



**PENGARUH *OVERSAMPLING* MULTI-VARIAN SMOTE DAN
ALGORITMA GENETIKA TERHADAP PERFORMA PREDIKSI
CACAT PERANGKAT LUNAK BERBASIS *MOBILE***

Skripsi

**Untuk memenuhi persyaratan
Dalam menyelesaikan Strata-1 Ilmu Komputer**

Oleh

ADE RAHMAD

NIM 2211016110002

**PROGRAM STUDI S-1 ILMU KOMPUTER
FAKULTAS MATEMATIKA DAN ILMU PENGETAHUAN ALAM
UNIVERSITAS LAMBUNG MANGKURAT
BANJARBARU**

JULI 2026

SKRIPSI

PENGARUH *OVERSAMPLING* MULTI-VARIAN SMOTE DAN ALGORITMA GENETIKA TERHADAP PERFORMA PREDIKSI CACAT PERANGKAT LUNAK BERBASIS *MOBILE*

Oleh:

ADE RAHMAD
2211016110002

Telah dipertahankan di depan Dosen Penguji pada tanggal 17 Juni 2026.
Susunan Penguji:

Pembimbing Utama



Rudy Herteno, S.Kom., M.Kom.
NIP. 198809252022031003

Penguji

1.



Triando Hamonangan Saragih, S.Kom., M.Kom.
NIP. 199308242019031012

Pembimbing Pendamping



Irwan Budiman, S.T., M.Kom.
NIP. 197703252008121001

2.



Dodon Turlanto Nugrahadi, S.Kom, M.Eng.
NIP. 198001122009121002

Banjarbaru, 7 Juli 2026

Koordinator Program Studi Ilmu Komputer



Dwi Kartini, S.Kom., M.Kom.
NIP. 198704212012122003

PERNYATAAN

Dengan ini saya menyatakan bahwa dalam skripsi ini tidak terdapat karya yang pernah diajukan untuk memperoleh gelar kesarjanaan di suatu Perguruan Tinggi, dan sepanjang pengetahuan saya juga tidak terdapat karya atau pendapat yang pernah ditulis atau diterbitkan oleh orang lain, kecuali yang secara tertulis diacu dalam naskah ini dan disebutkan dalam daftar pustaka.

Banjarbaru, 17 Juni 2026
Yang Menyatakan,



Ade Rahmad
NIM. 2211016110002

ABSTRAK

PENGARUH *OVERSAMPLING* MULTI-VARIAN SMOTE DAN ALGORITMA GENETIKA TERHADAP PERFORMA PREDIKSI CACAT PERANGKAT LUNAK BERBASIS *MOBILE* (Oleh: Ade Rahmad; Pembimbing: Rudy Herteno, S.Kom, M.Kom. dan Irwan Budiman, S.T., M.Kom.; 2026 ; 92 halaman)

Prediksi cacat *Just-in-Time* (JIT) memungkinkan identifikasi dini terhadap *commit* penyebab cacat selama proses pengembangan perangkat lunak. Namun, dataset JIT Android umumnya memiliki ketidakseimbangan kelas yang tinggi, yang dapat menurunkan performa model dan menyebabkan prediksi cenderung bias terhadap kelas mayoritas. Penelitian ini menganalisis pengaruh *Synthetic Minority Over-sampling Technique* (SMOTE) multi-varian dan optimasi hyperparameter berbasis *Genetic Algorithm* (GA) terhadap prediksi cacat JIT pada aplikasi Android. Eksperimen dilakukan pada 12 dataset benchmark Android menggunakan tiga model klasifikasi, yaitu *Logistic Regression* (LR), *Random Forest* (RF), dan *XGBoost* (XGB). Empat skenario eksperimen dievaluasi, yaitu *Baseline*, *Oversampling*, *GA*, dan *Oversampling + GA*. Performa model dievaluasi menggunakan *Stratified 10-Fold Cross-Validation* dengan metrik AUC dan *Matthews Correlation Coefficient* (MCC), sedangkan signifikansi statistik dianalisis menggunakan uji Friedman yang dilanjutkan dengan uji *Wilcoxon signed-rank* disertai koreksi Holm. Hasil penelitian menunjukkan bahwa skenario *Oversampling + GA* menghasilkan rata-rata AUC tertinggi pada seluruh model, yaitu 0,746 untuk LR, 0,813 untuk RF, dan 0,803 untuk XGB. Pada metrik MCC, performa terbaik diperoleh oleh *Oversampling* pada LR (0,311) serta oleh *Oversampling + GA* pada RF (0,440) dan XGB (0,422). Analisis statistik mengonfirmasi adanya perbedaan yang signifikan di antara keempat skenario eksperimen, baik pada AUC ($p = 2,16 \times 10^{-5}$) maupun MCC ($p = 4,19 \times 10^{-4}$). Secara keseluruhan, temuan penelitian ini menunjukkan bahwa penanganan ketidakseimbangan kelas melalui oversampling memberikan kontribusi yang lebih konsisten dalam meningkatkan performa prediksi cacat JIT Android dibandingkan optimasi hyperparameter saja, sementara kombinasi keduanya memberikan performa keseluruhan terbaik.

