Journal of Information System Research (JOSH)
Volume 6, No. 3, April 2025, pp 1564−1572
ISSN 2686-228X (media online)
https://ejurnal.seminar-id.com/index.php/josh/
DOI 10.47065/josh.v6i3.6898

Analisis Perbandingan Metode Random Forest dan Adaptive Boosting Untuk Prediksi Leukemia dengan Data Microarray

Juleha Irianti Heremba*, Christian Dwi Suhendra, Marlinda Sanglise

Teknik, Teknik Informatika, Universitas Papua, Manokwari
Jl. Gunung Salju, Amban, Kecamatan Manokwari Barat, Kabupaten Manokwari, Papua Barat, Indonesia
Email: [1,*]juleha.heremba@gmail.com, [2]c.suhendra@unipa.ac.id, [3]m.sanglise@unipa.ac.id
Email Penulis Korespondensi: juleha.heremba@gmail.com
Submitted: **02/02/2025;** Accepted: **02/04/2025;** Published: **06/04/2025**

Abstrak−Kanker merupakan pertumbuhan sel-sel yang tidak terkendali dan menyebar ke bagian tubuh lainnya. Ada berbagai jenis kanker yang diberi nama sesuai dengan organ asalnya. Salah satunya adalah kanker darah atau leukemia, merupakan kanker sumsum tulang yang disebabkan oleh mutasi genetik. Menurut data dari Global Cancer Statistics pada tahun 2020, diperkirakan terdapat 19,3 juta kasus kanker baru dan kasus kematian akibat kanker sebanyak 10 juta, dan diperkirakan pada tahun 2040 akan meningkat secara global sebanyak 47% dari 19,3 juta menjadi 28,4 juta kasus kanker baru. Leukemia merupakan salah satu jenis kanker dengan peringkat ke sembilan di Indonesia pada tahun 2020, terdapat 14.979 kasus baru dan 11.530 kasus kematian yang diakibatkan oleh leukemia. Salah satu upaya pencegahan leukemia dapat dilakukan dengan mendiagnosis kategori leukemia akut dengan menggunakan DNA dan informasi genetik. Tujuan penelitian ini adalah Menganalisis kinerja perbandingan antara metode Random Forest dan Adaptive Boosting dalam prediksi jenis leukemia dengan menggunakan dataset microarray untuk menentukan metode mana yang lebih efektif dalam melakukan klasifikasi. Dalam penelitian ini, dataset yang digunakan merupakan ekspresi gen pada sumsum tulang dan darah yang terdiri dari dua kategori leukemia akut yaitu Acute Myeloid Leukemia (AML) dan Acute Lymphoblastic Leukemia (ALL) yang diperoleh dengan teknologi microarray DNA. Gen ini akan diklasifikasi menggunakan metode Random Forest dan Adaboost untuk memprediksi kategori leukemia akut. Hasil dari proses analisis menunjukan bahwa metode random forest merupakan metode yang lebih baik untuk prediksi leukemia akut dengan nilai Area Under Curve 100%, Accuracy 92,9%, Precision 93,7%, Recall 92,9%, dan F1-Score 92,7% dibandingkan dengan metode AdaBoost dengan nilai Area Under Curve 83,3%, Accuracy 85,7%, Precision 88,6%, Recall 85,7%, dan F1-Score 85,1%.

Kata Kunci: Random Forest; AdaBoost; Leukemia; Microarray

Abstract−Cancer is the uncontrolled growth of cells that spread to other parts of the body. There are different types of cancer that are named after the organ they originate from. One of them is blood cancer or leukemia, which is bone marrow cancer caused by genetic mutations. According to data from Global Cancer Statistics in 2020, there were an estimated 19.3 million new cancer cases and 10 million cancer deaths, and it is estimated that by 2040 it will increase globally by 47% from 19.3 million to 28.4 million new cancer cases. Leukemia is one type of cancer with the ninth rank in Indonesia in 2020, there are 14,979 new cases and 11,530 cases of death caused by leukemia. One of the efforts to prevent leukemia can be done by diagnosing the acute leukemia category using DNA and genetic information. The purpose of this study is to analyze the comparative performance between Random Forest and Adaptive Boosting methods in predicting leukemia types using microarray datasets to determine which method is more effective in performing classification. In this study, the dataset used is gene expression in bone marrow and blood consisting of two categories of acute leukemia, namely Acute Myeloid Leukemia (AML) and Acute Lymphoblastic Leukemia (ALL) obtained with DNA microarray technology. These genes will be classified using Random Forest and Adaboost methods to predict acute leukemia categories. The results of the analysis process show that the random forest method is a better method for predicting acute leukemia with an Area Under Curve value of 100%, Accuracy 92.9%, Precision 93.7%, Recall 92.9%, and F1-Score 92.7% compared to the AdaBoost method with an Area Under Curve value of 83.3%, Accuracy 85.7%, Precision 88.6%, Recall 85.7%, and F1-Score 85.1%.

Keywords: Random Forest; AdaBoost; Leukemia; Microarray

1. PENDAHULUAN

Kanker merupakan pertumbuhan sel-sel tubuh yang tidak terkendali dan menyebar ke bagian tubuh lainnya. Kanker bisa muncul pada bagian tubuh mana pun dari tubuh manusia, dikarenakan tubuh manusia memiliki triliunan sel. Ada berbagai jenis kanker dan diberi nama sesuai sengan organ asalnya. Salah satunya adalah kanker darah yang disebut juga dengan istilah leukemia yaitu jenis kanker sumsum tulang yang disebabkan oleh mutasi genetik [1] [2].

Menurut data dari Global Cancer Statistics (GLOBOCAN) pada tahun 2020, diperkirakan terdapat 19,3 juta kasus kanker baru dan kasus kematian akibat kanker sebanyak 10 juta. Pada tahun 2040, diperkirakan akan meningkat secara global sebanyak 47% dari 19,3 juta menjadi 28,4 juta kasus kanker baru. Sementara itu, di Indonesia menurut data dari Global Cancer Statistics (GLOBOCAN) pada tahun 2020, leukemia merupakan jenis kanker terbanyak yang ada pada peringkat ke sembilan, diperkirakan terdapat 14.979 kasus baru dan 11.530 kasus kematian yang diakibatkan oleh leukemia [3], [4].

