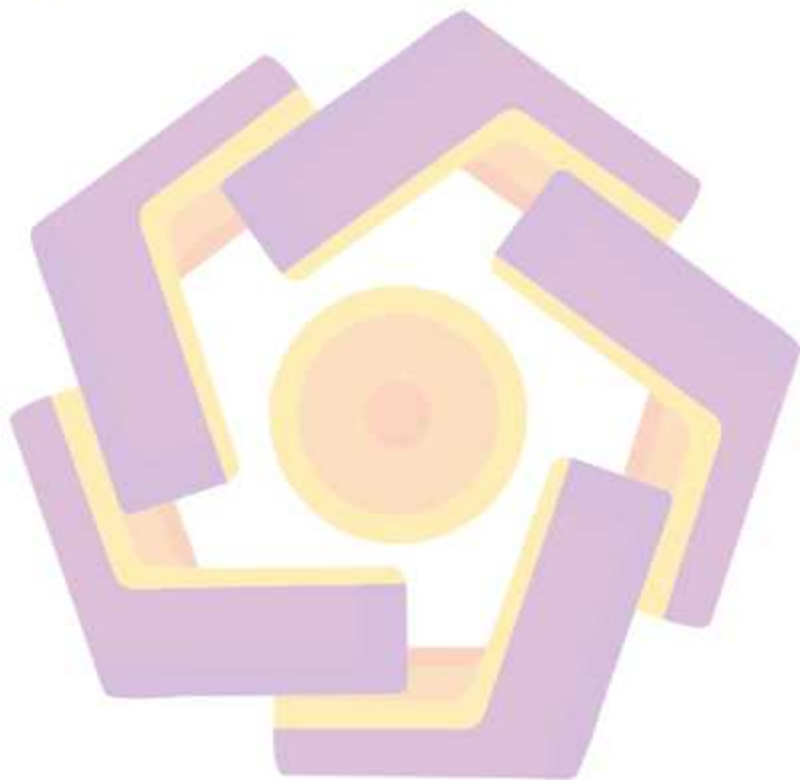
2.2.9. Local Outlier Factor (LOF)	13
2.2.10. SMOTE	13
2.2.11. Seleksi Fitur ANOVA (SelectKBest).....	14
2.2.12. Evaluasi Model.....	14
2.2.13. DeLong Test	14
2.2.14. Interpretabilitas Model menggunakan SHAP	14
BAB III Metode Penelitian	16
3.1. Objek Penelitian.....	16
3.2. Metode Penelitian.....	16
3.2.1. Akuisisi Data	17
3.2.2. Pra-pemrosesan Data.....	18
3.2.3. Exploratory Data Analysis (EDA)	19
3.2.4. Split Data.....	21
3.2.5. Preprocessing	22
3.2.6. Pemodelan.....	25
3.2.7. Evaluasi Model.....	26
3.2.8. Uji Signifikan (DeLong Test).....	27
3.2.9. Interpretabilitas Model.....	30
BAB IV Pembahasan	33
4.1. Exploratory Data Analysis (EDA)	33
4.2. Split Data.....	36
4.3. Preprocessing	37
4.4. Pemodelan.....	39
4.5. Evaluasi Model.....	40
4.5.1. Evaluasi CatBoost	40
4.5.2. Evaluasi HistGradientBoosting.....	43
4.5.3. Evaluasi LightGBM	45
4.5.4. Evaluasi GradientBoosting.....	47
4.5.5. Evaluasi XGBoost.....	49
4.6. Uji Signifikan Statistik (DeLong Test).....	52
4.7. Interpretabilitas Model.....	53
4.8. Deployment.....	55

BAB V Kesimpulan	57
5.1. Kesimpulan	57
5.2. Saran.....	57
Referensi	59
Curriculum Vitae	64
Lampiran dan Bukti Pendukung.....	66
a. Letter of Acceptance (LOA).....	66
b. Lembar Review	67
c. Bukti Terbit/Terindex	71
d. Bukti pembayaran	72



DAFTAR TABEL

Tabel 4. 1 Contoh Statistik Deskriptif 10 Fitur	34
Tabel 4. 2 Hasil Split Data	37
Tabel 4. 3 Hasil Peringkat Fitur (10 Fitur).....	38
Tabel 4. 4 Tabel Evaluasi Model.....	40