Kata kunci: *Just-in-Time defect prediction, SMOTE, algoritma genetika, ketidakseimbangan kelas, prediksi cacat perangkat lunak*

ABSTRACT

THE EFFECT OF MULTI-VARIANT SMOTE OVERSAMPLING AND GENETIC ALGORITHM ON THE PERFORMANCE OF MOBILE-BASED SOFTWARE DEFECT PREDICTION (Ade Rahmad; Advisor: Rudy Herteno, S.Kom, M.Kom. and Irwan Budiman, S.T., M.Kom.; 2026 ; 92 pages)

Just-in-Time (JIT) defect prediction enables early identification of defect-inducing commits during software development. However, Android JIT datasets are commonly characterized by severe class imbalance, which may reduce model performance and bias predictions toward the majority class. This study investigates the impact of multi-variant Synthetic Minority Over-sampling Technique (SMOTE) and Genetic Algorithm (GA)-based hyperparameter optimization on Android JIT defect prediction. Experiments were conducted on 12 Android benchmark datasets using three classification models, namely Logistic Regression (LR), Random Forest (RF), and XGBoost (XGB). Four experimental scenarios were evaluated: Baseline, Oversampling, GA, and Oversampling+GA. Model performance was assessed using Stratified 10-Fold Cross-Validation with AUC and Matthews Correlation Coefficient (MCC), while statistical significance was analyzed using the Friedman test followed by the Wilcoxon signed-rank test with Holm correction. The results show that the Oversampling+GA scenario achieved the highest average AUC for all models, reaching 0.746 for LR, 0.813 for RF, and 0.803 for XGB. For MCC, the best performance was obtained by Oversampling for LR (0.311) and by Oversampling+GA for RF (0.440) and XGB (0.422). Statistical analysis confirmed significant differences among the four experimental scenarios for both AUC ($p = 2.16 \times 10^{-5}$) and MCC ($p = 4.19 \times 10^{-4}$). Overall, the findings indicate that handling class imbalance through oversampling contributes more consistently to improving Android JIT defect prediction performance than hyperparameter optimization alone, while combining both approaches provides the best overall performance.

Keywords: Just-in-Time defect prediction, SMOTE, genetic algorithm, class imbalance, software defect prediction

PRAKATA

Puji syukur penulis panjatkan kehadiran Tuhan Yang Maha Esa karena atas rahmat dan karunia-Nya penulis dapat menyelesaikan skripsi yang berjudul “Pengaruh *Oversampling* Multi-Varian SMOTE dan Algoritma Genetika terhadap Performa Prediksi Cacat Perangkat Lunak Berbasis *Mobile*”. Skripsi ini disusun sebagai salah satu syarat untuk memperoleh gelar Sarjana Komputer pada Program Studi Ilmu Komputer Fakultas Matematika dan Ilmu Pengetahuan Alam Universitas Lambung Mangkurat.