Orang yang menderita leukemia akan menghasilkan jumlah sel darah putih yang belum matang dengan jumlah yang tidak normal, dan merusak sumsum tulang serta menghambat pembentukan sel darah penting lainnya untuk sistem kekebalan tubuh yang seimbang dan menghambat pembentukan sel darah yang sehat. Leukemia

Journal of Information System Research (JOSH)
Volume 6, No. 3, April 2025, pp 1564−1572
ISSN 2686-228X (media online)
https://ejurnal.seminar-id.com/index.php/josh/
DOI 10.47065/josh.v6i3.6898

mempunyai dua jenis kategori leukemia akut yang dapat muncul secara tiba-tiba dan berkembang dengan cepat, yaitu Leukemia Mieloid Akut (AML) dan Leukemia limfoblastik akut (ALL) [5].

Salah satu upaya pencegahan leukemia akut dapat dilakukan dengan mendiagnosis Leukemia Mieloid Akut (AML) dan Leukemia limfoblastik akut (ALL) dengan menggunakan DNA seseorang dan informasi genetiknya. Diagnosis Leukemia merupakan hal penting yang berhubungan dengan prognosis dan pengobatan.

Dengan menggunakan data microarray, penulis dapat mendiagnosis Leukemia, serta mengukur tingkat ekspresi ribuan gen secara bersamaan. Data microarray telah banyak digunakan untuk penemuan biomarker kanker atau penanda gen dan diagnosis kanker, data microarray memiliki jumlah gen secara signifikan lebih besar daripada jumlah sampel. Sebagian besar gen dalam data microarray bersifat redundan dan beberapa gen yang relevan mungkin berguna untuk diagnosis kanker dan pemilihan terapi yang tepat dalam manajemen klinis [6].

Namun, data microarray memiliki dimensi tinggi karena mengandung ribuan fitur dengan sedikit sampel sehingga klasifikasi data microarray menjadi sulit dilakukan. Besarnya dimensi dapat berakibat pada tingkat performansi yang rendah. Untuk menangani klasifikasi data berdimensi tinggi, biasanya dilakukan reduksi dimensi data microarray.

Metode yang digunakan untuk reduksi yaitu t-Distributed Scochastic Neighbor Embedding (t-SNE). t-SNE merupakan metode analisis faktor non-linier, unsupervised, dan berbasis manifold dimana data berdemensi tinggi dipetakan ke data berdimensi rendah dengan tetap mempertahankan struktur signifikan dari data asli. Pada dasarnya, t-SNE digunakan untuk eksploitasi dan visualisasi data. Dengan kata lain, t-SNE menyediakan intuisi dengan bagaimana data diatur dalam ruang dimensi tinggi dan berguna untuk memvisualisasikan data dimensi tinggi dengan mempertahankan struktur data yang signifikan [7], [8], [9].

Beberapa penelitian terdahulu telah melakukan prediksi leukemia menggunakan berbagai algoritma Machine Learning dan menggunakan data microarray, penulis mengambil referensi dari penelitian sebelumnya untuk menjadi bahan pertimbangan sehingga dapat memberikan referensi terhadap penelitian yang akan dilakukan. Misalnya penelitian oleh Shidqi Aqil Naufal, Adiwijaya dan Widi Astuti pada tahun 2020, menggunakan teknik reduksi dimensi Partial Least Square (PLS), Support Vector Machine (SVM) dan K-Nearest Neighbors untuk Deteksi Kanker dengan Data Microarray dapat menghasilkan akurasi tertinggi sebesar 100% pada data lung dengan KNN[10].

Selanjutnya penelitian oleh Kecheng Zhu pada tahun 2021 menggunakan pengklasifikasi SVM linier dengan pengambilan sampel kueri secara acak (pembelajaran mesin pasif) dan kueri berbasis kumpulan (pembelajaran mesin aktif) untuk klasifikasi leukemia berbasis microarray, hasil penelitian ini menunjukkan tingkat akurasi rata-rata 83% dan recall 82% pada SVM linier dengan sampel kueri berbasis kumpulan (pembelajaran mesin aktif) [1].

Selanjutnya penelitian yang dilakukan oleh Razieh Sheikhpour, Roohallah Fazli, Saaz Mehrabani pada tahun 2021 dengan menggunakan k-nearest neighbor (KNN), support vector machine (SVM), pengklasifikasi berbasis estimasi densitas kernel Gaussian (GKDEC), dan pengklasifikasi diskriminan linier (LDC), untuk mengidentifikasi gen agar dapat mendiagnosis leukemia myeloid dan limfoblastik akut dan hasil menunjukkan bahwa pengklasifikasi GKDE dan LDC mendapatkan akurasi 100% [6].

Selanjutnya penelitian yang dilakukan oleh Noor Bahjat Tayfor dan Snoor Jamal Mohammed pada tahun 2021 menggunakan Naïve Bayes, Regresi Logistik, SVM, dan model Multi-Layer Perceptron (MPL) untuk mengklasifikasikan dataset kanker menjadi kanker darah dan kanker non-darah, hasil penelitian ini menunjukkan bahwa pengklasifikasi yang paling sesuai dengan kemampuan yang terbaik untuk memprediksi dataset kanker adalah Multilayer Perceptron dengan akurasi 99,3967% [11]. Terakhir penelitian yang dilakukan oleh Samira Ratnawati dan Siti Sunendiari pada tahun 2021 dengan judul Penggunaan Metode Logistic Regression Ensemble (LORENS) pada Klasifikasi Leukemia Akut bertujuan untuk menerapkan metode LORENS pada klasifikasi leukemia akut, hasil penelitian ini menunjukkan nilai akurasi sebesar 78,57% [12].

Dalam penelitian ini, data Microarray akan diolah dengan metode ensemble. Metode ensemble merupakan sebuah pendekatan untuk meningkatkan akurasi prediksi dengan menggabungkan hasil dari beberapa pengklasifikasi. Metode ensemble cenderung lebih akurat daripada pengklasifikasi individu dan mengurangi masalah overfitting [13]. penullis akan menggunakan beberapa metode ensemble yaitu Adaptive Boosting dan Random Forest.

