Technical Report : Image-Based Anemia Classification using Random Forest

RAIHAN AKIRA RAHMAPUTRA
105222040@student.universitaspertamina.ac.id
GEMA FITRI RAMADANI
105222009@student.universitaspertamina.ac.id

1. Pendahuluan

Anemia merupakan salah satu gangguan darah yang paling umum di dunia dan dapat mempengaruhi produktivitas serta kualitas hidup seseorang. Proses diagnosis secara tradisional umumnya memerlukan pemeriksaan laboratorium lengkap, yang tidak selalu tersedia di daerah terpencil atau wilayah dengan keterbatasan infrastruktur medis. Oleh karena itu, pengembangan sistem berbasis citra sebagai alat skrining awal menjadi sangat penting.

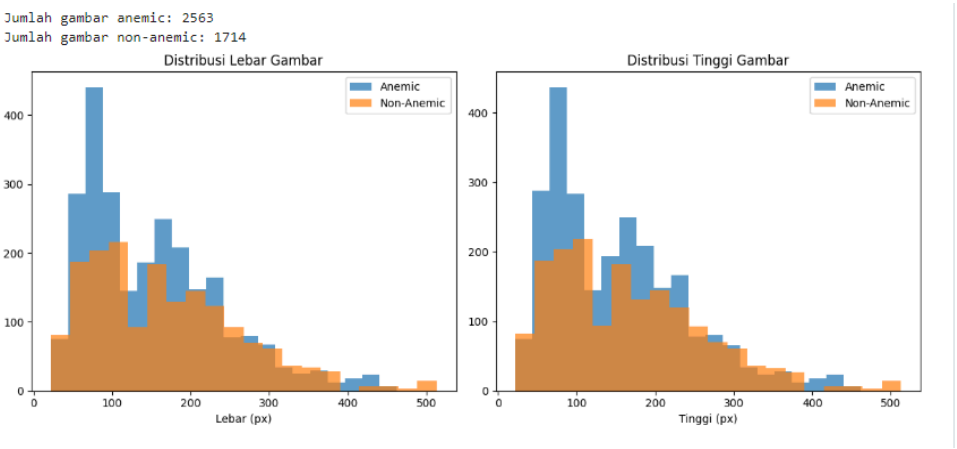
Model ini mengusulkan pengembangan sistem klasifikasi anemia berdasarkan gambar mikroskopis sel darah yang dianalisis menggunakan algoritma *Random Forest*. Algoritma ini dipilih karena kemampuannya dalam mengolah data tabular hasil ekstraksi fitur citra, ketahanannya terhadap *noise* dan *outlier*, serta kemudahan dalam interpretasi—khususnya dalam konteks klasifikasi biner. Melalui pendekatan ini, diharapkan dapat tercipta sistem deteksi anemia yang efisien, terjangkau, dan mudah diakses oleh masyarakat luas, terutama di wilayah dengan keterbatasan sumber daya medis.

2. Dataset dan Pra-pemrosesan

Dataset yang digunakan dalam model ini terdiri dari sekitar 4.277 gambar, yang mencakup 2.563 gambar berlabel *anemia* dan 1.714 gambar berlabel *non-anemia*. Gambar-gambar tersebut memiliki format .jpg, .jpeg, atau .png, dengan resolusi asli yang beragam. Untuk menyamakan dimensi input, seluruh gambar dikonversi ke format *grayscale* dan diubah ukurannya menjadi 128×128 piksel.