Dalam proses penyusunan skripsi ini, penulis memperoleh banyak bantuan, dukungan, serta bimbingan dari berbagai pihak. Oleh karena itu, pada kesempatan ini penulis menyampaikan ucapan terima kasih kepada:

- (1) Tuhan Yang Maha Esa atas segala rahmat, kesehatan, dan kesempatan yang diberikan selama proses penyusunan skripsi.
- (2) Keluarga tercinta yang selalu memberikan doa, dukungan, perhatian, dan semangat kepada penulis.
- (3) Bapak Rudy Herteno, S.Kom., M.Kom., selaku dosen pembimbing utama yang telah memberikan arahan, bimbingan, serta masukan selama proses penelitian dan penyusunan skripsi.
- (4) Bapak Irwan Budiman, S.T., M.Kom., selaku dosen pembimbing pendamping yang telah memberikan saran dan dukungan dalam penyusunan skripsi.
- (5) Ibu Dwi Kartini, S.Kom., M.Kom., selaku Koordinator Program Studi Ilmu Komputer FMIPA Universitas Lambung Mangkurat.
- (6) Seluruh dosen Program Studi Ilmu Komputer FMIPA Universitas Lambung Mangkurat yang telah memberikan ilmu dan pengalaman selama masa perkuliahan.
- (7) Teman-teman Ilmu Komputer angkatan 2022 yang telah memberikan dukungan dan semangat selama proses perkuliahan dan penyusunan skripsi.

Penulis menyadari bahwa skripsi ini masih memiliki berbagai kekurangan.

Oleh karena itu, penulis mengharapkan kritik dan saran yang membangun untuk perbaikan di masa mendatang. Penulis berharap skripsi ini dapat memberikan manfaat bagi pembaca dan menjadi salah satu referensi dalam pengembangan penelitian di bidang *Software Defect Prediction*.

Banjarbaru, 17 Juni 2026

A handwritten signature in black ink, consisting of a stylized 'A' followed by a cursive 'de' and 'Rahmad'.

Ade Rahmad

DAFTAR TABEL

Tabel 1. Keaslian Penelitian.....	9
Tabel 2. Perancangan Penelitian	10
Tabel 3. Deskripsi Dataset	31
Tabel 4. Deskripsi Fitur.....	32
Tabel 5. Ringkasan Karakteristik Dataset.....	37
Tabel 6. Ringkasan Missing value Dataset	38
Tabel 7. Contoh Data Sebelum Imputasi	38
Tabel 8. Data Setelah Normalisasi	39
Tabel 9. Jarak Tetangga Terdekat dari Data Missing	40
Tabel 10. Data Setelah Imputasi	40
Tabel 11. Perubahan Kelas Sebelum dan Sesudah Transformasi	41
Tabel 12. Contoh Data Awal Simulasi Stratified Cross Validation.....	42
Tabel 13. Contoh Pembagian Fold Stratified.....	42
Tabel 14. Contoh Distribusi Kelas per Fold (SMOTE-Firewall).....	44
Tabel 15. Sebelum dan Sesudah Oversampling (SMOTE).....	45
Tabel 16. Contoh Data Awal Simulasi SMOTE	45
Tabel 17. Jarak Tetangga Minoritas terhadap ID 6.....	46
Tabel 18. Hasil data sintetis simulasi SMOTE	47
Tabel 19. Sebelum dan Sesudah Oversampling (ADASYN)	47
Tabel 20. Contoh Data Awal Simulasi ADASYN	47
Tabel 21. Tingkat Kesulitan Data Minoritas pada Simulasi ADASYN	48
Tabel 22. Hasil data sintetis simulasi ADASYN	48
Tabel 23. Sebelum dan Sesudah Oversampling (Borderline-SMOTE)	49
Tabel 24. Contoh Data Awal Simulasi Borderline-SMOTE.....	49
Tabel 25. Status Data Minoritas pada Borderline-SMOTE	50
Tabel 26. Hasil data sintetis simulasi Borderline-SMOTE.....	50
Tabel 27. Sebelum dan Sesudah Oversampling (K-Means SMOTE).....	51
Tabel 28. Contoh Data Awal Simulasi K-Means SMOTE	51
Tabel 29. Hasil Clustering pada K-Means SMOTE.....	52
Tabel 30. Imbalance Ratio Tiap Cluster	52
Tabel 31. Hasil Data Sintetis Simulasi K-Means SMOTE	52
Tabel 32. Sebelum dan Sesudah Oversampling (SMOTE-Tomek).....	53
Tabel 33. Contoh Data Awal Simulasi SMOTE-Tomek	53
Tabel 34. Hasil data sintetis simulasi SMOTE-Tomek.....	54
Tabel 35. Parameter Yang Dituning	54
Tabel 36. Contoh Hasil Best parameter (GA).....	55
Tabel 37. Contoh Populasi Awal	55
Tabel 38. Hasil Crossover	56
Tabel 39. Hasil Mutasi	56
Tabel 40. Contoh Best parameter (Oversampling + GA)	57
Tabel 41. Data Simulasi Model Klasifikasi	57

