p-ISSN: 2986-3929
e-ISSN: 3032-1948

Evaluasi KNN, SVM dan Random Forest untuk Klasifikasi Leukemia Berdasarkan Citra Sel Darah

Angelica Audeska Sali[a1], I Ketut Gede Suhartana[a2], I Komang Arya Ganda Wiguna[a3]

[a]Program Studi Informatika, Fakultas Matematika dan Ilmu Pengetahuan Alam, Universitas Udayana
Jalan Raya Kampus Udayana, Bukit Jimbaran, Kuta Selatan, Badung, Bali, Indonesia
[1]sali.2308561047@student.unud.ac.id
[2]ikg.suhartana@unud.ac.id
[3]arya.ganda@unud.ac.id

Abstract

Leukemia is a type of cancer that affects the blood-forming system and requires early detection to improve patient outcomes. One of the primary indicators of leukemia is the presence of blast cells in blood smears. Manual detection by hematologists is time-consuming and requires specialized expertise, prompting the need for automated classification methods. This study evaluates and compares the performance of three machine learning algorithms like K-Nearest Neighbor (KNN), Support Vector Machine (SVM), and Random Forest for detecting leukemia blast cells from microscopic blood images. The dataset used consists of 15,000 labeled images classified as either normal or blast cells. Feature extraction involved RGB and HSV color histograms, along with texture features derived from the Gray-Level Co-occurrence Matrix (GLCM). Model performance was assessed using confusion matrices and evaluated through accuracy, precision, recall, and F1-score. Among the models tested, Random Forest achieved the highest accuracy at 86.31%, followed by SVM at 83.61% and KNN at 81.40%. These results indicate that Random Forest is the most effective model for automated detection of leukemia blast cells in this context.

Keywords: *Leukemia, Blast Cells, Machine Learning, Random Forest, Support Vector Machine, K-Nearest Neighbor, Microscopic Image Classification, Image Processing.*

1. Pendahuluan

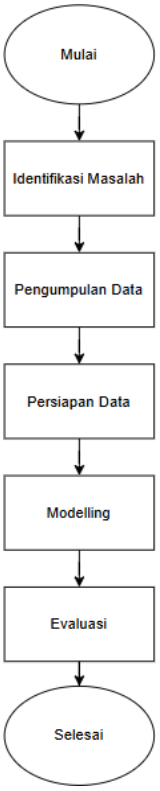
Leukemia adalah jenis kanker yang menyerang sistem pembentukan darah dan dapat mengancam jiwa jika tidak dideteksi secara dini. Salah satu indikator penting dalam diagnosis leukemia adalah kehadiran sel blast dalam darah. Deteksi manual oleh ahli hematologi membutuhkan keahlian tinggi dan waktu yang lama. Oleh karena itu, dibutuhkan metode otomatis yang dapat membantu dalam proses deteksi menggunakan citra mikroskopis. Seiring perkembangan teknologi, banyak penelitian telah mengaplikasikan metode machine learning dalam klasifikasi sel darah [1], termasuk untuk mengenali jenis-jenis leukosit [1] dan mendeteksi sel blast pada kasus leukemia limfoblastik akut [2]. Khandekar et al. [2] menunjukkan bahwa pendekatan otomatis mampu mendeteksi sel blast secara akurat dari citra mikroskopis. Pendekatan ini memanfaatkan fitur warna, bentuk, dan tekstur yang diekstrak dari sel darah putih. Bodzas et al. [3] juga menekankan pentingnya akurasi visual berdasarkan persepsi manusia dalam proses klasifikasi citra darah. Dalam penelitian ini, kami membandingkan tiga algoritma populer yaitu K-Nearest Neighbor (KNN), Support Vector Machine (SVM), dan Random Forest. Keempat algoritma ini sebelumnya telah digunakan dalam klasifikasi berbasis citra [4][5]. Dengan membandingkan performa model terhadap dataset leukemia berbasis citra mikroskopis, penelitian ini bertujuan untuk mengidentifikasi metode terbaik dalam mendeteksi sel blast secara otomatis dan akurat.

p-ISSN: 2986-3929
e-ISSN: 3032-1948

2. Metode Penelitian

2.1. Tahapan Penelitian

Penelitian ini melibatkan lima tahap utama: (1) Identifikasi Masalah, (2) Pengumpulan Data, (3) Persiapan Data, (4) Pembentukan Model, dan (5) Evaluasi Model. Diagram alur penelitian ditunjukkan pada Gambar 1.