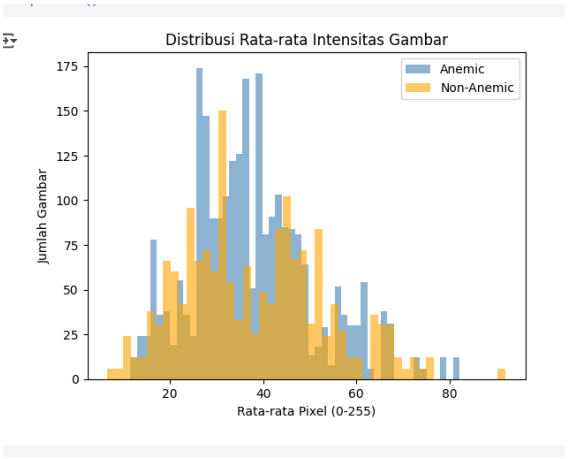
Exploratory Data Analysis (EDA)

Histogram Distribusi Ukuran Gambar (Tinggi dan Lebar Sebelum Resize)



Histogram ini menunjukkan sebaran ukuran gambar dari segi tinggi dan lebar sebelum dilakukan proses *resize*. Terlihat bahwa ukuran gambar sangat bervariasi; beberapa gambar memiliki resolusi yang kecil, sementara yang lain berukuran jauh lebih besar. Variasi ini perlu dinormalisasi, karena perbedaan resolusi dapat mempengaruhi hasil ekstraksi fitur dan performa model. Oleh karena itu, proses *resize* ke ukuran yang seragam menjadi langkah penting dalam tahap pra proses

Histogram Distribusi Rata-Rata Intensitas Piksel



Histogram ini menggambarkan distribusi rata-rata intensitas piksel dari gambar pada kedua kelas. Meskipun terdapat tumpang tindih antara kelas *Anemia* dan *Non-Anemia*, ada pola umum yang dapat diamati: gambar dari kelas *Anemia* cenderung memiliki intensitas rata-rata yang lebih rendah, mengindikasikan warna yang lebih gelap atau pucat. Perbedaan ini berpotensi menjadi indikator visual yang bisa dimanfaatkan dalam klasifikasi otomatis menggunakan algoritma pembelajaran mesin.

Proses Pra-pemrosesan

1. **Pengubahan Ukuran (Resize)**

   Seluruh gambar diubah ukurannya menjadi **128×128 piksel** guna memastikan dimensi input yang konsisten pada setiap citra.

2. **Konversi ke Grayscale**
   Gambar dikonversi ke dalam skala abu-abu untuk menyederhanakan informasi visual, mengurangi kompleksitas data, serta menurunkan jumlah saluran warna dari tiga (RGB) menjadi satu.
3. **Perataan (Flattening)**
   Gambar yang telah dikonversi ke grayscale kemudian diratakan menjadi vektor satu dimensi sepanjang **16.384 piksel**, yang merepresentasikan nilai intensitas dari setiap piksel.
4. **Normalisasi**
   Setiap nilai piksel dinormalisasi ke dalam rentang **[0, 1]** menggunakan metode **MinMaxScaler**. Langkah ini bertujuan untuk menyamakan skala nilai input dan mempercepat proses konvergensi pada algoritma pembelajaran.
5. **Penghapusan Outlier**
   Untuk meningkatkan kualitas data, dilakukan penghapusan outlier berdasarkan nilai **Z-score**, dengan ambang batas $|z| > 3$.
6. **Penyandian Label (Label Encoding)**
   Label kelas dikodekan secara biner, yaitu **Anemia = 1** dan **Non-Anemia = 0**.
7. **Oversampling**
   Mengingat terdapat ketidakseimbangan jumlah data antar kelas (kelas Anemia lebih dominan), diterapkan teknik **Random Oversampling** pada data pelatihan untuk menyeimbangkan distribusi kelas.

3. Metodologi

Pipeline Pengolahan

1. Melakukan preprocessing gambar (resize → grayscale → flatten → normalisasi).
2. Membagi dataset menjadi 80% data pelatihan dan 20% data pengujian.
3. Melakukan penghapusan outlier pada data pelatihan.
4. Menerapkan random oversampling pada data pelatihan untuk menyeimbangkan distribusi kelas.
5. Melatih model Random Forest menggunakan data pelatihan.
6. Mengevaluasi performa model menggunakan:
   - Classification Report (Precision, Recall, F1-score)
   - Confusion Matrix
   - 5-Fold Cross-Validation

Model: Random Forest

Random Forest dipilih karena karakteristik berikut:

- Robust terhadap noise dan outlier.
- Tidak bergantung pada asumsi distribusi data.
- Dapat menangani data tabular seperti fitur-fitur sederhana dari citra.

