

Komparasi Algoritma Support Vector Machine, Random Forest, dan XGBoost: Prediksi Penyakit Huntington Berdasarkan Data Ekspresi Gen

Comparison of Support Vector Machine, Random Forest, and XGBoost Algorithms: Predicting Huntington's Disease Based on Gene Expression Data

Reza Zaynury Winata^{1*}, Erliyan Redy Susanto²

¹ Fakultas Teknik Dan Ilmu Komputer, Prodi Sistem Informasi, Universitas Teknokrat Indonesia, Bandar Lampung, Indonesia

² Fakultas Teknik Dan Ilmu Komputer, Prodi Ilmu Komputer, Universitas Teknokrat Indonesia, Bandar Lampung, Indonesia

Email: ¹*reza_zaynury_winata@teknokrat.ac.id, ²erliyan.redy@teknokrat.ac.id

Email Penulis Korespondensi: reza_zaynury_winata@teknokrat.ac.id

Received: 28/08/2026, Revised: 13/09/2026, Accepted: 21/09/2026, Published: 22/09/2026

Abstrak—Penyakit Huntington adalah gangguan neurodegeneratif progresif akibat mutasi gen Huntingtin (HTT), dan diagnosis dini masih menjadi tantangan yang berkelanjutan. Pemanfaatan data ekspresi gen melalui pembelajaran mesin berpotensi meningkatkan akurasi prediksi penyakit ini, namun karakteristik data microarray dimensi fitur sangat tinggi dengan ukuran sampel terbatas (small-n large-p) serta ketidakseimbangan kelas meningkatkan risiko overfitting dan kebocoran data. Studi ini membandingkan kinerja Support Vector Machine (SVM), Random Forest, dan Extreme Gradient Boosting (XGBoost) dalam memprediksi status penyakit Huntington menggunakan dataset microarray GSE135589. Pipeline penelitian mencakup standarisasi data, seleksi fitur (SelectKBest berbasis uji F ANOVA), penyeimbangan kelas (SMOTE), dan penyetelan hyperparameter melalui GridSearchCV dengan Stratified 5-Fold Cross-Validation di mana jumlah fitur (k) turut dituning secara nested bersama parameter model lainnya untuk mencegah kebocoran informasi antara data latih dan data uji. Seluruh tahapan pra-pemrosesan diterapkan secara eksklusif pada data pelatihan (fit-on-train, transform-on-test) guna menjamin estimasi kinerja yang tidak bias secara optimis. Hasil menunjukkan bahwa SVM kernel linier memberikan kinerja terbaik, dengan akurasi 88,89%, presisi 89,66%, recall 96,30%, F1-score 92,86%, dan ROC-AUC 90,95%, mengungguli Random Forest dan XGBoost pada data uji independen. Temuan ini mengindikasikan bahwa pendekatan linier yang diregulasi lebih unggul dibanding model berbasis pohon pada kondisi dimensi fitur jauh melebihi ukuran sampel. Kebaruan utama penelitian ini terletak pada penyetelan jumlah fitur (k) sebagai hyperparameter nested di dalam skema validasi silang bukan langkah terpisah sebelum pelatihan model seperti pada mayoritas studi sejenis sehingga menghasilkan estimasi generalisasi yang lebih ketat bebas dari kebocoran data, sekaligus menjadi tolok ukur pembandingan pertama yang komprehensif antara ketiga algoritma tersebut secara khusus pada domain data ekspresi gen penyakit Huntington.

Kata Kunci : Penyakit Huntington; Pembelajaran Mesin; Support Vector Machine; Random Forest; XGBoost

Abstract—Huntington's disease is a progressive neurodegenerative disorder caused by mutations in the Huntingtin (HTT) gene, and early diagnosis remains an ongoing challenge. Leveraging gene expression data through machine learning holds potential for improving the prediction accuracy of this disease; however, the characteristics of microarray data very high feature dimensionality combined with a limited sample size (small-n large-p) together with class imbalance, increase the risk of overfitting and data leakage. This study compares the performance of Support Vector Machine (SVM), Random Forest, and Extreme Gradient Boosting (XGBoost) in predicting Huntington's disease status using the GSE135589 microarray dataset. The research pipeline includes data standardization, feature selection (SelectKBest based on the ANOVA F-test), class balancing (SMOTE), and hyperparameter tuning via GridSearchCV with Stratified 5-Fold Cross-Validation, in which the number of features (k) is also tuned as a nested hyperparameter alongside other model parameters to prevent information leakage between the training and test data. All preprocessing steps were applied exclusively to the training data (fit-on-train, transform-on-test) to ensure performance estimates free from optimistic bias. The results show that the linear-kernel SVM achieved the best performance, with an accuracy of 88.89%, precision of 89.66%, recall of 96.30%, F1-score of 92.86%, and ROC-AUC of 90.95%, outperforming Random Forest and XGBoost on the independent test set. These findings indicate that regularized linear approaches outperform tree-based models under conditions where feature dimensionality far exceeds sample size. The main novelty of this study lies in tuning the number of features (k) as a nested hyperparameter within the cross-validation scheme rather than as a separate step prior to model training, as is common in most comparable studies, thereby producing a stricter, leakage-free generalization estimate, while also serving as the first comprehensive benchmark comparing the three algorithms specifically within the domain of Huntington's disease gene expression data.

Keywords: Huntington's Disease; Machine Learning; Support Vector Machine; Random Forest; XGBoost

1. PENDAHULUAN

Penyakit Huntington adalah gangguan *neurodegeneratif progresif*, *autosomal* dominan yang disebabkan oleh perluasan abnormal pengulangan *trinukleotida* CAG dalam gen *Huntingtin* (HTT) pada *kromosom* 4 [1]. Perluasan ini menghasilkan protein *Huntingtin* mutan toksik yang secara progresif merusak *neuron*, terutama di *striatum* dan *korteks serebral*, sehingga menimbulkan gejala motorik (*korea*), kognitif, dan kejiwaan yang memburuk seiring waktu [2]. Kemajuan dalam teknologi *microarray* dan sekuensing generasi berikutnya telah memungkinkan para peneliti

untuk secara bersamaan mengukur profil ekspresi ribuan gen dari jaringan atau darah pasien HD, membuka peluang besar bagi pendekatan *bioinformatika* dan pembelajaran mesin untuk mengungkap pola molekuler yang membedakan individu dengan HD dari kontrol, sekaligus berpotensi mengidentifikasi biomarker baru yang mendukung diagnosis dini dan pemahaman yang lebih baik tentang patogenesis penyakit [3][4].

Meskipun data ekspresi gen skala besar semakin mudah diakses melalui repositori publik seperti *Gene Expression Omnibus (GEO)*, penggunaannya untuk klasifikasi berbasis pembelajaran mesin menghadapi tantangan metodologis yang signifikan [5]. Data microarray umumnya dicirikan oleh struktur “n kecil-p besar”, di mana jumlah fitur (puluhan ribu probe gen) jauh melebihi jumlah sampel (hanya puluhan hingga ratusan individu), sehingga model sangat rentan terhadap *overfitting* jika tidak diatasi melalui pemilihan dan validasi fitur yang tepat [6]. Masalah ini diperparah oleh distribusi kelas yang tidak seimbang antara kelompok HD dan kontrol, yang dapat membiaskan model ke arah kelas mayoritas dan mengurangi kemampuan mereka untuk mengenali kelas minoritas, meskipun kelas tersebut seringkali merupakan target klinis utama [7]. Teknik penyeimbangan kelas populer seperti SMOTE, yang bekerja melalui interpolasi linier antar sampel kelas minoritas, juga memiliki keterbatasan yang signifikan ketika diterapkan pada data berdimensi sangat tinggi seperti data microarray. Temuan ini didukung oleh kajian-kajian yang lebih mutakhir: Lyu et al. menyoroti bahwa interpolasi linier yang menjadi dasar SMOTE sulit menangkap hubungan non-linier antar sampel, khususnya pada ruang fitur berdimensi tinggi [8], sementara Hassanat et al. dalam tinjauan kritisnya merekomendasikan kehati-hatian dalam penggunaan oversampling sintesis karena berisiko menimbulkan *overfitting* dan generalisasi yang buruk [9]. Selain itu, risiko kebocoran data, misalnya, ketika penskalaan, pemilihan fitur, atau pengambilan sampel berlebih dilakukan sebelum membagi data menjadi set pelatihan dan pengujian seringkali diabaikan, menyebabkan perkiraan kinerja model menjadi terlalu optimis dan tidak mencerminkan kemampuan generalisasi yang sebenarnya [10]. Sampai saat ini, literatur yang secara khusus membandingkan algoritma SVM, Random Forest, dan XGBoost pada data ekspresi gen penyakit Huntington di bawah protokol anti-kebocoran yang ketat masih terbatas [11].

Beberapa penelitian terdahulu telah mengeksplorasi penerapan pembelajaran mesin pada data ekspresi gen untuk mendukung diagnosis maupun karakterisasi penyakit neurodegeneratif dan kanker. Cheng et al. mengidentifikasi gen-gen yang berkontribusi terhadap penyakit Huntington melalui pendekatan pembelajaran mesin, namun belum melakukan perbandingan sistematis antar-algoritma klasifikasi maupun penerapan protokol anti-kebocoran data secara eksplisit [11]. Sejalan dengan itu, Dash et al. mengidentifikasi biomarker baru pada penyakit Huntington melalui meta-analisis ekspresi gen diferensial yang dikombinasikan dengan pembelajaran mesin, tetapi fokus penelitian tersebut lebih diarahkan pada agregasi data lintas studi daripada evaluasi komparatif algoritma pada satu dataset microarray tunggal [12]. Ganesh et al. dalam tinjauan komprehensifnya mengenai kecerdasan buatan untuk diagnosis Huntington juga menyoroti bahwa sebagian besar studi yang ada masih jarang menekankan isu kebocoran data dan validasi yang ketat pada kondisi dimensi fitur yang jauh melebihi jumlah sampel ($p \gg n$) [4]. Di luar konteks Huntington, perbandingan algoritma SVM, Random Forest, dan XGBoost telah banyak dilakukan pada domain penyakit lain, misalnya Sugihartono et al. yang membandingkan keempat algoritma tersebut (termasuk k-NN) dengan SMOTE untuk deteksi stunting [13], serta Gao et al. yang memanfaatkan Random Forest dan pendekatan ensemble learning untuk seleksi gen biomarker pada data ekspresi kanker dengan karakteristik ketidakseimbangan kelas yang serupa [14]. Meskipun studi-studi tersebut menunjukkan bahwa model berbasis pohon maupun model linier dapat memberikan performa kompetitif pada data berdimensi tinggi, sebagian besar belum menerapkan skema *fit-on-train*, *transform-on-test* secara konsisten pada seluruh tahapan praproses (standarisasi, seleksi fitur, dan SMOTE), sehingga risiko estimasi kinerja yang terlalu optimis akibat kebocoran data masih berpotensi terjadi [15]. Berdasarkan celah tersebut, penelitian ini memposisikan diri sebagai studi yang secara khusus membandingkan kinerja SVM, Random Forest, dan XGBoost pada dataset ekspresi gen penyakit Huntington (GSE135589) dengan protokol anti-kebocoran data yang diterapkan secara ketat di setiap tahap, sekaligus menyajikan evaluasi transparan dan dapat direproduksi yang belum banyak ditekankan pada literatur sejenis.