Adaptive Boosting atau AdaBoost merupakan salah satu pendekatan klasifikasi boosting yang umum digunakan. AdaBoost menggunakan pendekatan boosting, prinsip dari metode boosting yaitu menghasilkan prediksi yang akurat dengan mengkombinasikan pengklasifikasi-pengklasifikasi lemah. Kelebihan dari AdaBoost yaitu mengurangi variance dan mengurangi bias [14]. Sedangkan Random Forest merupakan metode ensemble dari decision tree yang digunakan untuk klasifikasi dan regresi. Algoritma ini bekerja dengan menggabungkan beberapa pohon keputusan yang dibuat secara acak dari subset data. Setiap pohon memberikan prediksi dan hasil akhir diperoleh dengan menggabungkan prediksi dari semua pohon. Kelebihan dari Random Forest yaitu resistensi terhadap overfitting, akurasi yang tinggi, dan efisiensi pada dataset besar [15]. Random forests biasanya berkinerja lebih baik daripada bagging dan sebanding dengan boosting [16].

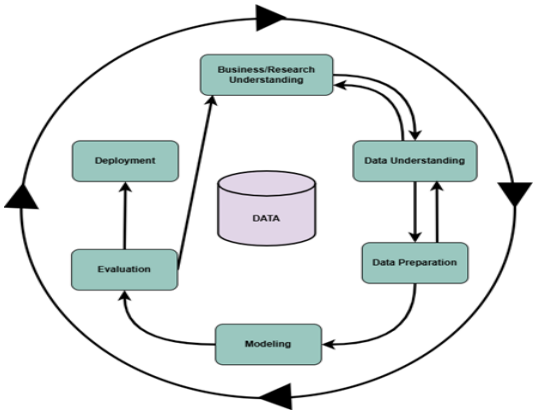
Oleh karena itu penelitian ini akan menggunakan metode Adaptive Boosting sebagai metode pembanding dari metode Random Forest [14]. Tujuan dari penelitian ini untuk membandingkan hasil akurasi mana yang paling baik dari dua metode klasifikasi ensemble dalam memprediksi leukemia akut dengan data microarray.

Journal of Information System Research (JOSH)
Volume 6, No. 3, April 2025, pp 1564−1572
ISSN 2686-228X (media online)
https://ejurnal.seminar-id.com/index.php/josh/
DOI 10.47065/josh.v6i3.6898

2. METODOLOGI PENELITIAN

2.1 Tahapan Penelitian

Metode yang digunakan pada penelitian ini menerapkan tahapan CRISP-DM (Cross Industry Standard Process for Data Mining) merupakan model proses yang sistematis untuk melakukan analisis data. Model ini terdiri dari enam tahapan berulang dari pemahaman masalah penelitian hingga penerapan. Menentukan tujuan penelitian yang jelas dan masalah yang ingin diselesaikan. Dalam hal ini, tujuan penelitian adalah untuk menganalisis dan membandingkan efektivitas metode Random Forest dan Adaptive Boosting dalam memprediksi jenis leukemia menggunakan data microarray.



Gambar 1. Tahapan CRISP-DM

Gambar 1 merupakan tahapan CRISP-DM yang memiliki enam tahapan yaitu business/research understanding merupakan tahap memahami masalah penelitian secara mendalam untuk menentukan tujuan dari penelitian, penentuan tujuan adalah salah satu aspek terpenting dalam tahap ini. Data understanding merupakan tahap mengumpulkan data dari sumber data, mengeksplorasi dan mendeskripsikannya, serta memeriksa kualitas data. Data preparation merupakan tahap persiapan data pada penelitian yang akan dilakukan untuk memastikan data sudah siap diolah dan memiliki kualitas yang baik. Modeling merupakan tahap memilih dan membuat model dilakukan untuk mengelompokkan data berdasarkan karakteristiknya agar sesuai dengan kebutuhan model yang akan digunakan. Evaluation merupakan tahap hasil yang diperoleh akan diperiksa dengan tujuan penelitian yang telah ditetapkan. Deployment merupakan hasil akhir mengenai informasi yang sudah didapatkan dari penerapan model sebelumnya [17].

2.2 Pemahaman Bisnis (Business atau Research Understanding)

Tahapan ini merupakan tahap memahami masalah penelitian secara mendalam untuk menentukan tujuan dari penelitian. Sehingga dapat menyiapkan informasi tertentu yang terkait dalam memprediksi kanker. Langkah-langkah yang akan dilakukan untuk mendapatkan informasi tertentu yaitu seperti tinjauan pustaka dan mencari referensi dari penelitian lain yang berkaitan dengan diagnosis leukemia. Penulis mengambil referensi dari penelitian sebelumnya untuk menjadi bahan pertimbangan agar dapat memberikan referensi atau tinjauan tertulis terhadap penelitian yang akan dilakukan. Sehingga, tujuan dari penelitian ini untuk menerapkan teknik data mining dalam memprediksi kanker dengan metode klasifikasi ensemble yaitu Adaptive Boosting dan Random Forest menggunakan data Microarray. Setelah itu membandingkan kedua metode algoritma gabungan yang memiliki hasil akurasi paling efektif dalam mengklasifikasi data berdasarkan kasus dalam penelitian ini.

2.3 Pemahaman Data (Data Understanding)

Pada penelitian ini, Data yang digunakan adalah data sekunder yang berasal dari platform Kaggle yang menyediakan serangkaian basis data yang relevan serta akses ke informasi biomedis dan genomik. Semua basis data tersebut tersedia secara online sehingga data bisa dikumpulkan dan diolah untuk mendapatkan kesimpulan dari hasil penelitian. Dataset yang diperoleh dari platform kaggel dengan judul "Gene expression dataset (Golub et al.)" berisi tentang ekspresi gen pada sumsum tulang dan darah tepi yang digunakan untuk klasifikasikan kanker melalui microarray DNA. Dataset yang digunakan dalam penelitian dapat diakses dan diunduh melalui link kaggel berikut https://www.kaggle.com/crawford/gene-expression [18].