Tabel 42. Data Simulasi AUC.....	60
Tabel 43. Hasil Prediksi Threshold 0,5	60
Tabel 44. Perhitungan TP, TN, FP, FN, TPR, FPR dan Titik ROC Seluruh Threshold	61
Tabel 45. Titik-titik ROC	62
Tabel 46. Data Simulasi MCC	62
Tabel 47. Data Mentah Performa Baseline	63
Tabel 48. Data Mentah Performa Oversampling (ADASYN)	64
Tabel 49. Data Mentah Performa Oversampling (Borderline-SMOTE).....	65
Tabel 50. Data Mentah Performa Oversampling (K-Means SMOTE)	65
Tabel 51. Data Mentah Performa Oversampling (SMOTE)	66
Tabel 52. Data Mentah Performa Oversampling (SMOTE-Tomek)	66
Tabel 53. Data Mentah Performa GA	67
Tabel 54. Data Mentah Performa Oversampling + GA (ADASYN)	68
Tabel 55. Data Mentah Performa Oversampling + GA (Borderline-SMOTE).....	69
Tabel 56. Data Mentah Performa Oversampling + GA (K-Means SMOTE)	69
Tabel 57. Data Mentah Performa Oversampling + GA (SMOTE)	70
Tabel 58. Data Mentah Performa Oversampling + GA (SMOTE-Tomek)	70
Tabel 59. Hasil Best-of-5 Oversampling (AUC)	71
Tabel 60. Hasil Best-of-5 Oversampling (MCC)	72
Tabel 61. Hasil Best-of-5 Oversampling + GA (AUC)	73
Tabel 62. Hasil Best-of-5 Oversampling + GA (MCC)	74
Tabel 63. Perbandingan AUC Tiap Skenario Per Dataset	76
Tabel 64. Perbandingan MCC Tiap Skenario Per Dataset	77
Tabel 65. Perbandingan AUC Tiap Skenario Per Model	79
Tabel 66. Perbandingan MCC Tiap Skenario Per Model	79
Tabel 69. Data Simulasi Uji Friedman.....	81
Tabel 70. <i>Ranking</i> Setiap Skenario	81
Tabel 71. Rata-rata <i>Ranking</i> Friedman	81
Tabel 72. Data Simulasi Wilcoxon	82
Tabel 73. Selisih Performa	82
Tabel 74. Selisih Absolut dan <i>Ranking</i>	83
Tabel 75. Nilai R	83
Tabel 76. <i>Average rank</i> AUC Antar Skenario	85
Tabel 77. Hasil Uji <i>Post-hoc</i> AUC	86
Tabel 78. <i>Average rank</i> MCC Antar Skenario.....	86
Tabel 79. Hasil Uji <i>Post-hoc</i> MCC	87