Berdasarkan masalah-masalah tersebut, penelitian ini bertujuan untuk membandingkan kinerja tiga algoritma klasifikasi populer SVM, Random Forest, dan XGBoost dalam memprediksi status penyakit Huntington dari data ekspresi gen, serta untuk mengidentifikasi gen dengan kontribusi prediktif tertinggi sebagai kandidat biomarker. Metode yang digunakan meliputi praproses data (pemeriksaan nilai yang hilang dan standarisasi fitur), pemilihan fitur menggunakan *SelectKBest* berdasarkan uji *F ANOVA* untuk mengurangi dimensi fitur yang sangat tinggi, penyeimbangan kelas pada data pelatihan menggunakan *Synthetic Minority Over-sampling Technique (SMOTE)*, dan penyetelan hyperparameter melalui *GridSearchCV* dengan skema validasi silang *k-fold* bertingkat [16]. Semua langkah praproses penskalaan, pemilihan fitur, dan SMOTE diterapkan secara ketat hanya pada data pelatihan (*fit-on-train*, *transform-on-test*) untuk mencegah kebocoran informasi ke dalam data uji. Secara umum, tahapan studi meliputi: pengumpulan dan eksplorasi data, pembagian data latih dan uji yang bertingkat, praproses dan pemilihan fitur, penyeimbangan kelas, pelatihan dan penyetelan model, dan akhirnya evaluasi kinerja pada data uji independen [13].

Pengujian pada data uji independen menunjukkan bahwa model SVM kernel linier memberikan kinerja klasifikasi terbaik dibandingkan dengan Random Forest dan XGBoost pada dataset ini, mencapai akurasi, F1 Score, dan ROC-AUC yang lebih tinggi, menunjukkan keunggulan pendekatan linier yang diregulasi dalam kondisi di mana dimensi fitur jauh melebihi ukuran sampel [12].

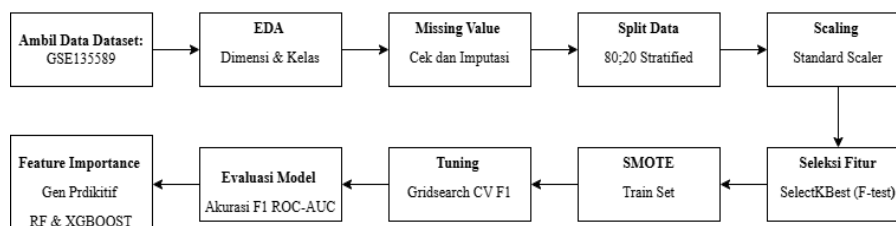
Keunggulan SVM kernel linier atas Random Forest dan XGBoost pada kondisi small-n large-p ini dapat dijelaskan lebih lanjut melalui kerangka bias-variance tradeoff. Model ensemble berbasis pohon seperti Random Forest dan XGBoost membentuk batas keputusan yang sangat fleksibel melalui pemisahan (split) rekursif terhadap ribuan kandidat fitur; ketika jumlah fitur jauh melebihi jumlah sampel, sebagian split cenderung memanfaatkan pola gen yang kebetulan tampak informatif pada subset data pelatihan tertentu (chance correlation), sehingga menghasilkan model dengan varians estimasi yang tinggi meskipun mekanisme agregasi bootstrap pada Random Forest maupun regularisasi bertahap pada XGBoost telah diterapkan untuk meredamnya. Cantor et al. menunjukkan secara empiris bahwa akurasi Random Forest menurun signifikan pada data ekspresi gen dengan rasio fitur-terhadap-sampel yang sangat besar dan ukuran sampel yang sangat kecil, karena pohon-pohon penyusun ensemble kesulitan membedakan sinyal biologis yang sebenarnya dari derau berdimensi tinggi [17]. Sebaliknya, SVM kernel linier mengasumsikan satu hiperbidang pemisah tunggal yang jauh lebih sederhana (bias lebih tinggi), namun kompleksitasnya dikendalikan secara eksplisit melalui parameter regularisasi C sehingga variansnya tetap rendah. Kondisi ini diperkuat oleh sifat geometris ruang fitur berdimensi tinggi, di mana data cenderung menjadi lebih mudah dipisahkan secara linier seiring bertambahnya dimensi, sehingga asumsi linieritas SVM lebih sesuai dengan struktur data microarray dibandingkan batas keputusan non-linier yang kompleks dari model berbasis pohon.

Studi ini memberikan evaluasi komparatif yang transparan dan dapat direproduksi dari tiga algoritma pembelajaran mesin populer pada data ekspresi gen HD, lengkap dengan protokol pra-pemrosesan yang secara eksplisit mencegah kebocoran data, suatu aspek yang seringkali kurang ditekankan dalam studi bioinformatika serupa [18]. Selain itu, daftar gen dengan skor kepentingan fitur tertinggi dari model berbasis pohon dapat berfungsi sebagai referensi awal untuk penelitian biologi dan klinis lebih lanjut tentang mekanisme molekuler HD, sementara kerangka kerja metodologis yang digunakan berpotensi untuk diadaptasi untuk studi klasifikasi penyakit lain berdasarkan data ekspresi gen sampel kecil dengan fitur besar [19].

Penelitian ini memberikan beberapa kontribusi utama. Pertama, penelitian ini menyajikan evaluasi komparatif yang transparan dan dapat direproduksi terhadap tiga algoritma pembelajaran mesin populer, yaitu Support Vector Machine, Random Forest, dan XGBoost, dalam memprediksi status penyakit Huntington berdasarkan data ekspresi gen, mengingat literatur yang secara khusus membandingkan ketiga algoritma tersebut pada domain ini di bawah protokol anti-kebocoran data yang ketat masih terbatas. Kedua, penelitian ini menerapkan protokol pra-pemrosesan anti-kebocoran data secara ketat, di mana seluruh tahapan penskalaan fitur, pemilihan fitur menggunakan SelectKBest berbasis uji F ANOVA, serta penyeimbangan kelas menggunakan SMOTE diterapkan secara eksklusif pada data pelatihan (fit-on-train, transform-on-test), sehingga estimasi kinerja model yang dihasilkan mencerminkan kemampuan generalisasi yang sebenarnya dan tidak bias secara optimis. Ketiga, penelitian ini menghasilkan daftar gen dengan skor kepentingan fitur tertinggi dari model berbasis pohon yang dapat menjadi kandidat biomarker awal, sekaligus referensi bagi penelitian biologi dan klinis lanjutan mengenai mekanisme molekuler penyakit Huntington. Keempat, penelitian ini menawarkan kerangka kerja metodologis yang mencakup praproses data, pemilihan fitur, penyeimbangan kelas, dan penyetelan hyperparameter melalui GridSearchCV dengan validasi silang k-fold bertingkat, yang berpotensi diadaptasi untuk studi klasifikasi penyakit lain dengan karakteristik data serupa, yaitu ukuran sampel kecil dengan dimensi fitur besar. Terakhir, temuan bahwa model SVM kernel linier memberikan kinerja klasifikasi terbaik dibandingkan Random Forest dan XGBoost pada dataset ini memberikan bukti empiris mengenai keunggulan pendekatan linier yang diregulasi dalam kondisi di mana dimensi fitur jauh melebihi ukuran sampel, sehingga dapat menjadi acuan praktis bagi peneliti dalam memilih algoritma yang sesuai untuk data ekspresi gen berskala kecil.

2. METODOLOGI PENELITIAN

Bagian ini menjelaskan metodologi yang digunakan dalam penelitian ini, meliputi tahapan mulai dari pengumpulan data, pra-pemrosesan, dan pengembangan model hingga evaluasi kinerja algoritma. Bagian ini juga menyajikan karakteristik dan sumber data yang digunakan, beserta dasar-dasar matematis dari algoritma klasifikasi dan metrik evaluasi yang diterapkan dalam penelitian ini. Alur kerja penelitian secara keseluruhan disajikan pada Gambar 1.



Gambar 1. Tahapan Penelitian

2.1 Tahapan Penelitian

Penelitian ini dimulai dengan pengumpulan data ekspresi gen dari repositori *Gene Expression Omnibus* (GEO) menggunakan dataset GSE135589, yang kemudian dimuat ke dalam lingkungan analisis untuk pemrosesan lebih

lanjut. Tahap selanjutnya adalah melakukan Analisis Data Eksplorasi (EDA), yang meliputi pemeriksaan dimensi dataset, menganalisis distribusi kelas target, dan menghitung statistik deskriptif untuk setiap fitur untuk mendapatkan gambaran awal karakteristik data [20]. Kemudian dilakukan pengecekan keberadaan nilai yang hilang, dan jika ditemukan, metode imputasi yang sesuai diterapkan untuk menjaga kualitas data. Setelah tahap pra-pemrosesan awal, data dipisahkan menjadi variabel fitur (X) dan variabel target (y) [21]. Dataset kemudian dibagi menjadi data pelatihan dan pengujian dengan rasio 80:20 menggunakan teknik pemisahan pelatihan-pengujian berstratifikasi, sehingga proporsi kelas dipertahankan di kedua subset [22]. Untuk model Support Vector Machine (SVM), fitur numerik dalam data pelatihan distandarisasi menggunakan metode *StandardScaler*, setelah itu parameter standarisasi yang dihasilkan diterapkan pada data uji tanpa *penyesuaian ulang*, untuk mencegah *kebocoran data* [15]. Tahap selanjutnya adalah pemilihan fitur menggunakan metode *SelectKBest* dengan fungsi penilaian [23]. Berbeda dari pendekatan yang menentukan jumlah fitur (k) satu kali melalui validasi silang terpisah lalu memfiksasinya untuk seluruh algoritma, jumlah fitur (k) pada *SelectKBest* dalam penelitian ini diperlakukan sebagai hyperparameter yang dituning secara nested bersama hyperparameter masing-masing model di dalam *GridSearchCV*. *SelectKBest*, *SMOTE*, dan (khusus untuk SVM) *StandardScaler* dibungkus dalam satu *imblearn.Pipeline* yang di-fit ulang pada setiap fold validasi silang (*Stratified 5-Fold Cross-Validation*), sehingga tidak ada informasi dari fold validasi yang bocor ke proses seleksi fitur (skema nested cross-validation). Konsekuensinya, nilai k optimal dapat berbeda-beda untuk setiap algoritma, karena masing-masing dituning secara independen dalam pipeline-nya sendiri [15][24]. Setelah pemilihan fitur, distribusi kelas dalam data pelatihan diseimbangkan menggunakan metode *Synthetic Minority Over-sampling Technique* (SMOTE) [25]. SMOTE diterapkan secara terpisah pada data standar untuk model SVM dan pada data non-standar untuk model Random Forest dan XGBoost [14], sedangkan data uji disimpan dalam keadaan aslinya yang tidak seimbang sehingga evaluasi model akan mencerminkan kondisi data dunia nyata. Penyetelan *hyperparameter* (penyetelan hyperparameter) kemudian dilakukan untuk setiap algoritma menggunakan *GridSearchCV* dengan skema *Validasi Silang 5-Fold Bertingkat* dan metrik F1 Score sebagai tujuan optimasi [26]. Model terbaik yang diperoleh dari proses penyetelan kemudian dilatih ulang menggunakan kombinasi *hyperparameter* optimal sebelum dievaluasi pada data uji independen [27]. Kinerja model dievaluasi menggunakan beberapa metrik, yaitu akurasi, presisi, recall, F1 Score, dan *Receiver Operating Characteristic–Area Under the Curve* (ROC-AUC), yang memberikan gambaran komprehensif tentang kinerja klasifikasi setiap algoritma [28]. Sebagai tahap akhir, analisis *kepentingan fitur* dilakukan pada model berbasis pohon, yaitu Random Forest dan XGBoost, untuk mengidentifikasi gen dengan kontribusi prediktif tertinggi terhadap klasifikasi penyakit Huntington [12].