Gambar 1. Tahap Penelitian

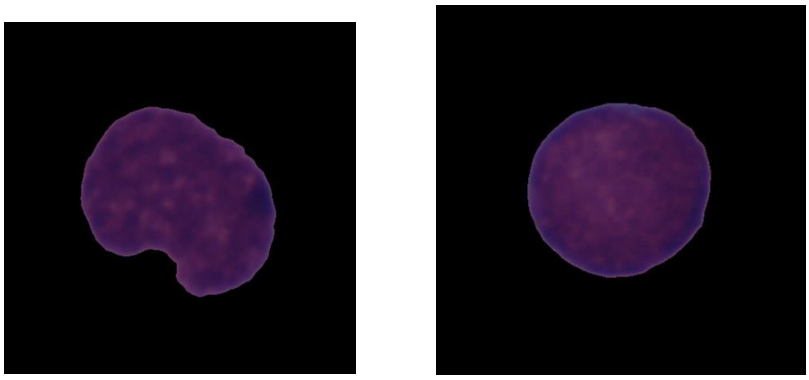
a. Identifkasi Masalah

Berdasarkan konteks yang telah disajikan, masalah yang diidentifikasi dalam penelitian ini menggunakan beberapa algoritma untuk membandingkan kinerja mereka. Penelitian ini memanfaatkan tiga algoritma, yakni K-Nearest Neighbor, *Support Vector Machine*, dan *Random Forrest* dengan tujuan memahami model klasifikasi mana yang memiliki akurasi tertinggi dalam menentukan cell blast.

b. Pengumpulan Data

Dataset C-NMC_Leukemia diunduh dari situs web kaggle.com, yang terdiri dari 15.000 gambar mikroskopis sel darah putih yang diklasifikasikan sebagai "normal" dan "blast"

p-ISSN: 2986-3929
e-ISSN: 3032-1948



Gambar 2. Dataset Penelitan, Sel Normal (Kiri) dan Sel Blast (Kanan)

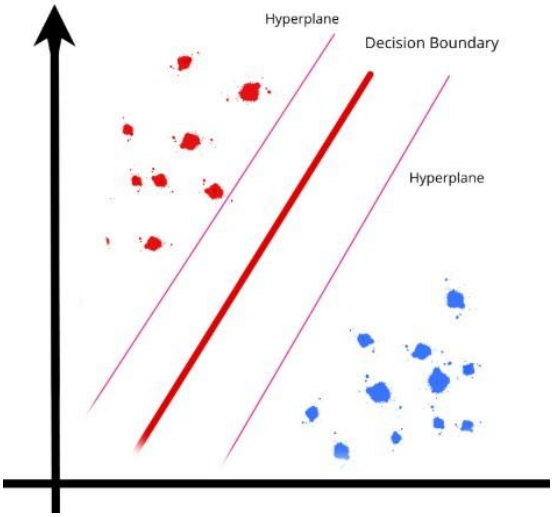
c. Persiapan Data

Setelah segmentasi berhasil, tahap selanjutnya adalah ekstraksi fitur yang mencerminkan karakteristik visual dan struktural dari setiap sel. Fitur warna diperoleh melalui perhitungan rata-rata kanal RGB dan histogram dari ruang warna HSV, yang masing-masing merepresentasikan aspek kecerahan, kejenuhan, dan rona warna. Pendekatan ini umum digunakan karena fitur warna terbukti membedakan antara sel normal dan blast yang memiliki perbedaan morfologi visual signifikan [3]. Selain warna, fitur tekstur juga diekstraksi menggunakan metode *Gray-Level Co-occurrence Matrix* (GLCM), yang mengukur hubungan spasial antar piksel dalam skala keabuan. Dari GLCM diperoleh metrik seperti kontras, homogenitas, dan energi, yang menggambarkan keragaman pola internal dari nukleus sel blast. Studi oleh Bodzas *et al.* [3] menekankan pentingnya fitur tekstur ini dalam mendukung klasifikasi otomatis leukemia karena struktur internal sel blast cenderung lebih kompleks dan padat dibandingkan sel normal. Dengan kombinasi fitur warna dan tekstur ini, sistem dapat mengidentifikasi perbedaan halus yang tidak mudah dikenali secara visual oleh manusia.