- Mudah diinterpretasikan dan efisien untuk klasifikasi biner.

Parameter Model

- **n estimators = 100**
- **random_state = 42**
- Tidak dilakukan tuning hyperparameter lanjutan karena hasil default sudah menunjukkan performa optimal.

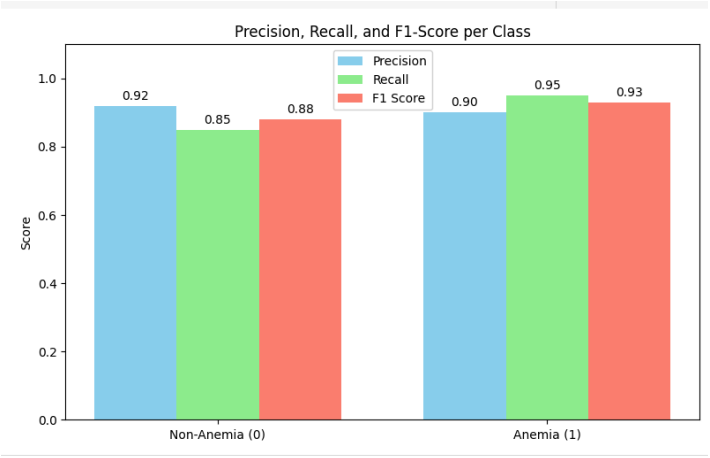
4. Experiments and Results

Evaluation Metrics – Cross-Validation (5-Fold)

Untuk mengevaluasi performa model secara menyeluruh dan menghindari bias akibat pembagian data, digunakan metode 5-Fold Cross-Validation. Proses ini membagi data pelatihan menjadi lima bagian, di mana empat bagian digunakan untuk pelatihan dan satu bagian untuk validasi, secara bergantian.

Rata-rata hasil dari kelima iterasi evaluasi disajikan pada tabel berikut:

```
Cross-Validation Results (5-Fold):
Accuracy:  0.8588 ± 0.0093
Precision: 0.8595 ± 0.0135
Recall:    0.9141 ± 0.0084
F1 Score:  0.8859 ± 0.0067
```



Model menunjukkan performa yang konsisten dengan nilai *recall* yang tinggi. Hal ini mengindikasikan bahwa model cukup sensitif dalam mendeteksi kasus anemia, meskipun terdapat sedikit kompromi pada nilai *precision* yang cenderung lebih rendah.

Evaluation Metrics – Final Model on Test Data

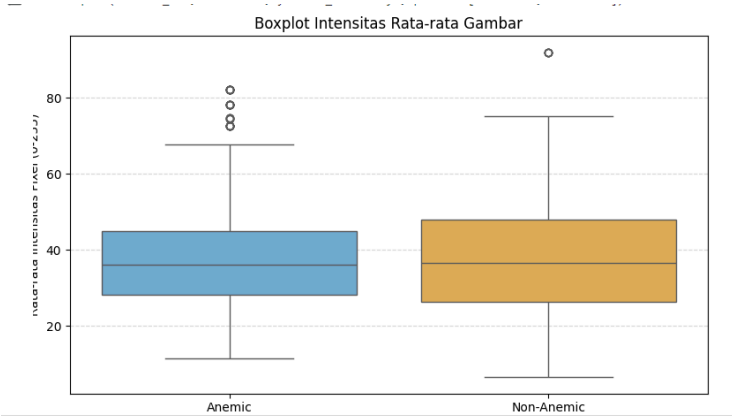
Setelah model dilatih secara menyeluruh menggunakan seluruh data latih, evaluasi dilakukan terhadap data uji untuk menilai performa model dalam kondisi nyata.