2.2 Data dan Sumber Data

Data yang digunakan dalam penelitian ini diperoleh dari basis data publik *Gene Expression Omnibus* (GEO) yang dikelola oleh *National Center for Biotechnology Information* (NCBI), dengan nomor akses GSE135589, yang berisi profil ekspresi gen turunan *microarray* yang terkait dengan penyakit *Huntington* [29]. Pembacaan dan pemeriksaan langsung file dataset (GSE135589.csv), yang telah menjalani tahap pra-pemrosesan awal yang siap untuk ML, menghasilkan 178 sampel (baris data) dengan 54.675 fitur ekspresi gen (kolom *probe microarray*) setelah menghapus kolom pengidentifikasi sampel dan label. Variabel target biner, *Label Encoded*, menunjukkan distribusi 134 sampel dalam kelas mayoritas dan 44 sampel dalam kelas minoritas, memberikan rasio ketidakseimbangan kelas sekitar 3:1. Tidak ditemukan nilai yang hilang dalam dataset ini. Karakteristik inilah, yaitu jumlah fitur yang jauh melebihi jumlah sampel ($p \gg n$), yang mendasari perlunya pemilihan fitur dan skema validasi anti-kebocoran yang diterapkan secara ketat pada tahap metodologi [30]. Dataset dibagi menjadi data pelatihan (142 sampel) dan data uji (36 sampel) menggunakan pembagian stratifikasi 80:20, sehingga proporsi kelas di kedua subset mewakili proporsi kelas dalam data keseluruhan. Distribusi kelas untuk penelitian ini disajikan pada Tabel 1.

Tabel 1. Distribusi kelas di seluruh data keseluruhan, data pelatihan, dan data pengujian.

Subset	Kelas Mayoritas	Kelas Minoritas	Total
Data keseluruhan	134	44	178
Data pelatihan (80%)	107	35	142
Data uji (20%)	27	9	36

2.3 Rumus Matematika

Bagian ini menjelaskan dasar matematis dari metode pemilihan fitur, penyeimbangan kelas, tiga algoritma klasifikasi, dan metrik evaluasi yang digunakan.

2.3.1 Pemilihan Fitur (Uji F ANOVA)

SelectKBest dengan skor $f_classif$ mengevaluasi setiap fitur secara independen menggunakan statistik uji F *ANOVA* satu arah, yang membandingkan varians antar kelas dengan varians dalam kelas [31]:

$$F = \frac{MS_{Between}}{MS_{Within}} \quad (1)$$

Dengan

$$MS_{Between} = \frac{\sum_{i=1}^K n_i (\bar{X}_i - \bar{X})^2}{K-1} \quad (2)$$

Dan

$$MS_{Within} = \frac{\sum_{i=1}^K \sum_{j=1}^{n_i} (X_{ij} - \bar{X}_i)^2}{N-K} \quad (3)$$

di mana K adalah jumlah kelas, n_i adalah jumlah sampel dalam kelas i , $\bar{X}_i$ adalah rata-rata fitur untuk kelas i , $\bar{X}$ adalah rata-rata fitur keseluruhan, dan N adalah jumlah total sampel. Fitur dengan nilai F tertinggi (asosiasi terkuat dengan label kelas) dipilih sebagai k fitur teratas. Nilai optimal k tidak difiksasi lebih dahulu, melainkan diperlakukan sebagai hyperparameter (`select_k`) yang dituning secara nested bersama hyperparameter setiap model klasifikasi di dalam `GridSearchCV`, sehingga `SelectKBest` di-fit ulang pada data pelatihan setiap fold dan k optimal dapat berbeda antar algoritma [32].

2.3.2 SMOTE (Synthetic Minority Over-sampling Technique)

SMOTE menyeimbangkan kelas minoritas dengan menghasilkan sampel sintetis melalui interpolasi linier antara sampel minoritas dan salah satu dari k tetangga terdekatnya di ruang fitur:

$$x_{baru} = x_i + \lambda(x_{nn} - x_i) \quad (4)$$

di mana x_i adalah sampel kelas minoritas yang dipilih, x_{nn} adalah salah satu tetangga terdekatnya (dipilih dari k tetangga terdekat), dan λ adalah angka acak yang diambil dari distribusi seragam pada interval $[0, 1]$. Proses ini diulang sampai jumlah sampel kelas minoritas sama dengan kelas mayoritas, dan hanya diterapkan pada data pelatihan [33].

2.3.3 Support Vector Machine (SVM)

SVM mencari *hyperplane* pemisah optimal yang memaksimalkan margin antar kelas. Untuk kasus pemisahan *non-linier* dengan margin lunak, formulasi optimasinya adalah:

$$\begin{aligned} \min_{w, b, \xi} \quad & \frac{1}{2} \|w\|^2 + C \sum_{i=1}^n \xi_i \\ \text{s.t.} \quad & y_i(w^T \phi(x_i) + b) \geq 1 - \xi_i, \\ & \xi_i \geq 0 \end{aligned} \quad (5)$$

di mana w adalah vektor bobot *hyperplane*, b adalah bias, ξ_i adalah variabel slack yang mentolerir kesalahan klasifikasi, C adalah parameter regularisasi yang mengontrol *trade-off* antara lebar margin dan kesalahan klasifikasi, dan $\phi(x)$ adalah fungsi pemetaan ke ruang berdimensi lebih tinggi. Untuk kernel *Radial Basis Function* (RBF), fungsi kernelnya adalah:

$$K(x_i, x_j) = \exp(-\gamma \|x_i - x_j\|^2) \quad (6)$$

di mana γ mengontrol pengaruh jarak antar sampel pada nilai kernel, sedangkan kernel linier menggunakan $K(x_i, x_j) = x_i \cdot x_j$.

2.3.4 Random Forest

Random Forest membangun sejumlah pohon keputusan pada subset data dan fitur acak (*bagging*), kemudian menggabungkan prediksinya melalui pemungutan suara mayoritas. Setiap pemisahan *node* dalam pohon keputusan dipilih untuk meminimalkan ketidakmurnian *Gini* [34]:

$$Gini(t) = 1 - \sum_k p_k^2 \quad (7)$$

di mana p_k adalah proporsi sampel kelas- k pada node t . Prediksi klasifikasi *Random Forest* akhir adalah hasil dari pemungutan suara mayoritas di seluruh T pohon:

$$\hat{y} = \text{mode}\{h_1(x), h_2(x) \dots h_T(x)\} \quad (8)$$

2.3.5 XGBoost (Extreme Gradient Boosting)

XGBoost membangun modelnya secara aditif dan iteratif, di mana setiap pohon baru dilatih untuk mengoreksi residual/kesalahan prediksi dari model sebelumnya, menggunakan fungsi tujuan yang diregulasi:

$$Obj^{(t)} = \sum_{i=1}^n l(y_i, \hat{y}_i^{(t-1)} + f_t(x_i)) + \Omega(f_t) \quad (9)$$

$$\Omega(f) = \gamma T + \frac{1}{2} \lambda \sum_{j=1}^T w_j^2 \quad (10)$$

di mana l adalah fungsi kerugian (misalnya, *log-loss* untuk klasifikasi biner), $\hat{y}_i = \sum_k f_k(x_i)$ adalah prediksi kumulatif dari K pohon, $\Omega(f)$ adalah istilah regularisasi yang memberikan penalti pada kompleksitas pohon (T = jumlah daun, w = bobot daun), dan γ dan λ adalah parameter regularisasi. Model diperbarui secara iteratif mengikuti arah *gradien* fungsi kerugian (*gradient boosting*), sehingga setiap pohon baru berfokus pada koreksi kesalahan pohon sebelumnya[35].

2.3.6 Metrik Evaluasi

Performa model dievaluasi menggunakan lima metrik klasifikasi biner standar, yang dihitung dari *True Positive (TP)*, *True Negative (TN)*, *False Positive (FP)*, dan *False Negative (FN)*:

$$\text{Accuracy} = \frac{TP+TN}{TP+TN+FP+FN} \quad (11)$$

$$\text{Precision} = \frac{TP}{TP+FP} \quad (12)$$

$$\text{Recall} = \frac{TP}{TP+FN} \quad (13)$$

$$F1 = 2 \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \quad (14)$$

$$TPR = \frac{TP}{TP+FN} \quad (15)$$

$$FPR = \frac{FP}{FP+TN} \quad (16)$$

$$ROC = TPR \text{ terhadap } FPR \quad (17)$$

$$AUC = \int_0^1 TPR(FPR) d(FPR) \quad (18)$$

Akurasi menunjukkan proporsi semua prediksi yang diklasifikasikan dengan benar relatif terhadap jumlah total sampel. Presisi mengukur akurasi model dalam mengidentifikasi sampel positif, sehingga presisi yang lebih tinggi berarti lebih sedikit positif palsu. Recall mengukur kemampuan model untuk menemukan semua sampel positif dan oleh karena itu penting dalam sistem deteksi penyakit, karena berkaitan dengan kemungkinan terlewatnya pasien (negatif palsu). F1 Score adalah rata-rata harmonik dari presisi dan recall, memberikan ukuran evaluasi yang lebih seimbang, khususnya untuk dataset dengan distribusi kelas yang tidak seimbang[36]. ROC menggambarkan hubungan antara Tingkat Positif Sejati dan Tingkat Positif Palsu di berbagai ambang batas keputusan, sedangkan AUC mengukur kemampuan keseluruhan model untuk membedakan antara kedua kelas. Nilai AUC yang mendekati 1 menunjukkan kemampuan diskriminatif yang lebih baik[37].