DAFTAR GAMBAR

Gambar 1. Diagram Alir Alur Kerja SMOTE.....	15
Gambar 2. Diagram Alir Alur Kerja Borderline-SMOTE	17
Gambar 3. Diagram Alir Alur Kerja ADASYN.....	19
Gambar 4. Diagram Alir Alur Kerja SMOTE-Tomek	21
Gambar 5. Diagram Alir Alur Kerja K-Means SMOTE	22
Gambar 6. Diagram Alir Alur Algoritma Genetika	24
Gambar 7. Diagram Alir Alur Logistic Regression	25
Gambar 8. Diagram Alir Alur Random Forest.....	26
Gambar 9. Diagram Alir Alur XGBoost.....	27
Gambar 10. Alur Penelitian.....	33
Gambar 12. Ilustrasi Pergantian Fold	43
Gambar 13. Perbandingan AUC Tiap Skenario Per Dataset (Bagian 1).....	77
Gambar 14. Perbandingan AUC Tiap Skenario Per Dataset (Bagian 2).....	77
Gambar 15. Perbandingan MCC Tiap Skenario Per Dataset (Bagian 1)	78
Gambar 16. Perbandingan MCC Tiap Skenario Per Dataset (Bagian 2)	78
Gambar 17. Perbandingan AUC Tiap Skenario Per Model	79
Gambar 18. Perbandingan MCC Tiap Skenario Per Model.....	80
Gambar 19. Diagram Batang <i>Average rank</i> AUC Antar Skenario	85
Gambar 20. Diagram Batang <i>Average rank</i> MCC Antar Skenario.....	87

DAFTAR LAMPIRAN

Lampiran 1. Source Code Pra-Processing Data	97
Lampiran 2. Source Code Skenario Baseline.....	102
Lampiran 3. Source Code Skenario Oversampling.....	104
Lampiran 4. Source Code Skenario GA.....	110
Lampiran 5. Source Code Oversampling + GA	119
Lampiran 6. Source Code Analisis Statistik (AUC & MCC)	130
Lampiran 7. Dokumentasi Kegiatan	136

DAFTAR ISI

HALAMAN JUDUL.....	i
HALAMAN PENGESAHAN.....	ii
PERNYATAAN.....	iii
ABSTRAK	iv
ABSTRACT	v
PRAKATA.....	vi
DAFTAR TABEL	viii
DAFTAR GAMBAR	x
DAFTAR LAMPIRAN.....	xi
DAFTAR ISI.....	xii
BAB I PENDAHULUAN.....	1
1.1 Latar Belakang	1
1.2 Rumusan Masalah	4
1.3 Batasan Penelitian	4
1.4 Tujuan Penelitian.....	5
1.5 Manfaat Penelitian.....	5
BAB II TINJAUAN PUSTAKA.....	7
2.1 Kajian Terdahulu	7
2.2 Keaslian Penelitian	8
2.3 Dataset JIT defect prediction Android	10
2.4 Software Defect Prediction dan JIT defect prediction	11
2.5 Stratified K-Fold Cross-Validation	13
2.6 Ketidakseimbangan Kelas	13
2.7 Varian SMOTE.....	14
2.7.1 SMOTE.....	14
2.7.2 Borderline-SMOTE	15
2.7.3 ADASYN.....	17
2.7.4 SMOTE-Tomek	19
2.7.5 K-Means SMOTE	21

2.8	Optimasi hyperparameter	22
2.9	Algoritma Genetika	23
2.10	Model Klasifikasi	24
2.10.1	Logistic Regression.....	24
2.10.2	Random Forest	25
2.10.3	XGBoost.....	26
2.11	Metrik Evaluasi	27
2.11.1	AUC (Area Under the ROC Curve).....	27
2.11.2	MCC (Matthews Correlation Coefficient)	27
2.12	Uji Friedman.....	28
2.13	Uji Wilcoxon & Holm.....	28
BAB III	METODE PENELITIAN	30
3.1	Alat Penelitian	30
3.2	Bahan Penelitian.....	30
3.3	Prosedur Penelitian.....	32
BAB IV	HASIL DAN PEMBAHASAN	37
4.1	Hasil.....	37
4.1.1	Pengumpulan Dataset	37
4.1.2	Pra-proses Data	38
4.1.3	Skema Evaluasi dan Alur Eksperimen.....	41
4.1.4	Penerapan Skenario Eksperimen	43
4.1.5	Pemodelan Klasifikasi	57
4.1.6	Evaluasi Dengan AUC dan MCC	60
4.1.7	Analisis Statistik	80
4.2	Pembahasan	88
BAB V	PENUTUP.....	91
5.1	Kesimpulan.....	91
5.2	Saran	92
DAFTAR PUSTAKA		93
DAFTAR LAMPIRAN.....		97