Tabel 1. Dataset Awal

No	Gene Description	Gene Accession Number	1	call (1)	2	call (2)	3	call (3)	4	call (4)	…	62	call (34)
1	AFFX-BioB-5_at (endogenous control)	AFFX-BioB-5_at	-214	A	-139	A	-76	A	-135	A	…	-176	A

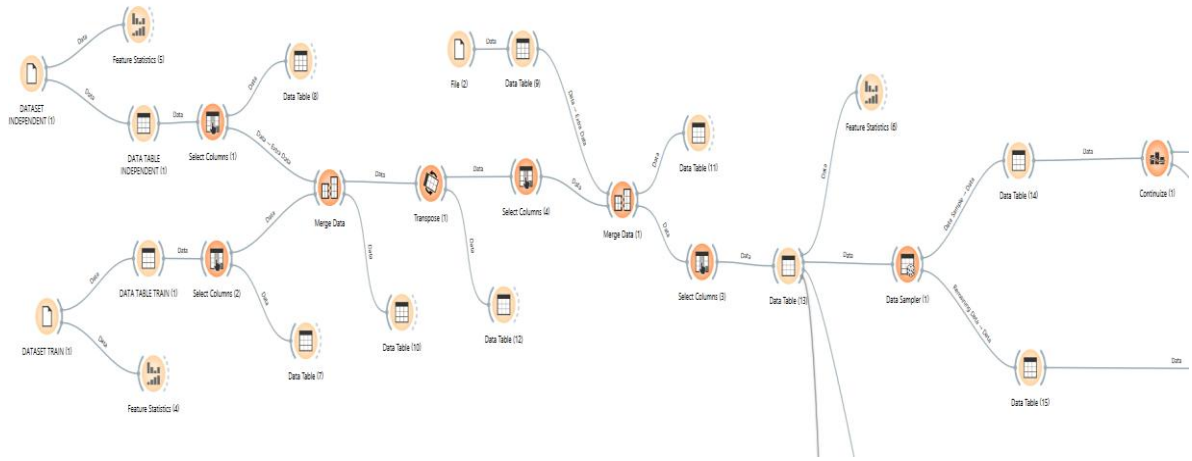
Journal of Information System Research (JOSH)
Volume 6, No. 3, April 2025, pp 1564−1572
ISSN 2686-228X (media online)
https://ejurnal.seminar-id.com/index.php/josh/
DOI 10.47065/josh.v6i3.6898

No	Gene Description	Gene Accession Number	1	call (1)	2	call (2)	3	call (3)	4	call (4)	…	62	call (34)
2	AFFX-BioB-M_at (endogenous control)	AFFX-BioB-M_at	-153	A	-73	A	-49	A	-114	A	…	-284	A
3	AFFX-BioB-3_at (endogenous control)	AFFX-BioB-3_at	-58	A	-1	A	-307	A	265	A	…	-81	A
…	…	…	…	…	…	…	…	…	…	…	…	…	…
7127	RB1 Retinoblastoma 1 (including osteosarcoma)	L49218_f_at	36	A	11	A	41	A	-50	A	…	20	A
7128	GB DEF = Glycophorin Sta (type A) exons 3 and 4; partial	M71243_f_at	191	A	76	A	228	A	126	A	…	379	A
7129	GB DEF = mRNA (clone 1A7)	Z78285_f_at	-37	A	-14	A	-41	A	-91	A	…	-60	A

Tabel 1 merupakan dataset penyakit leukemia dengan jumlah ekspresi gen 7129 dan 72 sample. Sample terdiri dari dua jenis leukemia yaitu 25 kasus acute myeloid leukemia (AML) dan 47 kasus acute lymphoblastic leukemia (ALL).

2.4 Persiapan Data (Data Preparation)

Setelah data diperoleh dari kaggel, tahapan persiapan data pada penelitian dilakukan untuk memastikan data yang akan siap diolah memiliki kualitas yang baik, implementasi proses ini akan menggunakan tools Orange.



Gambar 2. Preprocessing Data

Gambar 2 merupakan tahap preprocessing data, setelah data dimasukan ke dalam orange, proses awal dalam persiapan data yaitu data cleaning, proses ini dilakukan untuk menghapus beberapa kolom yang tidak diperlukan menggunakan widget select columns. Kedua, dilanjutkan dengan proses data integration, proses ini dilakukan untuk menggabungkan beberapa data variabel yang tabelnya terpisah dengan menggunakan widget merge data. Ketiga, table yang telah digabungkan akan ditransformation, proses ini dilakukan untuk mengubah struktur data dengan membalikan posisi pada baris dan kolom tabel dengan menggunakan widget transpose. Terakhir, dataset dibagi menjadi dua yaitu data training dan data testing dengan percobaan proporsi yaitu 80:20, pembagian data ini menggunakan widget data sampler, untuk data training ada 58 sample dan data testing ada 14 sample. Setelah itu, data akan dinormalisasi agar dapat meningkatkan kinerja model, tahap normalisasi data menggunakan widget continuize.

2.5 Pembentukan Model (Modeling)

Pada tahapan ini model klasifikasi dilakukan dengan menggunakan Adaptive Boosting dan Random Forest dengan menggunakan tools Orange. Dua model ini akan dilakukan dengan dua percobaan setelah itu dibandingkan untuk mendapatkan model yang memiliki akurasi yang terbaik dalam memprediksi kanker pada penelitian ini.

Random Forest, merupakan metode ensemble dari decision tree yang di dalam algoritma Random Forest membangun banyak pohon keputusan (T_1, T_2, ..., T_n) dari subset acak data dan subset acak fitur. Ensemble Voting digunakan untuk memprediksi kelas, algoritma Random Forest menggabungkan prediksi dari setiap pohon keputusan dan memilih kelas mayoritas.

$$y^{\wedge} = mode\ \{T_1(x), T_2(x), \dots, T_n(x)\} \quad (1)$$

Dimana Ti(x) adalah prediksi dari pohon keputusan ke-i. Pada pohon keputusan atau Decision Trees bekerja secara paralel. Setiap pohon memproses bagian berbeda dari data, dan hasil akhir didapatkan dari voting mayoritas dari semua pohon [19].