Mengingat dataset penelitian ini memiliki ketidakseimbangan kelas (rasio sekitar 3:1), akurasi standar berisiko memberikan gambaran kinerja yang terlalu optimis karena didominasi oleh kelas mayoritas. Oleh karena itu, penelitian ini turut melaporkan *Balanced Accuracy* sebagai metrik pelengkap yang lebih tahan terhadap ketidakseimbangan kelas, didefinisikan sebagai rata-rata dari recall pada tiap kelas:

$$\text{Balanced Accuracy} = \frac{1}{2} (\text{Recall}_{\text{kelas } 0} + \text{Recall}_{\text{kelas } 1}) \quad (19)$$

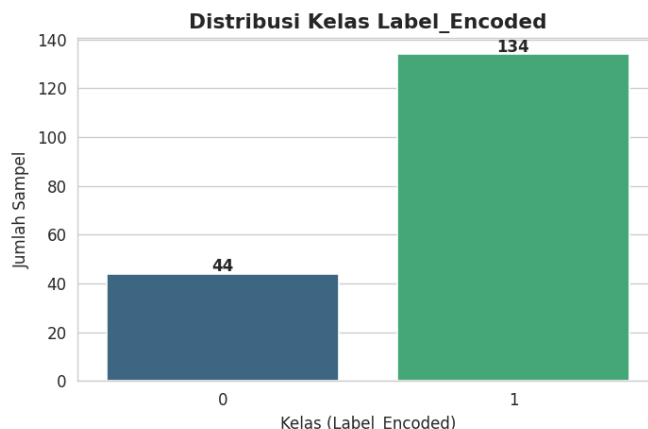
Berbeda dari akurasi standar yang dapat tetap tinggi meskipun model hanya mengenali kelas mayoritas dengan baik, *Balanced Accuracy* memberi bobot yang setara pada kemampuan model mengenali kelas mayoritas maupun minoritas, sehingga lebih representatif untuk menjustifikasi klaim keberhasilan model pada dataset dengan distribusi kelas yang tidak seimbang seperti pada penelitian ini.

3. HASIL DAN PEMBAHASAN

3.1 Karakteristik Dataset dan Pembagian Data

Dataset GSE135589 yang digunakan dalam penelitian ini terdiri dari 178 sampel ekspresi gen dengan 54.675 fitur (set probe), tanpa nilai yang hilang. Distribusi kelas tidak seimbang, dengan 134 sampel dalam kelas mayoritas (HD, *Label_Encoded* = 1) dan 44 sampel dalam kelas minoritas (Kontrol, *Label_Encoded* = 0), memberikan rasio sekitar 3:1. Kondisi di mana jumlah fitur jauh melebihi jumlah sampel ($p \gg n$) adalah karakteristik khas data microarray dan memiliki implikasi langsung terhadap risiko overfitting jika tidak ditangani dengan skema validasi yang ketat.

Dataset dibagi menggunakan pembagian bertingkat, dengan 80% data pelatihan (142 sampel: 107 kelas mayoritas, 35 kelas minoritas) dan 20% data uji (36 sampel: 27 kelas mayoritas, 9 kelas minoritas). Ukuran data uji yang relatif kecil, khususnya untuk kelas minoritas (9 sampel), merupakan poin penting yang perlu diperhatikan ketika menafsirkan stabilitas metrik evaluasi pada tahap selanjutnya. Distribusi kelas *Label_Encoded* untuk penelitian ini disajikan pada Gambar 2.

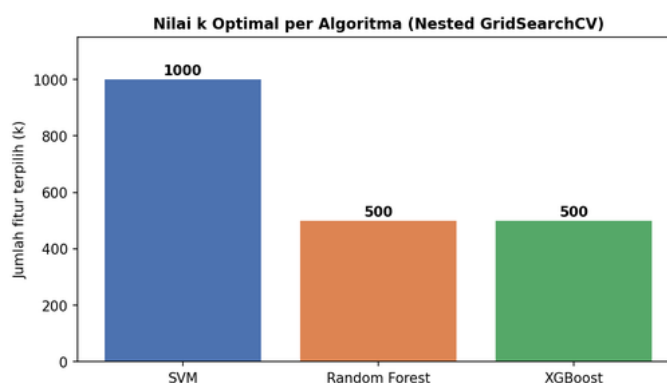


Gambar 2. Distribusi kelas Label_Encoded (0 = Kontrol, 1 = HD) dalam dataset GSE135589

3.2 Pemilihan Fitur

Pemilihan fitur dilakukan menggunakan SelectKBest dengan skor uji F ANOVA ($f_classif$). jumlah fitur (k) tidak lagi dipilih satu kali secara terpisah menggunakan satu estimator referensi lalu difiksasi untuk seluruh algoritma. Sebagai gantinya, k diperlakukan sebagai hyperparameter (`select_k`) dengan kandidat $k \in \{100, 300, 500, 1000\}$, dituning bersamaan dengan hyperparameter setiap model (SVM, Random Forest, XGBoost) di dalam GridSearchCV berskema *Stratified 5-Fold Cross-Validation* yang nested terhadap SelectKBest, SMOTE, dan (khusus SVM) StandardScaler seluruh tiga langkah tersebut di-fit ulang pada setiap fold pelatihan. Dengan skema ini, nilai k optimal untuk SVM, Random Forest, dan XGBoost ditentukan secara independen dan ternyata berbeda: $k = 1000$ untuk SVM, sedangkan Random Forest dan XGBoost sama-sama memilih $k = 500$. Karena k kini dituning bersamaan dengan hyperparameter lain di dalam satu proses pencarian gabungan, tidak ada lagi tabel akurasi CV tunggal per nilai k yang berlaku untuk seluruh model seperti pada draf sebelumnya; skor F1 hasil tuning gabungan (k dan hyperparameter model) untuk masing-masing algoritma dilaporkan pada Bagian 3.4

Pola umum bahwa penambahan fitur jauh di luar titik optimal cenderung menimbulkan *noise* yang tidak informatif dan dapat mengurangi kemampuan generalisasi model tetap relevan secara konseptual. Pengurangan dimensi fitur ini penting untuk mengurangi risiko *overfitting* yang timbul dari kondisi $p \gg n$. Pengurangan dimensi menjadi 500 fitur tetap menjadi pilihan optimal untuk Random Forest dan XGBoost, sedangkan SVM justru memperoleh performa CV terbaik pada dimensi yang lebih besar ($k = 1000$), yang konsisten dengan sifat SVM linier yang lebih toleran terhadap fitur berdimensi tinggi karena regularisasi eksplisit melalui parameter C , dibandingkan model berbasis pohon yang lebih rentan terhadap peningkatan varians ketika jumlah fitur bertambah. Nilai k optimal hasil tuning nested untuk masing-masing algoritma ditunjukkan pada Gambar 3, yang menggambarkan perbedaan preferensi dimensi fitur antara SVM dan model berbasis pohon.



Gambar 3. Pemilihan Fitur

3.3 Menangani Ketidakseimbangan Dengan SMOTE

Untuk mengatasi ketidakseimbangan kelas, SMOTE diterapkan secara eksklusif pada data pelatihan (setelah pemilihan fitur), menyeimbangkan distribusi kelas dalam data pelatihan dari 35:107 menjadi 107:107 sampel. Data uji dipertahankan dalam distribusi aslinya (9:27) sehingga evaluasi akan mencerminkan kondisi dunia nyata yang tidak seimbang. Penerapan SMOTE hanya pada data pelatihan merupakan langkah penting untuk mencegah kebocoran informasi (kebocoran data) dari data uji ke dalam proses pelatihan. Distribusi kelas sebelum dan sesudah SMOTE untuk penelitian ini disajikan pada Gambar 4.

```
=====
SMOTE (Synthetic Minority Over-sampling Technique)
=====
Distribusi kelas SEBELUM SMOTE (data training):
Label_Encoded
0    35
1   107
Name: count, dtype: int64

Distribusi kelas SESUDAH SMOTE (data training):
Label_Encoded
0    107
1    107
Name: count, dtype: int64
```

Gambar 4. Distribusi kelas sebelum dan setelah *SMOTE*

3.4 Penyetelan Hyperparameter (GridSearchCV)

Penyetelan hyperparameter dilakukan menggunakan *GridSearchCV* dengan skema validasi silang bertingkat 5-fold dan metrik optimasi F1 Score pada data pelatihan. ruang pencarian (param_grid) untuk setiap model kini turut menyertakan select_k (kandidat 100/300/500/1000) sebagai bagian dari hyperparameter yang dituning bersamaan di dalam pipeline nested (scaler→select→SMOTE→classifier untuk SVM; select→SMOTE→classifier untuk Random Forest dan XGBoost). Parameter terbaik yang diperoleh untuk setiap model, hasil pencarian nested yang sebenarnya, adalah sebagai berikut: SVM dengan kernel linier, C = 0,1, gamma = 'scale', dan k = 1000 (F1 Score CV Terbaik = 84,42%); Random Forest dengan n_estimators = 300, max_depth = None, min_samples_split = 5, dan k = 500 (F1 Score CV Terbaik = 81,62%); dan XGBoost dengan n_estimators = 100, learning_rate = 0,1, max_depth = 5, subsample = 0,8, dan k = 500 (F1 Score CV Terbaik = 85,28%). Menariknya, parameter Random Forest dan XGBoost identik dengan yang diperoleh pada draf sebelumnya karena kebetulan kedua model tersebut tetap memilih k = 500 meski k kini dituning secara nested; SVM-lah yang berubah, dari k = 500 (draf sebelumnya) menjadi k = 1000 pada skema nested yang benar. Kernel linier tetap terpilih sebagai parameter SVM terbaik, dan ini tetap informatif: dalam kondisi $p \gg n$, ruang fitur berdimensi tinggi cenderung membuat data lebih mudah dipisahkan secara linier, sehingga model yang lebih sederhana sebenarnya lebih disukai oleh proses penyetelan karena kinerjanya lebih stabil di bawah validasi silang. Perbandingan kinerja validasi silang untuk penelitian ini disajikan pada Tabel 2.

Tabel 2. Perbandingan kinerja validasi silang

Model	Best F1 (CV Tuning)	CV Accuracy (Mean)	CV Std	Perbedaan vs. Akurasi Tes.
SVM	84,42%	74,58%	±0,0633	-0,1431
Random Forest	81,62%	72,44%	±0,1154	0,0300
XGBoost	85,28%	75,99%	±0,0706	0,0099

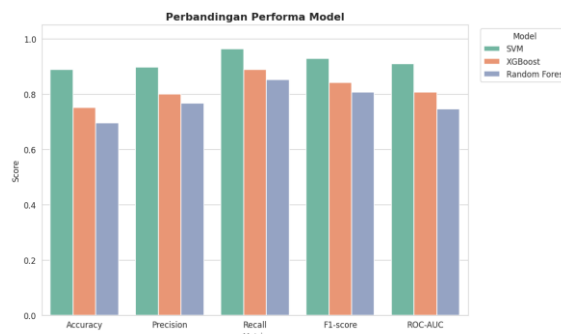
3.5 Evaluasi Model pada Data Uji dan Perbandingan Antar-Model

Evaluasi akhir dilakukan pada data uji (36 sampel), yang tidak pernah dilihat oleh model selama proses pelatihan maupun penyetelan hyperparameter. Hasil evaluasi untuk kelima metrik untuk ketiga model tersebut disajikan dalam Tabel 3.