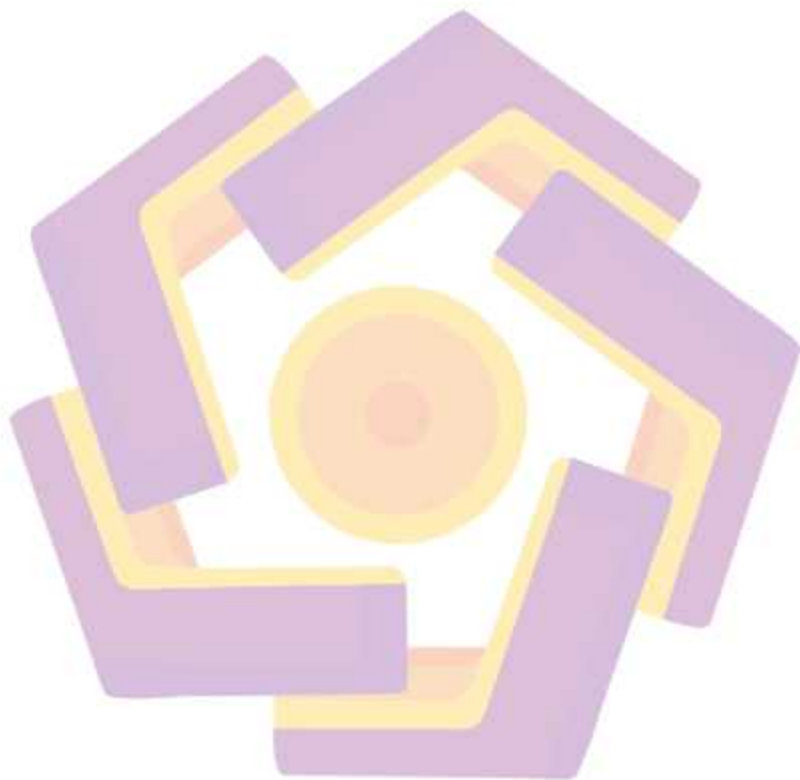


DAFTAR GAMBAR

Gambar 3. 1 Alur Penelitian.....	17
Gambar 4. 1 Analisis Distribusi Kelas	33
Gambar 4. 2 Visualisasi Distribusi Fitur	34
Gambar 4. 3 Heatmap Korelasi.....	35
Gambar 4. 4 Analisis Outlier Awal 5 Fitur.....	36
Gambar 4. 5 Distribusi Kelas Sebelum dan Sesudah SMOTE	38
Gambar 4. 6 Classification Report CatBoost.....	41
Gambar 4. 7 Confusion Matrix CatBoost	41
Gambar 4. 8 ROC Curve CatBoost.....	42
Gambar 4. 9 Classification Report HistGradientBoosting.....	43
Gambar 4. 10 Confusion Matrix HistGradientBoosting.....	44
Gambar 4. 11 ROC Curve HistGradientBoosting.....	45
Gambar 4. 12 Classification Report LightGBM.....	45
Gambar 4. 13 Confusion Matrix LightGBM	46
Gambar 4. 14 ROC Curve LightGBM.....	47
Gambar 4. 15 Classification Report GradientBoosting.....	48
Gambar 4. 16 Confusion Matrix GradientBoosting.....	48
Gambar 4. 17 ROC Curve GradientBoosting	49
Gambar 4. 18 Classification Report XGBoost.....	50
Gambar 4. 19 Confusion Matrix XGBoost	50
Gambar 4. 20 ROC Curve XGBoost.....	51
Gambar 4. 21 Hasil DeLong Test.....	52
Gambar 4. 22 SHAP Feature.....	54
Gambar 4. 23 SHAP Value.....	55

DAFTAR LAMPIRAN

Lampiran 1 Bukti LoA.....	66
Lampiran 2 Bukti Review	71
Lampiran 3 Bukti Terindeks.....	71
Lampiran 4 Bukti Pembayaran.....	72



DAFTAR LAMBANG DAN SINGKATAN



AI	<i>Artificial Intelligence (Kecerdasan Buatan)</i>
ML	<i>Machine Learning</i>
GBT	<i>Gradient Boosted Trees</i>
XGBoost	<i>eXtreme Gradient Boosting</i>
LightGBM	<i>Light Gradient Boosting Machine</i>
CatBoost	<i>Category Boosting</i>
HistGradientBoosting	<i>Histogram-based Gradient Boosting</i>
GradientBoosting	<i>Gradient Boosting Classifier</i>
EDA	<i>Exploratory Data Analysis</i>
LOF	<i>Local Outlier Factor</i>
SMOTE	<i>Synthetic Minority Over-sampling Technique</i>
ANOVA	<i>Analysis of Variance</i>
SelectKBest	<i>Metode seleksi fitur (memilih K fitur terbaik)</i>
ROC	<i>Receiver Operating Characteristic</i>
AUC	<i>Area Under the Curve</i>
SHAP	<i>Shapley Additive Explanations</i>
XAI	<i>Explainable Artificial Intelligence</i>
TP	<i>True Positive</i>
TN	<i>True Negative</i>
FP	<i>False Positive</i>
FN	<i>False Negative</i>
TPR	<i>True Positive Rate (Sensitivitas/Recall Positif)</i>
FPR	<i>False Positive Rate</i>
p-value	<i>Nilai probabilitas pada uji statistik</i>