2.2. Support Vector Machine (SVM)

Support Vector Machine (SVM) adalah algoritma yang berupaya menemukan hyperplane optimal untuk memisahkan dua kelas data dengan margin maksimum. SVM sangat efektif dalam mengklasifikasikan data berdimensi tinggi dan telah banyak digunakan dalam domain citra medis karena kemampuannya dalam menangani kasus non-linear melalui fungsi kernel. Dalam pengaplikasiannya, SVM mencari parameter yang memaksimalkan jarak antara titik data kelas berbeda ke hyperplane. Sebagaimana dijelaskan oleh Badillo et al. [4], pendekatan ini memungkinkan klasifikasi yang akurat bahkan dalam data yang kompleks dan tinggi dimensinya, seperti citra darah. Proses pembentukan decision boundary dan margin maksimum ini divisualisasikan pada Gambar 3.

p-ISSN: 2986-3929
e-ISSN: 3032-1948



Gambar 3. Batas Keputusan *SVM* dengan *Margin*

2.3. Random Forrest

Random Forest adalah metode ensemble learning yang menggabungkan banyak pohon keputusan (decision tree) yang dilatih pada subset acak dari data dan fitur untuk menghasilkan keputusan klasifikasi yang lebih stabil dan akurat. Dengan menggabungkan hasil dari banyak pohon, Random Forest mengurangi overfitting dan menghasilkan model yang lebih general. Untuk setiap pohon, pemisahan dilakukan berdasarkan kriteria seperti Gini Impurity.

2.4. K-Nearest Neighbour

K-Nearest Neighbor (KNN) merupakan algoritma klasifikasi non-parametrik yang menentukan kelas suatu data berdasarkan kedekatan jarak ke k data latih terdekat. KNN tidak memerlukan pelatihan eksplisit, dan klasifikasi dilakukan saat data uji diberikan. Proses pengambilan keputusan bergantung pada mayoritas kelas dari tetangga terdekat, di mana jarak biasanya dihitung menggunakan Euclidean distance.

2.5. Evaluasi Model

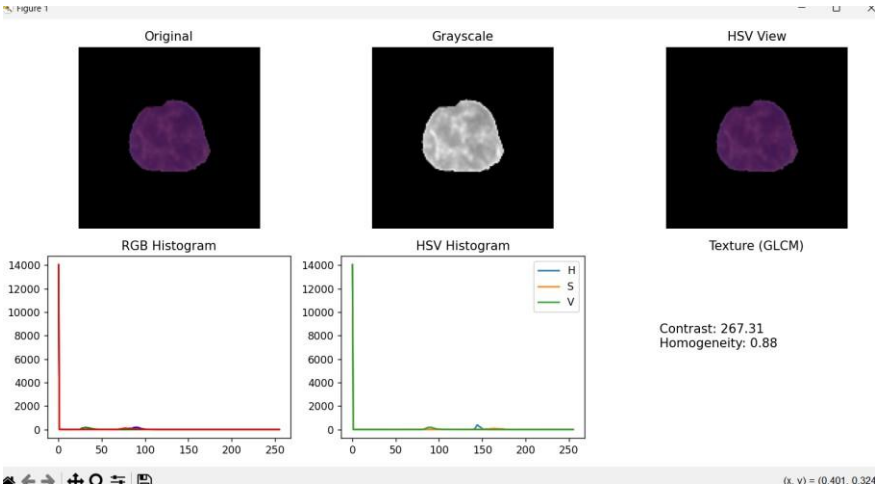
Evaluasi dilakukan dengan menghitung confusion matrix untuk masing-masing model. Confusion matrix menggambarkan performa model dalam membedakan antara sel blast dan sel normal. Metrik evaluasi yang digunakan adalah akurasi, precision, recall, dan f1-score.

3. Hasil dan Diskusi

3.1. Preprocessing Data

Setiap citra pada dataset diubah ke dalam ukuran 128×128 piksel untuk standarisasi dimensi. Konversi ke format grayscale dilakukan untuk memudahkan perhitungan tekstur menggunakan GLCM, sementara transformasi ke ruang warna HSV dilakukan untuk mengekstrak karakteristik warna yang lebih representatif terhadap variasi rona dan intensitas dalam struktur sel darah. Selanjutnya, dilakukan normalisasi dan augmentasi sederhana guna meningkatkan generalisasi model saat pelatihan.