Tabel 3. Perbandingan kinerja SVM, Random Forest, dan XGBoost pada data uji

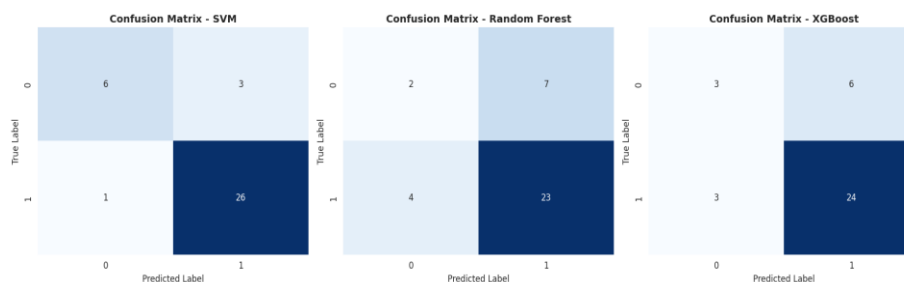
Model	Accuracy	Precision	Recall	F1-score	ROC-AUC	Balanced Accuracy
SVM	88,89%	89,66%	96,30%	92,86%	90,95%	81,48%
XGBoost	75,00%	80,00%	88,89%	84,21%	80,66%	61,11%
Random Forest	69,44%	76,67%	85,19%	80,70%	74,49%	53,70%

Berdasarkan Tabel 3, SVM mencatat kinerja terbaik pada data uji, dengan akurasi 88,89%, presisi 89,66%, recall 96,30%, F1-score 92,86%, dan ROC-AUC 90,95%, diikuti oleh XGBoost (akurasi 75,00%; ROC-AUC 80,66%) dan Random Forest sebagai model dengan kinerja terendah (akurasi 69,44%; ROC-AUC 74,49%). Keunggulan SVM linier ini konsisten dengan karakteristik data ekspresi gen berdimensi tinggi: dalam ruang fitur berdimensi tinggi, data cenderung lebih mudah dipisahkan secara linier, sehingga model linier seperti SVM seringkali lebih stabil dan kurang rentan terhadap overfitting dibandingkan model berbasis pohon, yang mempelajari batas keputusan non-linier yang lebih kompleks dalam ruang fitur yang sama. Karena akurasi standar dapat menyembunyikan kinerja yang buruk pada kelas minoritas ketika distribusi kelas tidak seimbang, Tabel 3 juga melaporkan Balanced Accuracy sebagai metrik pembandingan yang lebih tahan terhadap bias kelas mayoritas. Ketika dilihat melalui Balanced Accuracy, urutan peringkat ketiga model tetap konsisten, namun kesenjangan antar-model justru semakin tegas: SVM mencapai Balanced Accuracy 81,48%, jauh lebih tinggi daripada XGBoost (61,11%) dan Random Forest (53,70%). Nilai Balanced Accuracy Random Forest yang mendekati 0,50 (setara tebakan acak jika kedua kelas diberi bobot setara) menunjukkan bahwa akurasi standarnya (69,44%) sebagian besar ditopang oleh kinerja pada kelas mayoritas (HD), bukan oleh kemampuan diskriminatif model yang sesungguhnya seimbang antar kelas. Temuan ini memperkuat, dan bukan sekadar mengulang, klaim keunggulan SVM linier yang telah dikemukakan pada bagian Pembahasan: keunggulan tersebut tetap konsisten meskipun dievaluasi dengan metrik yang secara eksplisit dirancang untuk tidak bias terhadap kelas mayoritas.



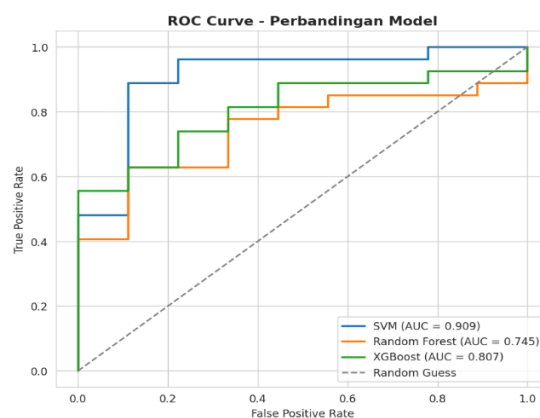
Gambar 5. Perbandingan kinerja SVM, Random Forest, dan XGBoost pada data uji (Akurasi, Presisi, Recall, F1-score, ROC-AUC)

Gambar 5 menyajikan perbandingan kinerja dari tiga model klasifikasi (SVM, XGBoost, dan Random Forest) berdasarkan lima metrik evaluasi akurasi, presisi, recall, F1-score, dan ROC-AUC yang diukur pada data uji. Secara keseluruhan, model SVM kernel linier ($C = 0,1$) menunjukkan kinerja terbaik di hampir semua metrik, dengan akurasi 88,89%, F1-score 92,86%, dan ROC-AUC 90,95%. Model XGBoost berada di peringkat kedua dengan kinerja yang cukup kompetitif (akurasi 75,00%, F1-score 84,21%, ROC-AUC 80,66%), sementara Random Forest secara konsisten menunjukkan kinerja terendah di semua metrik, dengan akurasi hanya 69,44% dan ROC-AUC 74,49%. Pola ini menunjukkan bahwa SVM linier lebih mampu menangani karakteristik data ekspresi gen berdimensi tinggi (1000 fitur untuk SVM, 500 fitur untuk Random Forest dan XGBoost, dipilih dari 54.675 fitur asli) dibandingkan model ensemble pohon dalam kasus ini.



Gambar 6. Confusion Matrix dari ketiga model pada data uji

Gambar 6 menunjukkan Confusion Matrix dari ketiga model pada data uji (36 sampel, terdiri dari 9 sampel kelas Kontrol dan 27 sampel kelas HD). Dengan pipeline nested, Model SVM mengklasifikasikan dengan benar 26 dari 27 sampel kelas-1 (HD) dan 6 dari 9 sampel kelas-0 (Kontrol), menghasilkan kesalahan klasifikasi paling sedikit (4 dari 36 sampel) di antara ketiga model membaik dari 6 kesalahan pada draft sebelumnya karena SVM kini memakai $k = 1000$. Random Forest dan XGBoost, yang parameternya (termasuk $k = 500$) kebetulan tidak berubah dari draft sebelumnya, tetap menunjukkan pola yang sama: Random Forest memiliki jumlah positif palsu tertinggi untuk kelas 0, dengan 7 dari 9 sampel Kontrol salah diklasifikasikan sebagai HD, yang berkontribusi pada presisi dan akurasi yang lebih rendah. XGBoost berada di antara keduanya, dengan 6 dari 9 sampel Kontrol salah diklasifikasikan sementara masih mengenali dengan benar 24 dari 27 sampel HD. Secara keseluruhan, ketiga model masih cenderung lebih baik dalam mendeteksi kelas mayoritas (HD) daripada kelas minoritas (Kontrol), meskipun kesenjangan tersebut kini paling kecil pada SVM.



Gambar 7. Kurva ROC dan nilai AUC untuk SVM, Random Forest, dan XGBoost pada data uji

Gambar 7 menunjukkan kurva ROC (*Receiver Operating Characteristic*) dari ketiga model beserta nilai *Area Under the Curve* (AUC) masing-masing. Kurva SVM (AUC = 0,909) terletak paling dekat dengan sudut kiri atas plot dan secara konsisten berada di atas kurva dua model lainnya di sebagian besar rentang tingkat *false-positive*, menunjukkan kemampuan diskriminasi kelas terkuat di antara ketiga model tersebut. XGBoost berada di peringkat kedua dengan AUC 0,807, sedangkan Random Forest memiliki AUC terendah (0,745), dengan kurvanya terletak paling dekat dengan garis referensi diagonal (tebakan acak), menunjukkan kemampuan diskriminatif yang relatif lebih lemah. Ketiga kurva tersebut tetapi berada di atas garis referensi diagonal, menunjukkan bahwa semua model masih memiliki kemampuan prediksi yang lebih baik daripada tebakan acak, meskipun dengan tingkat keandalan yang berbeda.

3.6 Kesenjangan antara Validasi Silang dan Kinerja Data Uji

Kesenjangan antara akurasi validasi silang dan akurasi data uji pada draf sebelum penerapan nested cross-validation, yang berkisar antara 0,93 dan 0,98 untuk CV berbanding sekitar 0,83–0,89 pada data uji, merupakan indikasi bahwa estimasi kinerja pada skema sebelumnya bersifat terlalu optimis. Tabel 2 menunjukkan bahwa perbaikan metodologis pada versi ini secara langsung menyempitkan kesenjangan tersebut. Temuan penting yang justru menegaskan keberhasilan perbaikan metodologis pada versi ini, sebagaimana ditunjukkan pada Tabel 2, adalah bahwa setelah dipilih secara nested di dalam GridSearchCV, kesenjangan antara akurasi validasi silang dan akurasi data uji menyempit drastis dibandingkan draf sebelumnya bahkan berbalik arah untuk SVM. Akurasi CV kini berkisar 0,72–0,76 untuk ketiga model (SVM = $74,58\% \pm 0,0633$; Random Forest = $72,44\% \pm 0,1154$; XGBoost = $75,99\% \pm 0,0706$), jauh lebih rendah daripada angka CV pada draf sebelumnya (0,93–0,98) yang dihasilkan dari skema seleksi fitur di luar loop CV. Untuk SVM, akurasi CV (74,58%) justru lebih rendah daripada akurasi pada data uji (88,89%), menghasilkan selisih negatif (-0,1431); untuk Random Forest dan XGBoost, selisih CV terhadap uji menyusut menjadi hampir nol (masing-masing 0,0300 dan 0,0099). Pola ini konsisten dengan ekspektasi teoretis nested cross-validation: karena SelectKBest, SMOTE, dan (untuk SVM) StandardScaler kini di-fit ulang pada setiap fold, skor CV yang dilaporkan tidak lagi memiliki akses tidak sengaja ke informasi dari fold validasi, sehingga tidak lagi optimis secara sistematis. Sebaliknya, angka CV yang jauh lebih tinggi pada draf sebelumnya kini dapat dipahami sebagai bukti empiris dari bias seleksi fitur yang telah diperingatkan dalam literatur terkait kebocoran data [15],[24]: ketika SelectKBest dievaluasi menggunakan seluruh data pelatihan sebelum loop CV berjalan, setiap fold validasi pada skema lama turut memberi informasi pada pemilihan fitur, sehingga skor CV menjadi terlalu optimis dan tidak mencerminkan kemampuan generalisasi yang sebenarnya. Prosedur nested pada versi ini memperbaiki kelemahan metodologis tersebut secara langsung, meskipun rasio fitur-ke-sampel yang masih besar (hingga 1000 fitur untuk 142 sampel pelatihan pada SVM) dan ukuran data uji yang relatif kecil (36 sampel) tetap membatasi presisi *estimasi kinerja* secara umum.

Kesenjangan kecil yang masih tersisa pada Random Forest dan XGBoost, serta pola CV yang justru lebih rendah daripada uji pada SVM, sebagian dapat dijelaskan oleh variabilitas estimasi akibat ukuran sampel pelatihan yang kecil (142 sampel) relatif terhadap dimensi fitur yang dipilih (500–1000 fitur), serta oleh efek SMOTE: sampel sintesis dalam data pelatihan dapat membuat estimasi performa di dalam CV sedikit lebih bervariasi antar fold dibandingkan performa pada 36 sampel data uji tunggal yang tetap tidak seimbang dan terdiri dari sampel asli (*non-sintetis*). *Random Forest* tetap menunjukkan variabilitas CV tertinggi di antara ketiga model (std = 0,1154), konsisten dengan kecenderungan model ensemble pohon untuk lebih sensitif terhadap komposisi sampel dan lebih rentan mengalami *overfitting* pada data berdimensi tinggi dengan ukuran sampel terbatas, meskipun parameter *min_samples_split* diatur untuk membatasi kompleksitas pohon.