Journal of Information System Research (JOSH)
Volume 6, No. 3, April 2025, pp 1564−1572
ISSN 2686-228X (media online)
https://ejurnal.seminar-id.com/index.php/josh/
DOI 10.47065/josh.v6i3.6898

Adaptive Boosting merupakan metode yang menggunakan pendekatan boosting, prinsip dari metode boosting yaitu menghasilkan prediksi yang akurat dengan mengkombinasikan pengklasifikasi banyak model lemah menjadi satu model yang kuat, model lemah yang digunakan pada umumnya yaitu pohon keputusan dengan satu kedalaman atau decision stumps. Persamaan dari AdaBoost yang dirumuskan oleh Yoav Freund dan Robert Schapire yaitu sebagai berikut.

$$F(x) = \sum_{t=1}^{T} \propto_t h_t(x) \quad (2)$$

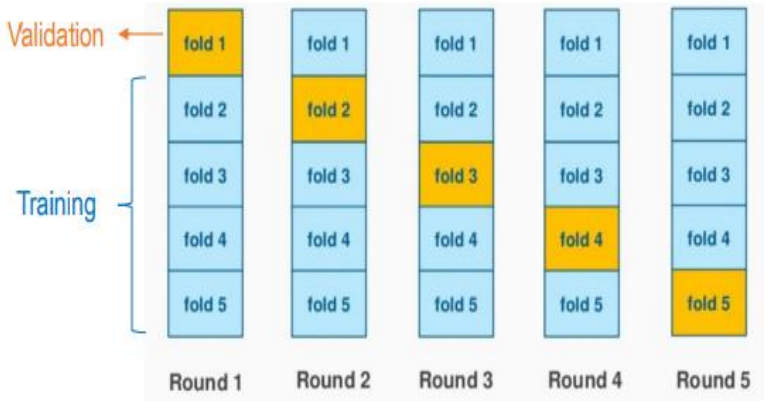
Untuk menghitung pengklasifikasian kuat maka F(x) sebagai prediksi, T sebagai jumlah model lemah, symbol sigma merepresentasikan penjumlahan dan $ht(x)$ sebagai pengklasifikasian lemah dikalikan dengan αt yang merupakan sebuah parameter penting atau learning rate [11], [20].

2.6 Evaluasi (Evaluation)

Evaluasi adalah proses untuk menilai kinerja dan keefektifan model klasifikasi yang telah dikembangkan. Untuk mengevaluasi setiap model, penulis akan menggunakan proses validasi untuk mengetahui seberapa baik model yang akan digunakan dalam memprediksi kanker proses validasi pada Orange menggunakan widget test and score. widget test and score dapat melakukan pengujian model dengan data. Widget dapat melakukan dua hal. Pertama, widget ini dapat menampilkan tabel dengan ukuran kinerja pengklasifikasi yang berbeda, seperti akurasi klasifikasi dan area di bawah kurva. Kedua, widget ini dapat mengeluarkan hasil evaluasi, yang dapat digunakan oleh widget lain untuk menganalisis kinerja pengklasifikasi, seperti Analisis ROC atau Confusion Matrix. Selain itu, Widget ini mendukung metode cross validation. Widget akan menghitung sejumlah statistik kinerja seperti berikut:

1) AUC atau Area under ROC adalah area di bawah kurva operasi penerima.
2) CA (Classification Accuracy) atau akurasi klasifikasi adalah proporsi contoh yang diklasifikasikan dengan benar.
3) F-1 score adalah rata-rata harmonik tertimbang dari presisi dan recall.
4) Precision atau Presisi adalah proporsi positif yang benar di antara contoh data yang diklasifikasikan sebagai positif, misalnya proporsi kanker yang diidentifikasi dengan benar sebagai kanker.
5) Recall adalah proporsi positif yang benar di antara semua contoh positif dalam data, misalnya jumlah orang sakit di antara semua yang didiagnosis sakit.
6) MCC atau Koefisien korelasi Matthews untuk memperhitungkan positif dan negatif yang benar dan salah dan secara umum dianggap sebagai ukuran yang seimbang yang dapat digunakan meskipun kelas-kelasnya memiliki ukuran yang sangat berbeda.

Proses validasi data akan diproses menggunakan metode K-fold Cross validation. K-Fold Cross Validation merupakan metode statistik yang digunakan untuk mengevaluasi kinerja model. K-Fold Cross Validation digunakan karena dapat mengurangi waktu kalkulasi dengan tetap menjaga keakuratan estimasi. Strategi paling umum dipakai adalah 10-fold, paling sedikit adalah 5-fold dan paling banyak 20-fold. Penulis akan menggunakan 10-fold-cross-validation yang akan membagi dataset secara acak menjadi sepuluh fold yaitu, Sembilan fold untuk training set dan satu fold untuk validation set [11], [20].



Gambar 3. Simulasi crossvalidation

Gambar 3 merupakan contoh alur kerja cross validation dengan penggunaan k-fold=5. Kemudian, untuk proses evaluasi dilakukan dengan menggunakan metode confusion matrix dan kurva ROC (Receiver Operating Characteristic). Confusion Matrix adalah sebuah instrumen yang digunakan untuk menilai efisiensi klasifikasi, memberikan ringkasan dari performa sebuah pengklasifikasi. Klasifikasi dengan Confusion Matrix ada empat konsep, yaitu True Positive (TP) menjelaskan jumlah data positif yang berhasil terdeteksi dengan benar, True Negative (TN) yaitu mencakup jumlah data negatif yang berhasil terdeteksi dengan benar, False Positive (FP) menjelaskan jumlah data positif yang tidak terdeteksi dengan benar, dan False Negative (FN) untuk mengindikasikan jumlah data negatif yang tidak terdeteksi dengan benar. Berikut merupakan tabel dari Confusion Matrix [11], [21], [22].

Journal of Information System Research (JOSH)
Volume 6, No. 3, April 2025, pp 1564−1572
ISSN 2686-228X (media online)
https://ejurnal.seminar-id.com/index.php/josh/
DOI 10.47065/josh.v6i3.6898

Tabel 2. Pengujian confusion matrix

		True Class	
		Positive	**Negative**
Predicted Class	Positive	TP	FP
	Negative	FN	TN

Dengan menggunakan Confusion Matrix, dapat memperoleh perhitungan untuk mengevaluasi kinerja model yang digunakan seperti berikut:

1) Accuracy, dihitung dengan rumus persamaan berikut:

$$Accuracy = \frac{(TP+TN)}{(TP+FP+FN+TN)} \quad (3)$$

2) Recall, dihitung dengan rumus persamaan berikut:

$$Recall = \frac{TP}{(TP+FN)} \quad (4)$$

3) Precision, dihitung dengan rumus persamaan berikut:

$$Precision = \frac{TP}{(TP+FP)} \quad (5)$$

4) F1-Score, dihitung dengan rumus persamaan berikut:

$$F1 - Score = \frac{2(Precision\ x\ Recall)}{(Precision+Recall)} \quad (6)$$

Sedangkan, kurva ROC akan menunjukkan perbandingan antara dua model klasifikasi [11].