p-ISSN: 2986-3929
e-ISSN: 3032-1948



Gambar 4. Visualisasi Preprocessing

3.2. Ekstraksi Fitur

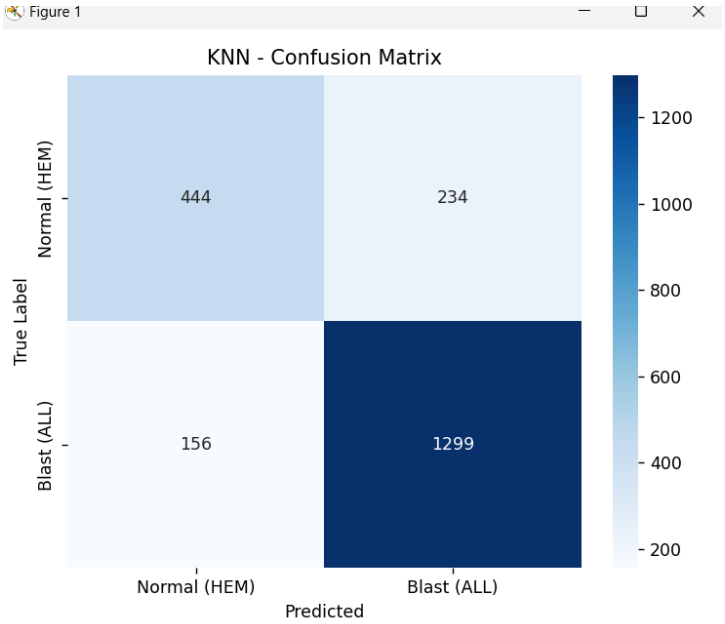
Setiap citra diproses untuk menghasilkan total 15 fitur utama. Tiga fitur pertama adalah rata-rata warna dari masing-masing kanal RGB: mean_R, mean_G, dan mean_B. Kemudian, sembilan fitur warna tambahan diperoleh dari histogram ruang warna HSV, masing-masing kanal (H, S, V) dibagi menjadi tiga bin: hist_H_0 hingga hist_V_2. Tiga fitur tekstur terakhir diekstraksi dari citra grayscale menggunakan metode Gray-Level Co-occurrence Matrix (GLCM), yaitu contrast, homogeneity, dan energy, yang memberikan informasi tentang distribusi intensitas dan struktur tekstur pada inti sel blast.

3.3. Pengujian Model

a. *K-Nearest Neighbour (KNN)*

Model KNN menghasilkan 444 prediksi yang benar untuk sel normal (True Negatives) dan 1299 prediksi yang benar untuk sel blast (True Positives). Namun, terdapat 234 kasus salah prediksi positif (False Positives) dan 156 kasus salah prediksi negatif (False Negatives). Berdasarkan nilai-nilai ini, model KNN memiliki precision sekitar 84.7%, recall sekitar 89.3%, dan f1-score sekitar 86.9%, yang menunjukkan performa yang cukup kuat dalam mendeteksi keberadaan sel blast meskipun tidak setinggi model lain seperti Random Forest. Gambar 5, menunjukkan confusion matrix dari model ini, yang menggambarkan keseimbangan relatif antara kedua kelas meskipun terdapat kecenderungan salah klasifikasi pada kelas normal (HEM).