3.7 Kinerja Per Kelas dan Bias terhadap Kelas Mayoritas

Rincian presisi, *recall*, dan *F1-score* per kelas untuk setiap model disajikan dalam Tabel 4 hingga Tabel 6.

Tabel 4. Kinerja SVM per kelas

Kelas	Precision	Recall	F1 Score	Dukungan (n)
0 (Kontrol)	0,86	0,67	0,75	9
1 (HD)	0,90	0,96	0,93	27

Tabel 5. Kinerja Random Forest per kelas

Kelas	Precision	Recall	F1 Score	Dukungan (n)
0 (Kontrol)	0,33	0,22	0,27	9
1 (HD)	0,77	0,85	0,81	27

Tabel 6. Kinerja XGBoost per kelas

Kelas	Precision	Recall	F1 Score	Dukungan (n)
0 (Kontrol)	0,50	0,33	0,40	9
1 (HD)	0,80	0,89	0,84	27

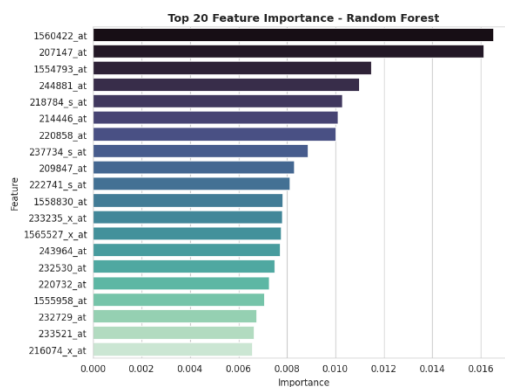
SVM mencapai recall kelas Kontrol tertinggi di antara ketiga model (0,67; 6 dari 9 sampel Kontrol diidentifikasi dengan benar), diikuti oleh XGBoost (0,33; 3 dari 9) dan Random Forest dengan nilai terendah (0,22; 2 dari 9). Sebaliknya, recall kelas HD tinggi di ketiga model (0,85–0,96), menunjukkan bahwa ketiga model tersebut relatif andal dalam mengenali sampel HD tetapi cenderung bias terhadap kelas mayoritas ketika mengklasifikasikan sampel Kontrol.

Pola ini merupakan temuan klinis penting: meskipun SMOTE diterapkan pada data pelatihan, hasil data uji masih menunjukkan bias terhadap kelas mayoritas di ketiga model tersebut. Dalam konteks skrining atau diagnostik, kegagalan mengenali sampel Kontrol (mengakibatkan positif palsu untuk HD) memiliki implikasi yang berbeda dari kegagalan mengenali sampel HD (negatif palsu), artinya kinerja model saat ini belum cukup andal untuk secara konsisten membedakan kelas Kontrol. Keterbatasan ini kemungkinan besar dipengaruhi oleh jumlah sampel kelas minoritas yang sangat kecil dalam data uji (hanya 9 sampel), sehingga setiap kesalahan klasifikasi memiliki dampak besar pada metrik recall dan presisi untuk kelas tersebut.

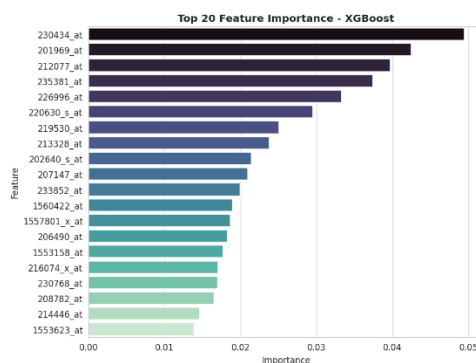
Bias terhadap kelas mayoritas ini konsisten dengan nilai Balanced Accuracy yang dilaporkan pada Bagian 3.5 (Tabel 3): SVM mencapai 0,8148, sedangkan XGBoost dan Random Forest masing-masing hanya 0,6111 dan 0,5370, jauh di bawah akurasi standarnya (75,00% dan 0,6944). Kesenjangan ini menunjukkan secara kuantitatif bahwa klaim keberhasilan model, khususnya untuk Random Forest dan XGBoost, tidak dapat disandarkan pada akurasi standar semata, karena metrik tersebut secara sistematis diuntungkan oleh dominasi kelas HD dalam data uji. Sebagai pelengkap Balanced Accuracy, kalibrasi probabilitas (misalnya melalui Platt scaling atau isotonic regression) berpotensi memperbaiki keandalan skor probabilitas yang dihasilkan ketiga model, khususnya untuk kelas Kontrol yang under-represented, namun penerapan dan evaluasinya (misalnya melalui reliability diagram atau Brier score) memerlukan pelatihan ulang model pada skema nested cross-validation yang berbeda dan diposisikan sebagai agenda penelitian lanjutan (lihat Bagian 3.9 dan 3.10), agar tidak menimbulkan kebocoran data baru jika dipaksakan pada data uji yang sama.

3.8 Implikasi Praktis dan Relevansi Biologis

Pengurangan dimensi dari 54.675 menjadi 500 fitur melalui *SelectKBest* menunjukkan bahwa sebagian besar set probe microarray tidak berkontribusi secara signifikan dalam membedakan HD dari Kontrol, dan bahwa sinyal ekspresi gen yang relevan terkonsentrasi pada subset fitur yang jauh lebih kecil. Mengidentifikasi gen spesifik yang paling berkontribusi pada prediksi (melalui Random Forest dan pentingnya fitur *XGBoost*) dapat memberikan wawasan tambahan tentang jalur biologis yang terlibat dalam patogenesis penyakit *Huntington*, dan direkomendasikan untuk pemeriksaan dan konfirmasi lebih lanjut melalui pemetaan probe ke simbol gen, tinjauan literatur biologi molekuler, serta studi validasi independen, sebelum dapat diinterpretasikan sebagai kandidat biomarker; daftar gen pada tahap ini bersifat eksploratif, bukan hasil validasi biologis.



Gambar 8. 20 fitur terpenting untuk model Random Forest



Gambar 9. 20 fitur terpenting untuk model XGBoost

3.9 Keterbatasan Studi

Studi ini memiliki beberapa keterbatasan yang perlu dipertimbangkan saat menafsirkan hasilnya. Pertama, ukuran sampel yang relatif kecil (178 sampel, dengan hanya 36 dalam data uji dan 9 di antaranya dalam kelas minoritas) menyebabkan estimasi metrik evaluasi memiliki varians yang tinggi dan rentan terhadap fluktuasi akibat perubahan kecil dalam komposisi data uji. Kedua, kesenjangan kinerja antara validasi silang dan data uji menunjukkan bahwa hasil ini perlu divalidasi lebih lanjut menggunakan skema validasi silang bertingkat atau kohort independen (dataset eksternal) untuk mengkonfirmasi generalisasi model. Ketiga, penerapan SMOTE pada data berdimensi tinggi berisiko menghasilkan sampel sintetis yang tidak sepenuhnya mewakili variasi biologis yang sebenarnya, sehingga pendekatan penyeimbangan kelas alternatif (misalnya, pembobotan kelas, pembelajaran sensitif biaya, atau kalibrasi ambang keputusan/threshold-moving yang belakangan terbukti lebih andal dan lebih murah secara komputasi dibandingkan augmentasi sintetis pada beberapa studi [38]) dapat dipertimbangkan sebagai perbandingan dalam penelitian selanjutnya. Keempat, penelitian ini menggunakan skor probabilitas mentah dari setiap model tanpa menerapkan kalibrasi probabilitas (misalnya Platt scaling atau isotonic regression). Meskipun Balanced Accuracy telah dilaporkan (Bagian 3.5 dan 3.7) sebagai metrik ringkasan yang tahan terhadap ketidakseimbangan kelas untuk menjustifikasi klaim keberhasilan model, kalibrasi probabilitas tetap relevan sebagai langkah tambahan untuk memastikan bahwa skor kepercayaan (confidence score) yang dihasilkan model, khususnya untuk kelas Kontrol yang under-represented, benar-benar mencerminkan kemungkinan sebenarnya, sehingga belum dievaluasi pada penelitian ini.

3.10 Rekomendasi untuk Penelitian Masa Depan

Berdasarkan temuan dan keterbatasan yang dijelaskan di atas, penelitian selanjutnya direkomendasikan untuk: (1) melakukan validasi eksternal menggunakan dataset ekspresi gen HD independen untuk menguji generalisasi model; (2) menerapkan skema validasi silang bertingkat sehingga proses pemilihan fitur dan penyetelan hiperparameter tidak menghasilkan hasil yang terlalu optimis; (3) mengeksplorasi metode penyeimbangan kelas alternatif selain SMOTE; (4) melakukan analisis biologis lebih lanjut terhadap fitur (gen) dengan kontribusi tertinggi, termasuk memetakan probe microarray yang teridentifikasi ke simbol gen resmi dan membandingkannya secara sistematis dengan literatur patofisiologi Huntington yang telah mapan, untuk mengonfirmasi (atau menyanggah) relevansi biologisnya sebelum dianggap sebagai kandidat biomarker; dan (5) menerapkan kalibrasi probabilitas (Platt scaling atau isotonic regression) yang dievaluasi menggunakan reliability diagram dan Brier score, dilengkapi pelaporan metrik yang tahan terhadap ketidakseimbangan kelas seperti Balanced Accuracy secara konsisten di seluruh perbandingan model, guna menjustifikasi klaim keberhasilan model secara lebih andal.

3.11 Pembahasan

Secara umum, temuan-temuan pada Bagian 3.1 hingga 3.10 menegaskan satu pola yang konsisten: pada dataset ekspresi gen Huntington GSE135589 yang berkarakter $p \gg n$, model linier teregularisasi (SVM kernel linier) memberikan generalisasi yang lebih baik dibandingkan model ensemble berbasis pohon (Random Forest dan XGBoost), baik dari segi akurasi data uji, F1-score, maupun ROC-AUC. Pola ini konsisten dengan kerangka *bias-variance tradeoff* yang telah diuraikan pada bagian Pendahuluan: ketika jumlah fitur jauh melebihi jumlah sampel, ruang fitur cenderung lebih mudah dipisahkan secara linier, sehingga model dengan asumsi lebih sederhana namun varians rendah, seperti SVM linier dengan parameter regularisasi C, lebih diuntungkan dibandingkan model berbasis pohon yang rentan menangkap korelasi kebetulan (*chance correlation*) pada data berdimensi tinggi [17].