DAFTAR ISTILAH

Benign	Tumor tidak ganas yang pertumbuhannya relatif tidak invasif.
Malignant	Tumor ganas yang berpotensi invasif dan dapat menyebar.
Klasifikasi	Proses memetakan data ke kelas tertentu.
Preprocessing	Tahap penyiapan data sebelum pelatihan model.
Standardisasi Fitur	Transformasi fitur agar memiliki rata-rata 0 dan standar deviasi 1.
Outlier	Data yang menyimpang jauh dari pola mayoritas dan berpotensi mengganggu pelatihan model.
Imbalance Kelas	Kondisi jumlah sampel tiap kelas tidak sebanding sehingga model cenderung bias ke kelas mayoritas.
Oversampling	Penambahan data buatan pada kelas minoritas untuk menyeimbangkan distribusi kelas.
Seleksi Fitur	Pemilihan fitur paling relevan.
Confusion Matrix	Tabel ringkas prediksi benar atau salah berdasarkan TP, TN, FP, dan FN.
ROC Curve	Kurva yang menunjukkan trade-off TPR dan FPR pada berbagai ambang keputusan.
DeLong Test	Uji statistic untuk membandingkan ROC-AUC dua model pada data uji yang sama.

INTISARI

Kanker payudara masih menjadi masalah kesehatan yang sering ditemukan, dan keterlambatan deteksi dini membuat banyak kasus teridentifikasi pada tahap lanjut sehingga meningkatkan risiko kematian, menurunkan kualitas hidup pasien, serta menambah beban layanan kesehatan (waktu diagnosis, kebutuhan terapi, dan biaya). Untuk membantu klasifikasi kanker secara lebih cepat dan konsisten, penelitian ini memanfaatkan Breast Cancer Dataset dari Kaggle yang berisi 569 data, terdiri dari 357 kasus kanker jinak (benign) dan 212 kasus kanker ganas (malignant). Data dibersihkan dengan penyeragaman nama atribut, penghapusan kolom tidak relevan, dan pengkodean label, lalu dibagi 80:20 untuk pelatihan dan pengujian. Selanjutnya dibangun alur pemrosesan yang mencakup standardisasi fitur, deteksi dan penghapusan outlier dengan Local Outlier Factor, penyeimbangan kelas menggunakan SMOTE, serta seleksi fitur ANOVA (SelectKBest). Lima model Gradient Boosted Trees (XGBoost, LightGBM, CatBoost, HistGradientBoosting, dan GradientBoosting) dilatih dan dievaluasi menggunakan akurasi, presisi, recall, F1-score, dan ROC-AUC, serta uji signifikansi DeLong untuk membandingkan AUC. Hasil menunjukkan performa seluruh model tinggi dan relatif sebanding, CatBoost memperoleh ROC-AUC 0,9960 tertinggi dengan akurasi 0,9649, namun perbedaannya tidak signifikan secara statistik ($p > 0.05$). Interpretabilitas SHAP menyoroti fitur ukuran dan ketidakaturan inti sel sebagai penentu utama prediksi. Temuan ini bermanfaat bagi peneliti dan praktisi kesehatan sebagai dasar pengembangan sistem pendukung keputusan yang akurat dan transparan, dengan rekomendasi penelitian lanjutan melakukan validasi eksternal pada data klinis dunia nyata sebelum penerapan klinis.

Kata kunci: Breast Cancer Classification, CatBoost, Gradient Boosted Trees, Machine Learning, SHAP Explainability.

ABSTRACT

Breast cancer remains a common health problem, and delays in early detection mean that many cases are identified at an advanced stage, increasing the risk of death, reducing patients' quality of life, and adding to the burden on healthcare services (diagnosis time, therapy requirements, and costs). To help classify cancer more quickly and consistently, this study utilizes the Breast Cancer Dataset from Kaggle, which contains 569 data points, consisting of 357 benign cancer cases and 212 malignant cancer cases. The data was cleaned by standardizing attribute names, removing irrelevant columns, and encoding labels, then divided 80:20 for training and testing. Next, a processing flow was built that included feature standardization, outlier detection and removal using Local Outlier Factor, class balancing using SMOTE, and ANOVA feature selection (SelectKBest). Five Gradient Boosted Trees models (XGBoost, LightGBM, CatBoost, HistGradientBoosting, and GradientBoosting) were trained and evaluated using accuracy, precision, recall, F1-score, and ROC-AUC, as well as DeLong's significance test to compare AUC. The results show that all models perform well and are relatively comparable, with CatBoost obtaining the highest ROC-AUC of 0.9960 with an accuracy of 0.9649, but the difference is not statistically significant ($p > 0.05$). SHAP interpretability highlights cell size and nuclear irregularity as the main determinants of prediction. These findings are useful for researchers and healthcare practitioners as a basis for developing accurate and transparent decision support systems, with recommendations for further research to conduct external validation on real-world clinical data before clinical application.

Keyword: Breast Cancer Classification, CatBoost, Gradient Boosted Trees, Machine Learning, SHAP Explainability.