p-ISSN: 2986-3929
e-ISSN: 3032-1948

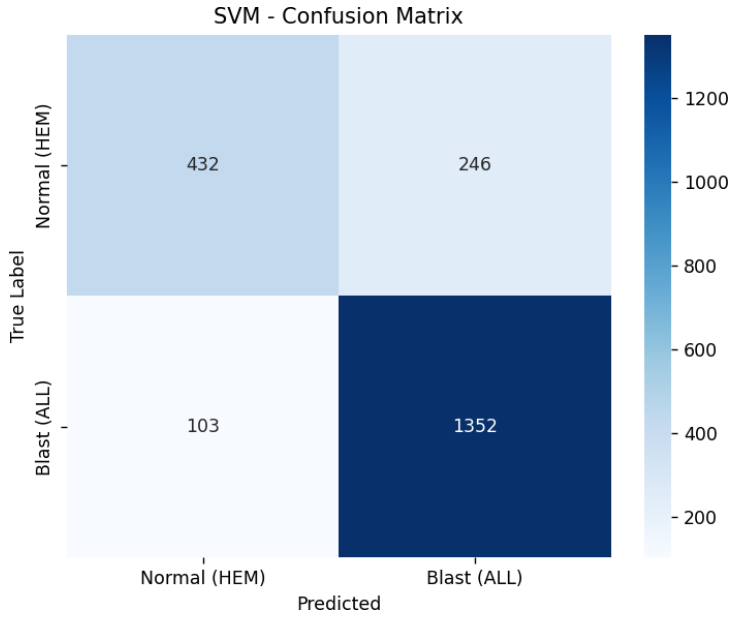


Gambar 5. Hasil Uji Coba Model *K-Nearest Neighbour (K-NN)*

b. *Support Vector Machine(SVM)*

Model SVM menunjukkan performa tinggi dengan 432 prediksi benar untuk sel normal (True Negatives) dan 1352 prediksi benar untuk sel blast (True Positives). Terdapat 246 kasus salah prediksi positif (False Positives) dan 103 kasus salah prediksi negatif (False Negatives). Berdasarkan data ini, akurasi keseluruhan model adalah 83.61%, dengan estimasi precision sekitar 84.6%, recall sekitar 92.9%, dan f1-score sekitar 88.6%, yang mencerminkan efektivitas tinggi model ini dalam mengklasifikasikan dua kelas, terutama dalam mengenali sel blast. Visualisasi evaluasi dari model SVM ditampilkan pada Gambar 6, yang menunjukkan keseimbangan performa antara kemampuan mendeteksi positif dan menghindari kesalahan pada kelas negatif.

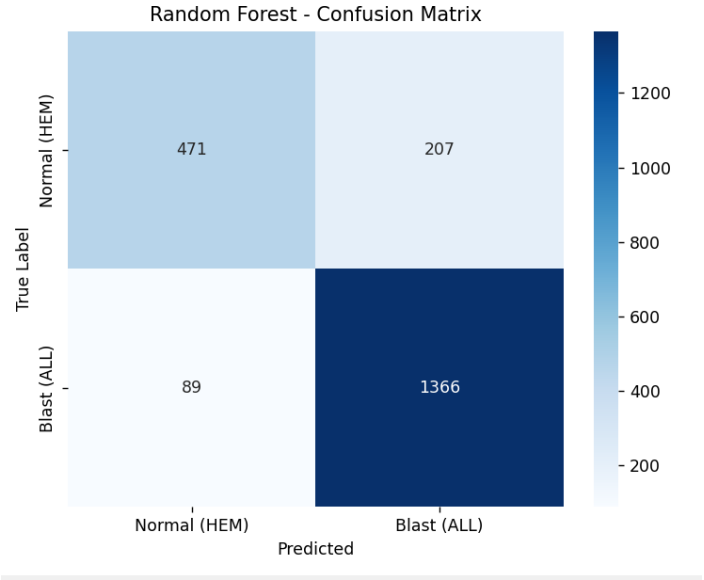
p-ISSN: 2986-3929
e-ISSN: 3032-1948



Gambar 6. Hasil Uji Coba Model *SVM*

c. Random Forrest

Model Random Forest mencatat performa terbaik di antara semua model, dengan 471 prediksi benar untuk sel normal (True Negatives) dan 1366 prediksi benar untuk sel blast (True Positives). Jumlah False Positives tercatat sebanyak 207 kasus, sedangkan False Negatives sebanyak 89 kasus. Dengan hasil ini, model mencapai akurasi sebesar 86.31%, dan estimasi metrik evaluasi menunjukkan nilai yang sangat kompetitif, yaitu precision sekitar 86.8%, recall sekitar 93.9%, dan f1-score sekitar 90.2%. Kinerja ini menegaskan efektivitas Random Forest dalam klasifikasi dua kelas, khususnya dalam mendeteksi sel blast dengan tingkat kesalahan yang relatif rendah. Confusion matrix dari model ini ditampilkan pada Gambar 7.



Gambar 7. Hasil Uji Coba Model *Random Forrest*

3.4. Akurasi Hasil Perbandingan

Tabel 1. Akurasi Hasil Perbandingan

Model	Akurasi
K-Nearest Neighbour	81.40%
Support Vector Machine	83.61%
Random Forrest	86.31%

Berdasarkan hasil perbandingan pada Tabel 1, model Random Forest menunjukkan performa terbaik dengan akurasi tertinggi sebesar 86.31%, mengindikasikan kemampuannya yang unggul dan konsisten dalam membedakan antara kondisi Normal (HEM) dan Blast (ALL), yang sangat krusial dalam konteks deteksi sel blast untuk diagnosis penyakit leukemia. Model Support Vector Machine (SVM) menempati posisi kedua dengan akurasi sebesar 83.61%, memperlihatkan kinerja yang solid dan keseimbangan yang baik antara prediksi positif dan negatif, menjadikannya alternatif yang handal. Sementara itu, model K-Nearest Neighbors (KNN) juga memberikan hasil yang cukup baik dengan akurasi 81.40%, menunjukkan bahwa pendekatan berbasis kedekatan data masih efektif untuk tugas klasifikasi ini, meskipun sedikit tertinggal dibandingkan dengan SVM dan Random Forest. Ketiga model ini memiliki potensi yang layak untuk digunakan dalam sistem pendukung keputusan medis, khususnya untuk mendeteksi keberadaan sel blast pada data sitologi atau hematologi, namun dengan mempertimbangkan kompleksitas dan kebutuhan akurasi tinggi dalam diagnosis klinis, Random Forest menjadi pilihan yang paling direkomendasikan.