2.7 Deployment

Tahapan deployment merupakan hasil akhir mengenai informasi yang sudah didapatkan dari penerapan model sebelumnya, data yang sudah diolah dan diuji coba hingga mendapatkan hasil akan divisualisasikan sehingga dapat memberikan informasi dan pengetahuan yang lebih mudah dipahami [23].

3. HASIL DAN PEMBAHASAN

3.1 Hasil Preprocessing Data

Pada Tabel 3 merupakan hasil dari Preprocessing data yang telah dibersihkan dengan menghapus beberapa kolom yang tidak diperlukan, setelah itu melakukan proses data integration yaitu menggabungkan beberapa data variabel yang tabelnya terpisah dengan table lainnya dan setelah itu table ditransformation yaitu mengubah struktur data dengan membalikan posisi pada baris dan kolom tabel.

Tabel 3. Hasil Preprocessing data

cancer	patient	1	2	3	4	5	…	7128	7129
ALL	1	-214	-153	-58	88	-295	…	191	-37
ALL	2	-139	-73	-1	283	-264	…	76	-14
ALL	3	-76	-49	-307	309	-376	…	228	-41
ALL	4	-135	-114	265	12	-419	…	126	-91
ALL	5	-106	-125	-76	168	-230	…	56	-25
…	…	…	…	…	…	…	…	…	…
AML	66	-58	-217	63	95	-191	…	1777	-49
ALL	67	-76	-98	-153	237	-215	…	80	-12
ALL	68	-154	-136	49	180	-257	…	-68	-1
ALL	69	-79	-118	-30	68	-110	…	109	-30
ALL	70	-55	-44	12	129	-108	…	176	40
ALL	71	-59	-114	23	146	-171	…	74	-12
ALL	72	-131	-126	-50	211	-206	…	237	-2

3.2 Hasil Cross Validation

Pada Gambar 4 merupakan hasil dari model Random Forest dan AdaBoost yang telah dilatih dengan reduksi t-SNE menggunakan widget test and score, mendapatkan hasil akurasi dan recall tertinggi pada model Random Forest yaitu dengan AUC 0.974 atau 97,4%, CA 0.879 atau 87,9%, F1-score 0.874 atau 87,4%, Precision 0.885 atau 88,5%, Recall 0.879 atau 87,9%, dan MCC 0.722 atau 72,2%, dan pada bagian bawah tabel dari Area under ROC curve menunjukan bahwa kinerja model Random Forest lebih baik daripada AdaBoost yaitu 0.966 atau 96.6%.

Journal of Information System Research (JOSH)
Volume 6, No. 3, April 2025, pp 1564−1572
ISSN 2686-228X (media online)
https://ejurnal.seminar-id.com/index.php/josh/
DOI 10.47065/josh.v6i3.6898

Evaluation results for target (None, show average over classes)

Model	AUC	CA	F1	Prec	Recall	MCC
Random Forest	0.974	0.879	0.874	0.885	0.879	0.722
AdaBoost	0.777	0.828	0.822	0.825	0.828	0.595

Compare models by: Area under ROC curve

	Random For...	AdaBoost
Random Forest		0.966
AdaBoost	0.034	

Gambar 4. cross validation menggunakan test and score

3.3 Hasil prediksi

Pada Gambar 5 menunjukan hasil prediksi dari model Random Forest mendapatkan nilai tertinggi dari model AdaBoost dengan AUC 1.000 atau 100%, CA 0.929 atau 92,9%, F1-score 0.927 atau 92,7%, Precision 0.937 atau 93,7%, Recall 0.929 atau 92,9%, dan MCC 0.861 atau 86,1%.

Show perfomance scores Target class: (Average over classes)

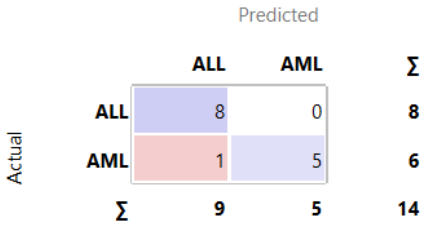
Model	AUC	CA	F1	Prec	Recall	MCC
Random Forest (1)	1.000	0.929	0.927	0.937	0.929	0.861
AdaBoost (1)	0.833	0.857	0.851	0.886	0.857	0.730

14 | 14 | 14 | 2×14

Gambar 5. hasil prediksi dengan reduksi t-SNE

3.4 Hasil Confusion Matrix

Selanjutnya, pada Gambar 6 menunjukkan jumlah data yang diprediksi dengan benar oleh model Random Forest.



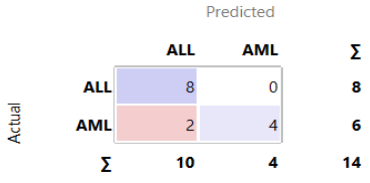
Gambar 6. confusion matrix random forest

Adapun keterangan dari nilai dalam table pada gambar diatas adalah sebagai berikut:
1) Nilai 8 adalah sampel ALL yang diprediksi benar sebagai ALL (True Positive atau TP)
2) Nilai 0 adalah sampel ALL yang diprediksi salah sebagai AML (False Positive atau FP)
3) Nilai 1 adalah sampel AML yang diprediksi salah sebagai ALL (False Negative atau FN)
4) Nilai 5 adalah sampel AML yang diprediksi benar sebagai ALL (True Negative atau TN)

Rumus yang digunakan untuk menghitung akurasi, precision, Recall, dan F1-Score pada Confusion Matrix adalah sebagai berikut:

1) $Accuracy = \frac{(TP+TN)}{(TP+FP+FN+TN)} = \frac{(8+5)}{(8+0+1+5)} = 0.929$
2) $Recall = \frac{TP}{(TP+FN)} = \frac{8}{(8+1)} = 0.889$
3) $Precision = \frac{TP}{(TP+FP)} = \frac{8}{(8+0)} = 1$
4) $F1 - Score = \frac{2(Precision\ x\ Recall)}{(Precision+Recall)} = \frac{2(1x0.889)}{(1+0.889)} = 0.941$

Pada Gambar 7 menunjukkan jumlah data yang diprediksi dengan benar oleh model AdaBoost.



