# Abstrak

Data berdimensi tinggi sering memiliki jumlah fitur yang jauh lebih besar daripada jumlah sampel, sehingga pembelajaran model menjadi tidak stabil dan biaya komputasi meningkat. Penelitian ini mengusulkan Genetic Algorithm–optimized Random Forest (GA-RF) untuk memperbaiki proses seleksi fitur pada kondisi tersebut. Berbeda dari Random Forest standar yang memilih fitur secara acak pada setiap pemisahan (split), pendekatan ini melakukan seleksi fitur pada level pohon: untuk setiap decision tree, Genetic Algorithm (GA) mencari mask biner fitur dan mengevaluasinya menggunakan akurasi decision tree berbasis cross-validation, serta menambahkan penalti jika jumlah fitur terpilih terlalu besar. Kromosom terbaik kemudian digunakan untuk melatih pohon yang teroptimasi; sejumlah pohon terbaik dipilih untuk membentuk ensemble dengan majority voting. Eksperimen dilakukan pada lima dataset benchmark (Arcene, Gisette, Madelon, Dorothea, dan Dexter) dan dibandingkan dengan baseline Random Forest, reduksi dimensi (PCA/TruncatedSVD) yang diikuti Random Forest, serta seleksi fitur konvensional (Variance Threshold dan SelectKBest). Hasil menunjukkan bahwa kinerja GA-RF sangat bergantung pada dataset: peningkatan besar terjadi pada Madelon (68,33% → 84,00%) dan peningkatan kecil pada Arcene (77,00% → 80,00%), namun performa GA-RF setara atau lebih rendah pada Gisette, Dorothea, dan Dexter. Pada Dorothea, TruncatedSVD+RF menghasilkan akurasi terbaik (95,14%). Analisis relevansi fitur menggunakan Mutual Information dan varians mendukung bahwa seleksi fitur berbasis GA lebih bermanfaat ketika banyak fitur bersifat lemah informasinya atau redundan.

# Abstract

High-dimensional datasets often contain far more features than samples, making learning unstable and computation expensive. This study proposes a Genetic Algorithm–optimized Random Forest (GA-RF) to improve feature selection in such settings. Unlike standard Random Forest that samples features randomly at each split, the proposed approach performs tree-level feature selection: for each decision tree, a GA searches a binary feature mask and evaluates it using cross-validated decision tree accuracy with a penalty for overly large subsets. The best mask is then used to train an optimized tree; the top-performing trees are assembled into an ensemble using majority voting. Experiments were conducted on five benchmark datasets (Arcene, Gisette, Madelon, Dorothea, and Dexter) and compared against baseline Random Forest, dimensionality reduction (PCA/TruncatedSVD) followed by Random Forest, and conventional filter-based selection (Variance Threshold and SelectKBest). Results show that GA-RF performance is dataset-dependent: it substantially improves Madelon (68.33% → 84.00%) and slightly improves Arcene (77.00% → 80.00%), while it is comparable or lower on Gisette, Dorothea, and Dexter. TruncatedSVD+RF yields the best accuracy on Dorothea (95.14%). Feature relevance analysis using Mutual Information and variance supports the observation that GA-based selection is most beneficial when many features are weakly informative or redundant.