KLASIFIKASI STADIUM KANKER PAYUDARA BERBASIS DATA EKSPRESI GEN MENGGUNAKAN ALGORITMA STACKING ENSEMBEL LEARNING

SKRIPSI

PRA YOGA WIJAYA
00000061956

PROGRAM STUDI INFORMATIKA
FAKULTAS TEKNIK DAN INFORMATIKA
UNIVERSITAS MULTIMEDIA NUSANTARA
TANGERANG
2025

KLASIFIKASI STADIUM KANKER PAYUDARA BERBASIS DATA EKSPRESI GEN MENGGUNAKAN ALGORITMA STACKING ENSEMBEL LEARNING

SKRIPSI

Diajukan sebagai salah satu syarat untuk memperoleh
Gelar Sarjana Komputer (S.Kom.)

PRA YOGA WIJAYA
00000061956

PROGRAM STUDI INFORMATIKA
FAKULTAS TEKNIK DAN INFORMATIKA
UNIVERSITAS MULTIMEDIA NUSANTARA
TANGERANG
2025

HALAMAN PERNYATAAN TIDAK PLAGIAT

Dengan ini saya,

Nama	:	Pra Yoga Wijaya
Nomor Induk Mahasiswa	:	00000061956
Program Studi	:	Informatika

Skripsi dengan judul:

KLASIFIKASI STADIUM KANKER PAYUDARA BERBASIS DATA EKSPRESI GEN MENGGUNAKAN ALGORITMA STACKING ENSEMBEL LEARNING

merupakan hasil karya saya sendiri bukan plagiat dari laporan karya tulis ilmiah yang ditulis oleh orang lain, dan semua sumber, baik yang dikutip maupun dirujuk, telah saya nyatakan dengan benar serta dicantumkan di Daftar Pustaka.

Jika di kemudian hari terbukti ditemukan kecurangan/penyimpangan, baik dalam pelaksanaan maupun dalam penulisan laporan karya tulis ilmiah, saya bersedia menerima konsekuensi dinyatakan TIDAK LULUS untuk mata kuliah yang telah saya tempuh.

Tangerang, 26 Juni 2025

(Pra Yoga Wijaya)

HALAMAN PENGESAHAN

Skripsi dengan judul

KLASIFIKASI STADIUM KANKER PAYUDARA BERBASIS DATA EKSPRESI GEN MENGGUNAKAN ALGORITMA STACKING ENSEMBEL LEARNING

oleh

Nama	:	Pra Yoga Wijaya
NIM	:	00000061956
Program Studi	:	Informatika
Fakultas	:	Fakultas Teknik dan Informatika

Telah diujikan pada hari Jumat, 18 Juli 2025
Pukul 13.00 s/d 15.00 dan dinyatakan
LULUS
Dengan susunan penguji sebagai berikut

Ketua Sidang

(Arya Wicaksana, S.Kom., M.Eng.Sc., OCA)
NIDN: 0315109103

Penguji

(Dennis Gunawan, S.Kom., M.Sc.)
NIDN: 0320059001

Pembimbing I

(Moeljono Widjaja, B.Sc., M.Sc., Ph.D)
NIDN: 0311106903

Pembimbing II

(David Agustriawan, S.Kom., M.Sc., Ph.D.)
NIDN: 0525088601

Ketua Program Studi Informatika,

(Arya Wicaksana, S.Kom., M.Eng.Sc., OCA)
NIDN: 0315109103

HALAMAN PERSETUJUAN PUBLIKASI KARYA ILMIAH UNTUK KEPENTINGAN AKADEMIS

Yang bertanda tangan di bawah ini:

Nama	:	Pra Yoga Wijaya
NIM	:	00000061956
Program Studi	:	Informatika
Jenjang	:	S1
Judul Karya Ilmiah	:	KLASIFIKASI STADIUM KANKER PAYUDARA BERBASIS DATA EKSPRESI GEN MENGGUNAKAN ALGORITMA STACKING ENSEMBEL LEARNING

Menyatakan dengan sesungguhnya bahwa saya bersedia (**pilih salah satu**):

☑ Saya bersedia memberikan izin sepenuhnya kepada Universitas Multimedia Nusantara untuk mempublikasikan hasil karya ilmiah saya ke dalam repositori Knowledge Center sehingga dapat diakses oleh Sivitas Akademika UMN/Publik. Saya menyatakan bahwa karya ilmiah yang saya buat tidak mengandung data yang bersifat konfidensial.