4. Kesimpulan

Berdasarkan hasil evaluasi terhadap tiga algoritma machine learning, Random Forest menunjukkan performa terbaik dengan akurasi sebesar 86.31%, diikuti oleh Support Vector Machine (83.61%) dan K-Nearest Neighbour (81.40%). Hal ini menunjukkan bahwa pendekatan ensemble learning seperti Random Forest lebih andal dalam mengklasifikasikan sel darah antara normal dan blast pada citra mikroskopis leukemia. Oleh karena itu, Random Forest dapat dipertimbangkan sebagai metode yang paling efektif dalam deteksi otomatis leukemia limfoblastik akut pada dataset ini.

Daftar Pustaka

[1] M. Sharif et al., “Recognition of different types of leukocytes using YOLoV2 and optimized bag-of-features,” IEEE Access, vol. 8, pp. 167448–167459, 2020, doi: 10.1109/ACCESS.2020.3021660.

[2] R. Khandekar, P. Shastry, S. Jaishankar, O. Faust, and N. Sampathila, “Automated blast cell detection for Acute Lymphoblastic Leukemia diagnosis,” *Biomedical Signal Processing and Control*, vol. 68, Jul. 2021, doi: 10.1016/j.bspc.2021.102690.

[3] A. Bodzas, P. Kodytek, and J. Zidek, “Automated Detection of Acute Lymphoblastic Leukemia From Microscopic Images Based on Human Visual Perception,” *Frontiers in Bioengineering and Biotechnology*, vol. 8, 2020, doi: 10.3389/fbioe.2020.01005.

[4] 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.

[5] S. Badillo *et al.*, “An Introduction to Machine Learning,” *Clinical Pharmacology and Therapeutics*, vol. 107, no. 4, pp. 871–885, Apr. 2020, doi: 10.1002/cpt.1796.

[6] M. R. Larijani, E. A. Asli-Ardeh, E. Kozegar, and R. Loni, “Evaluation of image processing technique in identifying rice blast disease in field conditions based on KNN algorithm improvement by K-means,” *Food Science and Nutrition*, vol. 7, no. 12, pp. 3922–3930, Dec. 2019, doi: 10.1002/fsn3.1251.

[7] M. Ghaderzadeh, F. Asadi, A. Hosseini, D. Bashash, H. Abolghasemi, and A.

Roshanpour, “Machine Learning in Detection and Classification of Leukemia Using Smear Blood Images: A Systematic Review,” *Scientific Programming*, vol. 2021. Hindawi Limited, 2021. doi: 10.1155/2021/9933481.

[8] K. K. Anilkumar, V. J. Manoj, and T. M. Sagi, “A survey on image segmentation of blood and bone marrow smear images with emphasis to automated detection of Leukemia,” *Biocybernetics and Biomedical Engineering*, vol. 40, no. 4. 2020. doi: 10.1016/j.bbe.2020.08.010.

[9] N. Rismayanti, A. Naswin, U. Zaky, M. Zakariyah, and D. A. Purnamasari, “Evaluating Thresholding-Based Segmentation and Humoment Feature Extraction in Acute Lymphoblastic Leukemia Classification using Gaussian Naive Bayes,” *International Journal of Artificial Intelligence in Medical Issues*, vol. 1, no. 2, pp. 74–83, Nov. 2023, doi: 10.56705/ijaimi.v1i2.99.

[10] M. N. H. Gumay, Y. Widiastiwi, M. M. Santoni, and Y. Yulnelly, “Perbandingan Metode Naïve Bayes dan K-Nearest Neighbor Pada Klasifikasi Morfologi Gen Sel Darah Putih,” *Informatik : Jurnal Ilmu Komputer*, vol. 18, no. 1, 2022, doi: 10.52958/iftk.v17i4.4576.

p-ISSN: 2986-3929
e-ISSN: 3032-1948

Halaman ini sengaja dibiarkan kosong