Gambar 7. confusion matrix adaboost

Journal of Information System Research (JOSH)
Volume 6, No. 3, April 2025, pp 1564−1572
ISSN 2686-228X (media online)
https://ejurnal.seminar-id.com/index.php/josh/
DOI 10.47065/josh.v6i3.6898

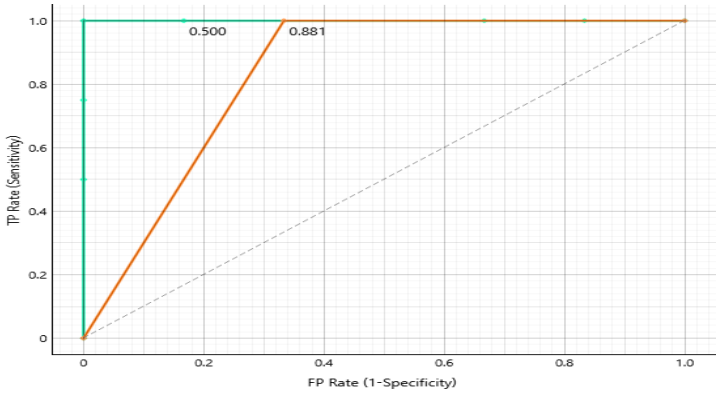
Adapun keterangan dari nilai dalam table pada gambar diatas adalah sebagai berikut:
1) Nilai 8 adalah sampel ALL yang diprediksi benar sebagai ALL (True Positive atau TP)
2) Nilai 0 adalah sampel ALL yang diprediksi salah sebagai AML (False Positive atau FP)
3) Nilai 2 adalah sampel AML yang diprediksi salah sebagai ALL (False Negative atau FN)
4) Nilai 4 adalah sampel AML yang diprediksi benar sebagai ALL (True Negative atau TN)

Rumus yang digunakan untuk menghitung akurasi, precision, Recall, dan F1-Score pada Confusion Matrix adalah sebagai berikut:

1) $Accuracy = \frac{(TP+TN)}{(TP+FP+FN+TN)} = \frac{(8+4)}{(8+0+2+4)} = 0.857$
2) $Recall = \frac{TP}{(TP+FN)} = \frac{8}{(8+2)} = 0.8$
3) $Precision = \frac{TP}{(TP+FP)} = \frac{8}{(8+0)} = 1$
4) $F1 - Score = \frac{2(Precision\ x\ Recall)}{(Precision+Recall)} = \frac{2(1x0.8)}{(1+0.8)} = 0.889$

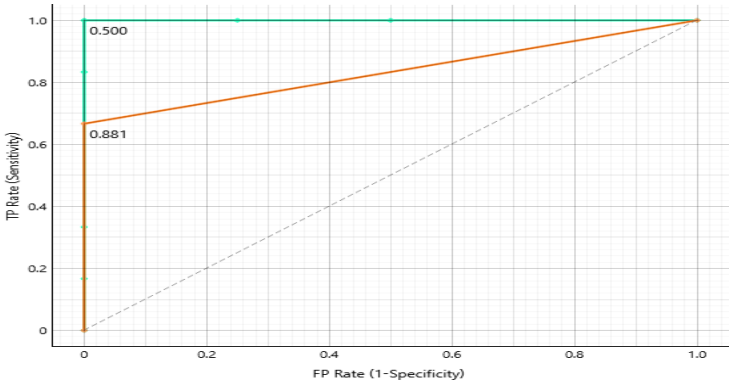
3.5 Kurva ROC

Pada Gambar 8 kurva ROC dalam prediksi ALL dan AML, untuk FP Rate (1-Specificity) merupakan sumbu X yang mewakili false positive dan FP Rate (Specificity) merupakan sumbu Y yang mewakili true positive. Sedangkan untuk kurva hijau mewakili model Random Forest dan kurva orange mewakili model AdaBoost.



Gambar 8. Hasil prediksi kanker ALL dengan Kurva ROC model random forest dan adaboost

Hasil perbandingan dalam memprediksi kanker ALL pada Gambar 8 menunjukan bahwa model Random Forest mendapatkan nilai AUC 0.881 yang menunjukkan bahwa model ini mendapatkan probabilitas 88.1% untuk mengklasifikasi sampel positif dengan benar dibandingkan dengan sampel negative.



Gambar 9. Hasil prediksi kanker AML dengan Kurva ROC model random forest dan adaboost

Hasil perbandingan dalam memprediksi kanker AML pada gambar 9 menunjukan bahwa model Random Forest mendapatkan nilai AUC 0.881 yang menunjukkan bahwa model ini mendapatkan probabilitas 88.1% untuk mengklasifikasi sampel positif dengan benar dibandingkan dengan sampel negative.

4. KESIMPULAN

Berdasarkan hasil dan pembahasan yang telah dijelaskan, dapat disimpulkan yaitu dengan proses analisis yang telah dilakukan untuk klasifikasi kategori leukemia akut menunjukkan bahwa metode klasifikasi ensemble dengan model Random Forest dan AdaBoost menggunakan reduksi t-SNE memiliki hasil akurasi yang tinggi pada model Random Forest dibandingkan dengan model AdaBoost, yaitu dengan hasil AUC 1.000 atau 100%, CA 0.929 atau

Journal of Information System Research (JOSH)
Volume 6, No. 3, April 2025, pp 1564−1572
ISSN 2686-228X (media online)
https://ejurnal.seminar-id.com/index.php/josh/
DOI 10.47065/josh.v6i3.6898

92,9%, F1-score 0.927 atau 92,7%, Precision 0.937 atau 93,7%, Recall 0.929 atau 92,9%, dan MCC 0.861 atau 86,1%, sehingga dapat simpulkan pada penelitian ini metode ensemble untuk model Random Forest dengan reduksi t-SNE lebih efektif dalam pengklasifikasi ekpresi gen menggunakan data microarray.