☐ Saya tidak bersedia mempublikasikan hasil karya ilmiah ini ke dalam repositori Knowledge Center, dikarenakan: dalam proses pengajuan publikasi ke jurnal/konferensi nasional/internasional (dibuktikan dengan *letter of acceptance*) **.

☐ Lainnya, pilih salah satu:

- Hanya dapat diakses secara internal Universitas Multimedia Nusantara
- Embargo publikasi karya ilmiah dalam kurun waktu tiga tahun.

Tangerang, 26 Juni 2025

Yang menyatakan

Pra Yoga Wijaya

**Jika tidak bisa membuktikan LoA jurnal/HKI, saya bersedia mengizinkan penuh karya ilmiah saya untuk dipublikasikan ke KC UMN dan menjadi hak institusi UMN.

HALAMAN PERSEMBAHAN / MOTTO

”It always seems impossible until it’s done.”

Nelson Mandela

KATA PENGANTAR

Puji Syukur atas berkat dan rahmat kepada Tuhan Yang Maha Esa, atas selesainya penulisan skripsi ini dengan judul: Klasifikasi Stadium Kanker Payudara Berbasis Data Ekspresi Gen Menggunakan Algoritma Stacking Ensembel Learning dilakukan untuk memenuhi salah satu syarat untuk mencapai gelar Sarjana Komputer Jurusan Informatika Pada Fakultas Teknik dan Informatika Universitas Multimedia Nusantara. Saya menyadari bahwa, tanpa bantuan dan bimbingan dari berbagai pihak, dari masa perkuliahan sampai pada penyusunan skripsi ini, sangatlah sulit bagi saya untuk menyelesaikan skripsi ini. Oleh karena itu, saya mengucapkan terima kasih kepad

Mengucapkan terima kasih

1. Bapak Dr. Ir. Andrey Andoko, M.Sc., selaku Rektor Universitas Multimedia Nusantara.
2. Bapak Dr. Eng. Niki Prastomo, S.T., M.Sc., selaku Dekan Fakultas Teknik dan Informatika Universitas Multimedia Nusantara.
3. Bapak Arya Wicaksana, S.Kom., M.Eng.Sc., OCA, selaku Ketua Program Studi Informatika Universitas Multimedia Nusantara.
4. Bapak Moeljono Widjaja, B.Sc., M.Sc., Ph.D, sebagai Pembimbing pertama yang telah memberikan bimbingan, arahan, dan motivasi atas terselesainya tugas akhir ini.
5. Bapak David Agustriawan, S.Kom., M.Sc., Ph.D., selaku Pembimbing Kedua yang telah memberikan masukan dan dukungan dalam penyusunan skripsi ini.
6. Keluarga saya yang telah memberikan bantuan dukungan material dan moral, sehingga penulis dapat menyelesaikan tugas akhir ini.

Semoga karya ilmiah ini dapat berkontribusi dalam pengembangan ilmu pengetahuan dan mendorong riset lanjutan di masa depan.

Tangerang, 26 Juni 2025

Pra Yoga Wijaya

KLASIFIKASI STADIUM KANKER PAYUDARA BERBASIS DATA EKSPRESI GEN MENGGUNAKAN ALGORITMA STACKING ENSEMBEL LEARNING

Pra Yoga Wijaya

ABSTRAK

Kanker payudara merupakan salah satu penyebab utama kematian pada wanita. Deteksi dini terhadap stadium kanker payudara sangat diperlukan sebagai tindakan untuk meningkatkan angka keberlangsungan hidup pasien. Penelitian ini bertujuan untuk membangun model klasifikasi stadium kanker payudara (*early stage* dan *late stage*) berbasis *machine learning* dengan menggunakan data ekspresi gen (RNA-seq) dari *The Cancer Genome Atlas Breast cancer* (TCGA-BRCA). Pendekatan yang dilakukan melibatkan seleksi fitur bioinformatika dan statistika, menggunakan *Differentially Expressed Genes* (DEG) dengan paket *limma* sebagai seleksi fitur bioinformatika dan menggunakan *Logistic Regression*, serta *Analysis of Variance* (ANOVA) sebagai pendekatan seleksi fitur statistika. Model klasifikasi dibangun dengan algoritma *Stacking Ensembel Learning* dengan *Random Forest*, *XGBoost*, *Support Vector Machine* (SVM), dan *Logistic Regression* sebagai *base learners*. Model terbaik pada penelitian ini mampu mencapai accuracy 93,47%, precision 93,52%, recall 93,47%, dan f1-score 93,49%. Berdasarkan model terbaik, terdapat 48 gen kandidat potensial biomarker, 7 gen diantaranya telah didukung oleh literatur terdahulu terkait kanker payudara. Pengujian biomarker juga dilakukan dengan menghitung nilai ROC-AUC individual gen dan didapatkan bahwa nilai tertinggi hanya mencapai 0.638, menunjukan bahwa kontribusi setiap gen dapat menghasilkan model yang lebih kuat dibandingkan secara individual. Penelitian ini mendukung potensi pendekatan *machine learning* dalam klasifikasi stadium kanker payudara pada data ekspresi gen RNA-seq secara efektif.