Perbaikan metodologis berupa penerapan skema *nested cross-validation* (Bagian 3.4–3.6) juga memberikan kontribusi penting di luar sekadar peningkatan angka kinerja. Menyempitnya kesenjangan antara akurasi validasi silang dan akurasi data uji, bahkan berbalik arah pada SVM, menjadi bukti empiris bahwa skema seleksi fitur di luar loop cross-validation pada draf sebelumnya menghasilkan estimasi kinerja yang bias secara optimis akibat kebocoran informasi seleksi fitur, sebagaimana diperingatkan oleh Kapoor dan Narayanan mengenai krisis reproduksibilitas dalam penelitian berbasis pembelajaran mesin [15]. Dengan demikian, penelitian ini tidak hanya melaporkan model dengan kinerja terbaik, tetapi juga menunjukkan secara konkret dampak metodologis dari kebocoran data yang sering luput diperhatikan pada studi bioinformatika serupa.

Temuan mengenai bias ketiga model terhadap kelas mayoritas (Bagian 3.7) turut memperkuat kekhawatiran yang telah diangkat pada bagian Pendahuluan mengenai keterbatasan SMOTE pada data berdimensi sangat tinggi [8], [9]. Meskipun SMOTE berhasil menyeimbangkan distribusi kelas pada data pelatihan, sampel sintetis yang dihasilkan melalui interpolasi linier tampaknya belum cukup merepresentasikan variasi biologis kelas Kontrol yang sesungguhnya, sehingga recall kelas minoritas pada data uji tetap rendah, terutama pada Random Forest dan XGBoost. Secara klinis, pola ini penting untuk diperhatikan karena kegagalan mengenali sampel Kontrol berkonsekuensi berbeda dibandingkan kegagalan mengenali sampel HD, dan menunjukkan bahwa pendekatan penyeimbangan kelas alternatif seperti pembobotan kelas atau kalibrasi ambang keputusan [38] layak dipertimbangkan pada penelitian selanjutnya, sebagaimana telah diuraikan pada Bagian 3.9 dan 3.10. Secara kuantitatif, kekhawatiran ini juga tercermin pada nilai Balanced Accuracy (Bagian 3.5 dan 3.7), yang jauh lebih rendah daripada akurasi standar untuk Random Forest (53,70% berbanding 69,44%) dan XGBoost (61,11% berbanding 75,00%); dengan demikian, penggunaan Balanced Accuracy sebagai metrik ringkasan yang tahan terhadap ketidakseimbangan kelas, dilengkapi kalibrasi

probabilitas sebagai agenda penelitian lanjutan, memberikan dasar yang lebih kuat untuk menjustifikasi bahwa keunggulan SVM linier bukan sekadar artefak dari akurasi standar yang bias terhadap kelas mayoritas.

Secara keseluruhan, hasil penelitian ini memperkuat posisi bahwa perbandingan algoritma pembelajaran mesin pada data ekspresi gen berdimensi tinggi tidak dapat dilepaskan dari protokol validasi yang ketat; algoritma dengan kinerja yang tampak unggul pada skema validasi yang longgar dapat kehilangan keunggulannya, atau bahkan terungkap sebaliknya, ketika dievaluasi di bawah skema anti-kebocoran data yang benar. Daftar gen dengan skor kepentingan fitur tertinggi (Bagian 3.8) tetap relevan sebagai kandidat biomarker awal, namun interpretasinya perlu mempertimbangkan keterbatasan ukuran sampel dan potensi bias kelas yang telah diuraikan di atas.

4. KESIMPULAN

Studi ini berhasil membandingkan kinerja tiga algoritma pembelajaran mesin Support Vector Machine (SVM), Random Forest, dan XGBoost dalam memprediksi status penyakit Huntington berdasarkan data ekspresi gen microarray dari dataset GSE135589, dengan menerapkan protokol pra-pemrosesan yang ketat (pemilihan fitur *SelectKBest*, penyeimbangan kelas SMOTE, dan penyetelan hyperparameter GridSearchCV) untuk mencegah kebocoran data. Hasil evaluasi pada data uji independen menunjukkan bahwa model SVM *kernel linier* ($C = 0,1$) memberikan kinerja terbaik, dengan akurasi 88,89%, F1 Score 92,86%, dan ROC-AUC 90,95%, mengungguli XGBoost dan Random Forest, menunjukkan bahwa model linier yang diregulasi lebih kuat dalam kondisi di mana dimensi fitur jauh melebihi ukuran sampel ($p \gg n$) daripada model berbasis pohon keputusan. Penerapan skema nested cross-validation, dengan jumlah fitur (k) dituning sebagai hyperparameter bersama parameter model lainnya, terbukti menghilangkan kesenjangan besar antara akurasi validasi silang dan data uji yang teramati pada draft sebelumnya; untuk SVM, akurasi CV bahkan sedikit lebih rendah daripada akurasi uji, sementara untuk Random Forest dan XGBoost selisihnya menyusut hingga mendekati nol, menunjukkan bahwa estimasi kinerja kini bebas dari bias optimis akibat kebocoran informasi seleksi fitur. Meskipun demikian, ketiga model tetap menunjukkan kecenderungan bias terhadap kelas mayoritas (HD) dalam mengenali sampel kelas minoritas (Kontrol), yang menunjukkan keterbatasan generalisasi yang timbul dari ukuran sampel yang kecil. Daftar gen dengan skor kepentingan fitur tertinggi dari model Random Forest dan XGBoost saat ini masih bersifat eksploratif, karena diperoleh murni dari peringkat statistik (feature importance) tanpa dikonfirmasi terhadap literatur patofisiologi yang telah mapan mengenai penyakit *Huntington*, sehingga belum dapat diklaim sebagai kandidat biomarker. Interpretasi dan konfirmasi relevansi biologisnya, termasuk pemetaan probe ke simbol gen dan penelusuran jalur molekuler terkait, direkomendasikan sebagai agenda penelitian lanjutan; validasi eksternal menggunakan dataset independen dan skema validasi silang bertingkat juga tetap diperlukan untuk memastikan keandalan dan generalisasi model sebelum aplikasi yang lebih luas. Sebagai arah pengembangan penelitian serupa selanjutnya, kerangka kerja metodologis berbasis skema anti-kebocoran data dan nested cross-validation ini berpotensi diadaptasi untuk klasifikasi penyakit neurodegeneratif atau kondisi lain berbasis data ekspresi gen dengan karakteristik serupa ($p \gg n$), serta dapat diperkaya melalui integrasi data multi-omik (mis. genomik, proteomik) atau pendekatan deep learning untuk meningkatkan akurasi prediksi dan memperdalam pemahaman biologis yang dihasilkan. Berdasarkan temuan dan keterbatasan tersebut, beberapa saran diajukan untuk pengembangan penelitian serupa selanjutnya. Pertama, validasi eksternal menggunakan dataset independen dari kohort atau platform microarray/RNA-seq yang berbeda perlu dilakukan untuk menguji generalisasi model di luar dataset GSE135589. Kedua, penelitian lanjutan disarankan mengeksplorasi teknik penanganan ketidakseimbangan kelas alternatif, seperti pembobotan kelas (class weighting) atau kalibrasi ambang keputusan, sebagai pembanding terhadap SMOTE, mengingat keterbatasan interpolasi linier pada data berdimensi tinggi. Ketiga, penerapan metode explainable AI seperti SHAP atau LIME disarankan untuk memperkuat interpretasi kontribusi setiap fitur terhadap prediksi model, sekaligus membangun kepercayaan terhadap kandidat biomarker yang dihasilkan. Keempat, pemetaan probe microarray ke simbol gen serta penelusuran jalur molekuler (pathway enrichment analysis) perlu dilakukan secara sistematis untuk mengonfirmasi relevansi biologis gen-gen dengan skor kepentingan fitur tertinggi. Kelima, kolaborasi multi-institusi untuk memperbesar ukuran sampel penelitian juga disarankan guna memperkuat keandalan dan generalisasi model pada studi klasifikasi penyakit Huntington maupun penyakit neurodegeneratif lainnya berbasis data ekspresi gen.

5. PERNYATAAN PENULIS

5.1 Kontribusi Penulis

Konseptualisasi: R.Z.W., dan E.R.S.;

Metodologi: R.Z.W., dan E.R.S.;

Perangkat Lunak: R.Z.W.;

Validasi: R.Z.W., dan E.R.S.;

Analisis Formal: R.Z.W., dan E.R.S.;

Investigasi: R.Z.W., dan E.R.S.;

Sumber Daya: R.Z.W.;

Kurasi Data: R.Z.W.;

Penulisan Draf Awal: R.Z.W., dan E.R.S.;

Penulisan, Peninjauan, dan Penyuntingan: R.Z.W., dan E.R.S.;

Visualisasi: R.Z.W.;

Seluruh penulis telah membaca dan menyetujui versi manuskrip yang telah diterbitkan.

5.2 Ketersediaan Data

Data yang disajikan dalam penelitian ini tersedia berdasarkan permintaan kepada penulis korespondensi.

5.3 Pendanaan

Penelitian ini tidak menerima pendanaan eksternal

5.4 Pernyataan Dewan Peninjau Institusional

Tidak Berlaku (Not applicable).

5.5 Pernyataan Persetujuan Berdasarkan Informasi

Tidak Berlaku (Not applicable).

5.6 Pernyataan Konflik Kepentingan

Para penulis menyatakan bahwa mereka tidak memiliki kepentingan keuangan yang bersaing atau hubungan pribadi yang diketahui yang dapat dianggap memengaruhi pekerjaan yang dilaporkan dalam artikel ini.

5.7 Penggunaan AI Generatif dan Teknologi Berbantuan AI dalam Proses Penulisan

Para penulis menggunakan AI Generatif untuk meningkatkan kejelasan penulisan dalam artikel ini. Para penulis telah meninjau dan menyunting konten yang dihasilkan dengan bantuan AI serta bertanggung jawab penuh atas versi akhir publikasi.