REFERENCES

[1] K. Zhu, "Active Learning for Microarray based Leukemia Classification," in 2021 8th International Conference on Biomedical and Bioinformatics Engineering, New York, NY, USA: ACM, Nov. 2021, pp. 77–81. doi: 10.1145/3502871.3502884.
[2] A. El-Baz and J. S. Suri, Artificial Intelligence in Cancer Diagnosis and Prognosis, Volume 1. IOP Publishing, 2022. doi: 10.1088/978-0-7503-3595-9.
[3] "Cancer statistics for the year 2020: An overview," Mar. 2021.
[4] E. Morgan et al., "Global burden of colorectal cancer in 2020 and 2040: incidence and mortality estimates from GLOBOCAN," Gut, vol. 72, no. 2, pp. 338–344, Feb. 2023, doi: 10.1136/gutjnl-2022-327736.
[5] D. Castillo et al., "Leukemia multiclass assessment and classification from Microarray and RNA-seq technologies integration at gene expression level," PLoS One, vol. 14, no. 2, p. e0212127, Feb. 2019, doi: 10.1371/journal.pone.0212127.
[6] R. Sheikhpour, R. Fazli, and S. Mehrabani, "Gene Identification from Microarray Data for Diagnosis of Acute Myeloid and Lymphoblastic Leukemia Using a Sparse Gene Selection Method," Iran J Ped Hematol Oncol, Mar. 2021, doi: 10.18502/ijpho.v11i2.5838.
[7] S. A. Naufal, A. Adiwijaya, and W. Astuti, "Analisis Perbandingan Klasifikasi Support Vector Machine (SVM) dan K-Nearest Neighbors (KNN) untuk Deteksi Kanker dengan Data Microarray," JURIKOM (Jurnal Riset Komputer), vol. 7, no. 1, p. 162, Feb. 2020, doi: 10.30865/jurikom.v7i1.2014.
[8] W. Astuti and A. Adiwijaya, "Principal Component Analysis Sebagai Ekstraksi Fitur Data Microarray Untuk Deteksi Kanker Berbasis Linear Discriminant Analysis," Jurnal Media Informatika Budidarma, vol. 3, no. 2, p. 72, Apr. 2019, doi: 10.30865/mib.v3i2.1161.
[9] F. Anowar, S. Sadaoui, and B. Selim, "Conceptual and empirical comparison of dimensionality reduction algorithms (PCA, KPCA, LDA, MDS, SVD, LLE, ISOMAP, LE, ICA, t-SNE)," Comput Sci Rev, vol. 40, p. 100378, May 2021, doi: 10.1016/j.cosrev.2021.100378.
[10] S. A. Naufal, A. Adiwijaya, and W. Astuti, "Analisis Perbandingan Klasifikasi Support Vector Machine (SVM) dan K-Nearest Neighbors (KNN) untuk Deteksi Kanker dengan Data Microarray," JURIKOM (Jurnal Riset Komputer), vol. 7, no. 1, p. 162, Feb. 2020, doi: 10.30865/jurikom.v7i1.2014.
[11] N. B. Tayfor and S. J. Mohammed, "A Comparison Study of Data Mining Algorithms for blood Cancer Prediction," Passer Journal of Basic and Applied Sciences, vol. 3, no. 2, pp. 174–179, Sep. 2021, doi: 10.24271/psr.29.
[12] S. Ratnawati, S. Sunendiari, P. Statistika, F. Matematika, D. Ilmu, and P. Alam, "Penggunaan Metode Logistic Regression Ensemble (LORENS) pada Klasifikasi Leukemia Akut", 2021, doi: 10.29313/.v7i1.25555.
[13] W. W. Piegorsch, "Statistical data analytics : foundations for data mining, informatics, and knowledge discovery," 2015.
[14] M. J. Paput, K. Suryowati, and M. T. Jatipaningrum, "Perbandingan Metode Random Forest Dan Adaptive Boosting Pada Klasifikasi Indeks Pembangunan Manusia Di Indonesia," Jurnal Statistika Industri dan Komputasi, vol. 8, no. 2, pp. 73–83, Jul. 2023, doi: 10.34151/statistika.v8i2.4458.
[15] A. C. Kurniawan and A. Salam, "Seleksi Fitur Information Gain untuk Optimasi Klasifikasi Penyakit Tuberkulosis," Jurnal Media Informatika Budidarma, vol. 8, no. 1, p. 70, Jan. 2024, doi: 10.30865/mib.v8i1.7122.
[16] C. C. Aggarwal, Data Mining. Cham: Springer International Publishing, 2015. doi: 10.1007/978-3-319-14142-8.
[17] C. Schröer, F. Kruse, and J. M. Gómez, "A Systematic Literature Review on Applying CRISP-DM Process Model," Procedia Comput Sci, vol. 181, pp. 526–534, 2021, doi: 10.1016/j.procs.2021.01.199.
[18] C. Crawford, "Gene expression dataset (Golub et al.)," Access Date Oct 2024, https://www.kaggle.com/datasets/crawford/gene-expression.
[19] Z. I. Bimawan, T. Astuti, and P. Arsi, "Comparison Of Random Forest, K-Nearest Neighbor, Decision Tree, And Xgboost Algorithms For Detecting Stunting In Toddlers Komparasi Algoritma Random Forest, K-Nearest Neighbor, Decision Tree, Xgboost Untuk Mendeteksi Penyakit Stunting Balita," Jurnal Teknik Informatika (JUTIF), vol. 5, no. 6, pp. 1599–1607, 2024, doi: 10.52436/1.jutif.2024.5.6.2629.
[20] S. Widaningsih, "Perbandingan Metode Data Mining Untuk Prediksi Nilai Dan Waktu Kelulusan Mahasiswa Prodi Teknik Informatika Dengan Algoritma C4,5, Naïve Bayes, Knn Dan Svm," Jurnal Tekno Insentif, vol. 13, no. 1, pp. 16–25, Apr. 2019, doi: 10.36787/jti.v13i1.78.
[21] H. Azis, P. Purnawansyah, F. Fattah, and I. P. Putri, "Performa Klasifikasi K-NN dan Cross Validation Pada Data Pasien Pengidap Penyakit Jantung," ILKOM Jurnal Ilmiah, vol. 12, no. 2, pp. 81–86, Aug. 2020, doi: 10.33096/ilkom.v12i2.507.81-86.
[22] D. Desyanti, J. Suarlin, and R. Faisal, "Otoritas Guru Dalam Prestasi Belajar Siswa Menggunakan Fuzzy Mamdani," Jurnal Media Informatika Budidarma, vol. 7, no. 3, p. 1323, Jul. 2023, doi: 10.30865/mib.v7i3.6368.
[23] F. Salsabila, I. Fitrianti, Y. Umaidah, and N. Heryana, "Penerapan Metode Crisp-Dm Untuk Analisa Pendapatan Bersih Bulanan Pekerja Informal Di Provinsi Jawa Barat Dengan Algoritma K-Means," Dinamik, vol. 28, no. 2, pp. 97–104, Jul. 2023, doi: 10.35315/dinamik.v28i2.9454.