Kata kunci: Biomarker, Kanker Payudara, *Machine Learning*, RNA-seq, Seleksi Fitur

BREAST CANCER STAGES CLASSIFICATION USING GENE EXPRESSION DATA WITH STACKING ENSEMBEL LEARNING ALGORITHM

Pra Yoga Wijaya

ABSTRACT

Breast cancer is one of the leading causes of death among women. Early detection of breast cancer stages is crucial to improve patient survival rates. This study aims to develop a machine learning-based classification model for breast cancer stages (early stage and late stage) using gene expression data (RNA-seq) from The Cancer Genome Atlas Breast Cancer (TCGA-BRCA). The approach involves bioinformatics and statistical feature selection, using Differentially Expressed Genes (DEG) with the limma package for bioinformatics-based selection, and Logistic Regression as well as Analysis of Variance (ANOVA) for statistical feature selection. The classification model was built using the Stacking Ensemble Learning algorithm, with Random Forest, XGBoost, Support Vector Machine (SVM), and Logistic Regression as base learners. The best model achieved an accuracy of 93.47%, precision of 93.52%, recall of 93.47%, and an F1-score of 93.49%. Based on the best-performing model, 48 genes were identified as potential biomarker candidates, 7 of which are supported by previous breast cancer studies. Biomarker validation was also conducted by calculating the individual gene ROC-AUC scores, with the highest value reaching only 0.638, indicating that the combination of genes provides stronger predictive power than individual genes. This study supports the potential of machine learning approaches for effective classification of breast cancer stages using RNA-seq gene expression data.

Keywords: *Biomarker*, *Breast Cancer*, *Feature Selection*, *Machine Learning*, *RNA-seq*

DAFTAR ISI

HALAMAN JUDUL i
PERNYATAAN TIDAK MELAKUKAN PLAGIAT ii
HALAMAN PENGESAHAN iii
HALAMAN PERSETUJUAN PUBLIKASI KARYA ILMIAH iv
HALAMAN PERSEMBAHAN/MOTO v
KATA PENGANTAR vi
ABSTRAK vii
ABSTRACT viii
DAFTAR ISI ix
DAFTAR TABEL xi
DAFTAR GAMBAR xii
DAFTAR KODE xiii
DAFTAR RUMUS xiv
DAFTAR LAMPIRAN xv
BAB 1 PENDAHULUAN 1
1.1 Latar Belakang Masalah 1
1.2 Rumusan Masalah 5
1.3 Batasan Permasalahan 5
1.4 Tujuan Penelitian 6
1.5 Manfaat Penelitian 6
1.5.1 Manfaat Teoritis 6
1.5.2 Manfaat Praktis 7
1.6 Sistematika Penulisan 7
BAB 2 LANDASAN TEORI 9
2.1 Kanker Payudara (Breast Cancer) 9
2.2 Ekspresi Gen RNA-seq 11
2.3 Differentially Expressed Genes (DEG) 11
2.4 Feature Selection 13
2.4.1 Analysis of Variance (ANOVA) 14
2.4.2 Recursive Feature Elimination (RFE) 15
2.5 Logistic Regression (LR) 15
2.6 Random Forest 19
2.7 Extreme Gradient Boosting (XGBoost) 20
2.8 Support Vector Machine (SVM) 23
2.9 Stacking Ensembel Learning 28
2.10 Metrik Evaluasi (Evaluation Metrix) 29
2.10.1 Confusion Matrix 29
2.10.2 Akurasi (Accuracy) 30
2.10.3 Presisi (Precision) 30
2.10.4 Recall 31
2.10.5 F1-score 31
2.10.6 ROC (Receiver Operating Characteristic) Curve 31
2.10.7 Area Under the Curve (AUC) 32
BAB 3 METODOLOGI PENELITIAN 33
3.1 Alur Kerja Penelitian 33
3.2 Perangkat Penunjang 34
3.3 Pengumpulan Data 34
3.4 Pra-pemrosesan Data (Data Preprocessing) 35