REFERENCES

- [1] A. Jurcau, "Molecular Pathophysiological Mechanisms in Huntington's Disease," *Biomedicines*, vol. 10, no. 6, p. 1432, Jun. 2022, doi: 10.3390/biomedicines10061432.
- [2] S. J. Tabrizi, M. D. Flower, C. A. Ross, and E. J. Wild, "Huntington disease: new insights into molecular pathogenesis and therapeutic opportunities," *Nat. Rev. Neurol.*, vol. 16, no. 10, pp. 529–546, 2020, doi: 10.1038/s41582-020-0389-4.
- [3] J. Ko, H. Furby, X. Ma, J. D. Long, X.-Y. Lu, D. Slowiejko, and R. Gandhi, "Clustering and prediction of disease progression trajectories in Huntington's disease: An analysis of Enroll-HD data using a machine learning approach," *Front. Neurol.*, vol. 13, art. 1034269, 2023, doi: 10.3389/fneur.2022.1034269.
- [4] S. Ganesh, T. Chithambaram, N. R. Krishnan, D. R. Vincent, J. Kaliappan, and K. Srinivasan, "Exploring Huntington's Disease Diagnosis via Artificial Intelligence Models: A Comprehensive Review," *Diagnostics*, vol. 13, no. 23, art. 3592, 2023, doi: 10.3390/diagnostics13233592.
- [5] M. M. H. Aziz and S. A. Mahmood, "Utilizing Machine Learning Techniques for Cancer Prediction and Classification Based on Gene Expression Data," *UHD J. Sci. Technol.*, vol. 9, no. 1, pp. 135–148, Jun. 2025, doi: 10.21928/uhdjt.v9n1y2025.pp135-148.
- [6] F. Wang, A. M. Zain, Y. Ren, M. Bahari, A. A. Samah, Z. B. Ali Shah, N. B. Yusup, R. A. Jalil, A. Mohamad, and N. F. M. Azmi, "Navigating the microarray landscape: a comprehensive review of feature selection techniques and their applications," *Front. Big Data*, vol. 8, art. 1624507, 2025, doi: 10.3389/fdata.2025.1624507.
- [7] M. Salmi, D. Atif, D. Oliva, A. Abraham, and S. Ventura, "Handling imbalanced medical datasets: review of a decade of research," *Artif. Intell. Rev.*, vol. 57, no. 10, art. 273, 2024, doi: 10.1007/s10462-024-10884-2.
- [8] J. Lyu, J. Yang, Z. Su, and Z. Zhu, "LD-SMOTE: A Novel Local Density Estimation-Based Oversampling Method for Imbalanced Datasets," *Symmetry*, vol. 17, no. 2, art. 160, 2025, doi: 10.3390/sym17020160.
- [9] A. S. Tarawneh and A. B. Hassanat, "Stop Oversampling for Class Imbalance Learning: A Review," *IEEE Access*, vol. 10, pp. 47643–47660, 2022, doi: 10.1109/ACCESS.2022.3169512.
- [10] A. Apicella, F. Isgrò, and R. Prevete, "Don't push the button! Exploring data leakage risks in machine learning and transfer learning," *Artif. Intell. Rev.*, vol. 58, no. 11, art. 339, 2025, doi: 10.1007/s10462-025-11326-3.
- [11] J. Cheng, H.-P. Liu, W.-Y. Lin, and F.-J. Tsai, "Identification of contributing genes of Huntington's disease by machine learning," *BMC Med. Genomics*, vol. 13, no. 1, art. 176, 2020, doi: 10.1186/s12920-020-00822-w.
- [12] N. Dash, M. A. Bashar, J. Lee, and R. Dash, "Identification of Novel Biomarkers in Huntington's Disease Based on Differential Gene Expression Meta-Analysis and Machine Learning Approach," *Appl. Sci.*, vol. 15, no. 15, p. 8286, Jul. 2025, doi: 10.3390/app15158286.
- [13] T. Sugihartono, B. Wijaya, Marini, A. F. Alkayes, and H. A. Anugrah, "Optimizing Stunting Detection through SMOTE and Machine Learning: A Comparative Study of XGBoost, Random Forest, SVM, and k-NN," *J. Appl. Data Sci.*, vol. 6, no. 1, pp. 667–682, 2025, doi: 10.47738/jads.v6i1.494.
- [14] Q. Gao et al., "Identification of Orphan Genes in Unbalanced Datasets Based on Ensemble Learning," *Front. Genet.*, vol. 11, art. 820, Oct. 2020, doi: 10.3389/fgene.2020.00820.
- [15] S. Kapoor and A. Narayanan, "Leakage and the reproducibility crisis in machine-learning-based science," *Patterns*, vol. 4, no. 9, p. 100804, Sep. 2023, doi: 10.1016/j.patter.2023.100804.

- [16] V. Shanmugam, R. Razavi-Far, and E. Hallaji, “Addressing Class Imbalance in Intrusion Detection: A Comprehensive Evaluation of Machine Learning Approaches,” *Electronics*, vol. 14, no. 1, art. 69, 2024, doi: 10.3390/electronics14010069.
- [17] E. Cantor, S. Guauque-Olarte, R. León, S. Chabert, and R. Salas, “Knowledge-slanted random forest method for high-dimensional data and small sample size with a feature selection application for gene expression data,” *BioData Min.*, vol. 17, art. 34, 2024, doi: 10.1186/s13040-024-00388-8.
- [18] J. Bennett et al., “Guiding questions to avoid data leakage in biological machine learning applications,” *Nat. Methods*, vol. 21, no. 8, pp. 1444–1453, 2024, doi: 10.1038/s41592-024-02362-y.
- [19] R. K. Thelagathoti, D. S. Chandel, W. A. Tom, C. Jiang, G. Krzyzanowski, A. Olou, and M. R. Fernando, “Machine Learning-Based Ensemble Feature Selection and Nested Cross-Validation for miRNA Biomarker Discovery in Usher Syndrome,” *Bioengineering*, vol. 12, no. 5, art. 497, 2025, doi: 10.3390/bioengineering12050497.
- [20] R. Urbanowicz, R. Zhang, Y. Cui, and P. Suri, “STREAMLINE: A Simple, Transparent, End-To-End Automated Machine Learning Pipeline Facilitating Data Analysis and Algorithm Comparison,” in *Genetic Programming Theory and Practice XIX*, Singapore: Springer Nature Singapore, 2023, vol.19, pp. 201–231, doi: 10.1007/978-981-19-8460-0_9.
- [21] C. C. Olisah, L. Smith, and M. Smith, “Diabetes mellitus prediction and diagnosis from a data preprocessing and machine learning perspective,” *Comput. Methods Programs Biomed.*, vol. 220, p. 106773, Jun. 2022, doi: 10.1016/j.cmpb.2022.106773.
- [22] M. Sivakumar, S. Parthasarathy, and T. Padmapriya, “Trade-off between training and testing ratio in machine learning for medical image processing,” *PeerJ Comput. Sci.*, vol. 10, p. e2245, Sep. 2024, doi: 10.7717/peerj-cs.2245.
- [23] K. Dissanayake and M. G. Md Johar, “Comparative Study on Heart Disease Prediction Using Feature Selection Techniques on Classification Algorithms,” *Appl. Comput. Intell. Soft Comput.*, vol. 2021, pp. 1–17, Nov. 2021, doi: 10.1155/2021/5581806.
- [24] M. J. Lewis, A. Spiliopoulou, K. Goldmann, C. Pitzalis, P. McKeigue, and M. R. Barnes, “nestedcv: an R package for fast implementation of nested cross-validation with embedded feature selection designed for transcriptomics and high-dimensional data,” *Bioinform. Adv.*, vol. 3, no. 1, art. vbad048, 2023, doi: 10.1093/bioadv/vbad048.
- [25] S. A. Alex, J. Jesu Vedha Nayahi, and S. Kaddoura, “Deep convolutional neural networks with genetic algorithm-based synthetic minority over-sampling technique for improved imbalanced data classification,” *Appl. Soft Comput.*, vol. 156, p. 111491, May 2024, doi: 10.1016/j.asoc.2024.111491.
- [26] C. Paramita and W. Prasetyaningtyas, “Enhanced Chronic Kidney Disease Prediction Using Optimized Support Vector Machine With Hyperparameter Tuning And Smote,” *Rabit J. Teknol. dan Sist. Inf. Univrab*, vol. 11, no. 1, pp. 964–978, Jan. 2026, doi: 10.36341/rabit.v11i1.7179.
- [27] D.-P. Tran, G.-N. Nguyen, and V.-D. Hoang, “Hyperparameter Optimization for Improving Recognition Efficiency of an Adaptive Learning System,” *IEEE Access*, vol. 8, pp. 160569–160580, 2020, doi: 10.1109/ACCESS.2020.3020930.
- [28] T. Islam et al., “Predictive modeling for breast cancer classification in the context of Bangladeshi patients by use of machine learning approach with explainable AI,” *Sci. Rep.*, vol. 14, no. 1, pp. 1–17, 2024, doi: 10.1038/s41598-024-57740-5.
- [29] C. T. Mitchell et al., “Longitudinal expression changes are weak correlates of disease progression in Huntington’s disease,” *Brain Commun.*, vol. 2, no. 2, art. fcaa172, Jul. 2020, doi: 10.1093/braincomms/fcaa172.
- [30] K. Kotlarz, D. Słomian, W. Zawadzka, and J. Szyda, “Feature Selection Strategies for Deep Learning-Based Classification in Ultra-High-Dimensional Genomic Data,” *Int. J. Mol. Sci.*, vol. 26, no. 16, p. 7961, Aug. 2025, doi: 10.3390/ijms26167961.
- [31] F. E. Rebbah, H. Chamlal, and T. Ouaderhman, “Accurate analysis for univariate-based filter methods for microarray data classification,” *J. Algorithm. Comput. Technol.*, vol. 18, art. 17483026241232295, Jan. 2024, doi: 10.1177/17483026241232295.
- [32] T. M. Ali, A. Mir, A. U. Rehman, M. Humayun, M. Shaheen, and R. T. S. Alshammari, “Revolutionizing Lung Cancer Detection: A High-Accuracy Machine Learning Framework for Early Diagnosis,” *Biomed Res. Int.*, vol. 2025, art. 9961773, Jan. 2025, doi: 10.1155/bmri/9961773.
- [33] T. Ahmad, M. M. Kalan, F. Portier, and G. Stupfler, “Concentration and excess risk bounds for imbalanced classification with synthetic oversampling,” in *Advances in Neural Information Processing Systems 38 (NeurIPS 2025)*, vol.38, 2025, pp. 157456–157497, doi: 10.52202/085713-5262.
- [34] A. Firmansyah, M. F. Syahidin, and Y. S. Triana, “Prediksi Kebakaran Hutan Berdasarkan Titik Panas dan Iklim Menggunakan Algoritma Random Forest,” *Jurnal Nasional Teknologi dan Sistem Informasi (TEKNOSI)*, vol. 10, no. 2, pp. 145–155, Sep. 2024, doi: 10.25077/TEKNOSI.v10i2.2024.145-155.
- [35] R. N. Alifah, M. K. Najib, S. Nurdianti, A. P. Sari, K. Herlambang, Noval, D. T. P. B. Ginting, and S. N. Sya’adah, “Perbandingan Metode Tree Based Classification untuk Masalah Klasifikasi Data Body Mass Index,” *Indonesian Journal of Mathematics and Natural Sciences*, vol. 47, no. 1, pp. 49–65, 2024, doi: 10.15294/m2k97436.
- [36] N. M. Lefi and M. Rahardi, “Hyperparameter Optimization and Feature Selection Analysis on the XGBoost Model for Hepatitis C Infection Prediction,” *Journal of Applied Informatics and Computing (JAIC)*, vol. 9, no. 6, pp. 3338–3345, Dec. 2025, doi: 10.30871/jaic.v9i6.10876.
- [37] M. Kristanaya, M. P. Azzahra, Trimono, and M. Idhom, “Klasifikasi Status Rujukan Pasien Poliklinik Bandara Berbasis Random Forest dan Interpretabilitas Model Menggunakan SHAP,” *Data Sciences Indonesia (DSI)*, vol. 5, no. 1, pp. 60–74, Jul. 2025, doi: 10.47709/dsi.v5i1.6078.
- [38] H. Gittlin, “Beyond Synthetic Augmentation: Group-Aware Threshold Calibration for Robust Balanced Accuracy in Imbalanced Learning,” in *Machine Learning and Principles and Practice of Knowledge Discovery in Databases (AIDEM Workshop, ECML PKDD 2025)*, vol. 2843, Communications in Computer and Information Science, Cham: Springer, 2026, pp. 461–470, doi: 10.1007/978-3-032-19108-3_34.