3.4.1 Memuat Dataset . . . 36
3.4.2 Gabungkan dataset . . . 36
3.4.3 Menghapus Nilai Kosong (NaN Value) . . . 36
3.4.4 Filter dataset berdasarkan Ras Kulit Putih (White Race) . . . 36
3.4.5 Filter dataset berdasarkan stadium . . . 37
3.4.6 Labeling Dataset . . . 37
3.5 Seleksi Fitur . . . 37
3.6 Pembangunan Model . . . 39
3.7 Evaluasi Model . . . 39
3.8 Skenario Percobaan . . . 40
3.9 Analisis Biomarker . . . 40
BAB 4 HASIL DAN DISKUSI . . . 41
4.1 Pengumpulan Data . . . 41
4.2 Pra-pemrosesan data . . . 41
4.2.1 Penggabungan Dataset . . . 42
4.2.2 Penanganan Missing Values . . . 43
4.2.3 Filter dataset berdasarkan ras . . . 44
4.2.4 Filter Berdasarkan Stadium & Labeling Dataset . . . 44
4.3 Seleksi Fitur . . . 46
4.3.1 Seleksi Fitur Dengan Pendekatan DEG . . . 46
4.3.2 Seleksi Fitur Dengan Pendekatan Informatika . . . 48
4.4 Pembangunan dan Evaluasi Model . . . 52
4.5 Hasil Skenario . . . 56
4.5.1 Hasil Klasifikasi dengan metode seleksi fitur DEG dengan paket Limma . . . 57
4.5.2 Hasil Klasifikasi dengan metode seleksi fitur Logistic Regression 2000 dan RFE . . . 58
4.5.3 Hasil Klasifikasi dengan metode seleksi fitur Logistic Regression 4000 dan RFE . . . 60
4.5.4 Hasil Klasifikasi dengan metode seleksi fitur ANOVA 2000 dan RFE . . . 62
4.5.5 Hasil Klasifikasi dengan metode seleksi fitur ANOVA 4000 dan RFE . . . 64
4.6 Perbandingan Hasil Terbaik Pada Setiap Skenario . . . 66
4.7 Analisis Potensial Kandidat Biomarker . . . 66
BAB 5 SIMPULAN DAN SARAN . . . 73
5.1 Simpulan . . . 73
5.2 Saran . . . 74
DAFTAR PUSTAKA . . . 75

DAFTAR TABEL

Tabel 2.1	Stage Groupings Based on the TNM classification of the Breast Cancer	10
Tabel 2.2	Fungsi-fungsi Kernel yang Umum Digunakan dalam SVM	27
Tabel 2.3	Confusion matrix	30
Tabel 2.4	Interpretasi Area Under the Curve (AUC)	32
Tabel 3.1	Kombinasi seleksi fitur yang dilakukan	39
Tabel 3.2	Skenario percobaan	40
Tabel 4.1	Rincian dataset yang digunakan dalam penelitian	41
Tabel 4.2	Distribusi ras dalam dataset	44
Tabel 4.3	Distribusi stadium kanker payudara dalam dataset	45
Tabel 4.4	Hasil Analisis Ekspresi Gen Diferensial (DEG) dengan limma	47
Tabel 4.5	Top 5 Klasifikasi dengan metode seleksi fitur Limma	58
Tabel 4.6	Top 5 klasifikasi dengan metode seleksi fitur Logistic Regression 2000 + RFE	59
Tabel 4.7	Top 5 klasifikasi dengan metode seleksi fitur Logistic Regression 4000 + RFE	61
Tabel 4.8	Top 5 klasifikasi dengan metode seleksi fitur ANOVA 2000 + RFE	63
Tabel 4.9	Top 5 klasifikasi dengan metode seleksi fitur ANOVA 4000 + RFE	65
Tabel 4.10	Daftar ensembel ID dan simbol gen	66

DAFTAR GAMBAR

Gambar 2.1 Ilustrasi hyperplnae optimal yang memaksimalkan margin dalam dua ruang dimensi 24
Gambar 2.2 Arsitektur Stacking Classifier 28
Gambar 3.1 Alur Penelitian 33
Gambar 3.2 Rangkaian proses pra-proses data 35
Gambar 3.3 Rangkaian proses seleksi fitur 38
Gambar 4.1 Hasil penghapusan *missing value* 43
Gambar 4.2 Plot akurasi model Stacking Classifier dengan metode seleksi fitur Limma 57
Gambar 4.3 Plot akurasi model Stacking Classifier dengan metode seleksi fitur Logistic Regression 2000 + RFE 59
Gambar 4.4 Plot akurasi model Stacking Classifier dengan metode seleksi fitur Logistic Regression 4000 + RFE 61
Gambar 4.5 Plot akurasi model Stacking Classifier dengan metode seleksi fitur ANOVA 2000 + RFE 63
Gambar 4.6 Plot akurasi model Stacking Classifier dengan metode seleksi fitur ANOVA 4000 + RFE 64
Gambar 4.7 Plot ROC-AUC 48 gen 71

DAFTAR KODE

Kode 4.1 Pemuatan dataset 41
Kode 4.2 Penggabungan dataset 42
Kode 4.3 Penanganan *missing values* 43
Kode 4.4 Filter dataset berdasarkan *white race* 44
Kode 4.5 Filter dataset berdasarkan stadium dan labeling dataset 45
Kode 4.6 Analisis DEG menggunakan paket limma 46
Kode 4.7 Normalisasi dataset 48
Kode 4.8 Fungsi seleksi fitur ANOVA 49
Kode 4.9 Fungsi seleksi fitur *logistic regression* 50
Kode 4.10 Fungsi seleksi fitur RFE 51
Kode 4.11 Fungsi klasifikasi *stacking classifier* 52
Kode 4.12 iterasi klasifikasi *stacking classifier* 55

DAFTAR RUMUS

Rumus 2.1 *Between-Class Sum of Squares (SSB)* . . . 14
Rumus 2.2 *Within-Class Sum of Squares (SSW)* . . . 14
Rumus 2.3 *F-Statistic (F-value)* . . . 14
Rumus 2.4 *Linear Combination (Logit)* . . . 16
Rumus 2.5 *Sigmoid Function (Probability Output)* . . . 17
Rumus 2.6 *Fungsi Prediksi (Keputusan Kelas)* . . . 17
Rumus 2.7 *Fungsi Loss: Cross-Entropy* . . . 17
Rumus 2.8 *Total Loss dengan L1 Regularization* . . . 18
Rumus 2.9 *Turunan Fungsi Loss* . . . 18
Rumus 2.10 *Gradien untuk Pembaruan Bobot (dengan L1)* . . . 18
Rumus 2.11 *Random Forest Prediction* . . . 19
Rumus 2.12 *Gini Index* . . . 20
Rumus 2.13 *Gini Index In Binary classification* . . . 20
Rumus 2.14 Prediksi Model pada Iterasi ke-p . . . 22
Rumus 2.15 Fungsi Objektif XGBoost . . . 22
Rumus 2.16 Fungsi Regularisasi Tree . . . 22
Rumus 2.17 Ekspansi Taylor Orde Kedua . . . 23
Rumus 2.18 Skor Gain pada Pemisahan Node . . . 23
Rumus 2.19 Fungsi Aktivasi Sigmoid untuk Klasifikasi Biner . . . 23
Rumus 2.20 Persamaan Hyperplane SVM . . . 24
Rumus 2.22 Fungsi Objektif SVM - Hard Margin . . . 25
Rumus 2.23 Kendala Klasifikasi Benar . . . 25
Rumus 2.24 Fungsi Objektif Soft-Margin SVM . . . 25
Rumus 2.25 Kendala Soft-Margin SVM . . . 25
Rumus 2.26 Fungsi Lagrangian pada SVM . . . 26
Rumus 2.27 Kondisi dari Turunan terhadap w dan b . . . 26
Rumus 2.28 Fungsi Objektif pada Dual Problem (Formulasi Dual) . . . 26
Rumus 2.29 Kendala dalam Optimasi Dual Soft-Margin . . . 27
Rumus 2.30 Definisi Fungsi Kernel . . . 27
Rumus 2.31 Fungsi Klasifikasi dengan Kernel . . . 27
Rumus 2.32 *Accuracy* . . . 30
Rumus 2.33 *Precision* . . . 30
Rumus 2.34 *Recall* . . . 31
Rumus 2.35 *F1 Score* . . . 31

DAFTAR LAMPIRAN

Lampiran 1 Hasil Persentase Turnitin 80
Lampiran 2 Formulir Bimbingan - Moeljono Widjaja, B.Sc., M.Sc., Ph.D . 94
Lampiran 3 Formulir Bimbingan - David Agustriawan, S.Kom., M.Sc., Ph.D. 97