Confusion Matrix

	Predicted: Non-Anemia	Predicted: Anemia
Actual: Non-Anemia	299	44
Actual: Anemia	30	483

Total sampel: **856**

Boxplot Intensitas Pixel



Visualisasi ini memperlihatkan perbedaan persebaran antara kedua kelas. Kelas *Anemia* memiliki nilai median intensitas yang lebih rendah serta sebaran yang lebih sempit dibandingkan dengan kelas *Non-Anemia*. Terlihat pula adanya beberapa *outlier*, terutama pada kelas *Anemia*, yang mencerminkan adanya variabilitas data. Hal ini kemungkinan disebabkan oleh gangguan pencahayaan atau kesalahan saat proses akuisisi citra.

Classification Report

```
Classification Report:
              precision    recall  f1-score   support

           0       0.91      0.87      0.89       343
           1       0.92      0.94      0.93       513

    accuracy                           0.91       856
   macro avg       0.91      0.91      0.91       856
weighted avg       0.91      0.91      0.91       856
```

Analisis:

Model menunjukkan performa yang seimbang dalam mendeteksi kedua kelas, dengan nilai akurasi akhir mencapai 91%. Precision dan recall tinggi pada kelas Anemia menunjukkan model sangat efektif untuk deteksi kasus anemia.

5. Diskusi

Analisis Hasil dan Kesalahan Klasifikasi

Berdasarkan hasil evaluasi model, diketahui bahwa sebagian besar kesalahan klasifikasi terjadi pada gambar kelas Non-Anemia yang diklasifikasikan sebagai Anemia (*false positive*). Artinya, model cenderung lebih sensitif dalam mendeteksi keberadaan anemia, namun memiliki kecenderungan untuk salah mengklasifikasikan individu sehat sebagai penderita anemia.

Beberapa kemungkinan penyebab dari kesalahan ini antara lain:

- Distribusi nilai intensitas piksel yang tumpang tindih antar kelas. Hasil eksplorasi data menunjukkan adanya kemiripan distribusi rata-rata piksel antara citra kelas Anemia dan Non-Anemia. Hal ini menyulitkan model dalam membedakan karakteristik visual yang benar-benar spesifik.
- Pencahayaan yang tidak merata pada citra. Beberapa citra menunjukkan pencahayaan yang kurang seragam, yang dapat memengaruhi nilai rata-rata intensitas piksel yang diekstraksi sebagai fitur utama.
- Variasi alami antar individu atau perangkat perekam. Perbedaan perangkat perekam citra (misalnya, kamera mikroskop digital atau kondisi pencahayaan laboratorium) serta variasi fisiologis antar individu dapat menghasilkan pola visual yang tumpang tindih antara kelas.

Kendala yang Dihadapi

- Ukuran Dataset Terbatas
  Dataset yang digunakan tergolong kecil untuk tugas klasifikasi citra berbasis pembelajaran mesin. Jumlah gambar yang terbatas membuat model rentan terhadap *overfitting*, terutama jika data tidak cukup mewakili variasi kondisi nyata.
- Kualitas Gambar Tidak Konsisten
  Tidak semua citra memiliki kualitas yang baik. Beberapa citra tampak buram, terlalu terang, atau terlalu gelap. Kualitas yang tidak seragam ini mempengaruhi akurasi proses ekstraksi fitur dan performa model secara keseluruhan.
- Pra-pemrosesan yang Terbatas
  Tahapan pra-pemrosesan citra yang dilakukan hanya mencakup konversi ke skala abu-abu (*grayscale*) dan pengubahan ukuran (*resizing*). Belum diterapkan teknik peningkatan kualitas citra (seperti *histogram equalization*), augmentasi data, atau normalisasi lanjutan yang dapat meningkatkan kualitas input bagi model.

6. Kesimpulan

Hasil model ini menunjukkan bahwa fitur visual sederhana, seperti rata-rata intensitas piksel dari citra darah, dapat dimanfaatkan untuk melakukan klasifikasi anemia secara otomatis. Dengan memanfaatkan algoritma *Random Forest*, model yang dikembangkan berhasil mencapai akurasi hingga 91%, dengan nilai *precision* dan *recall* yang relatif seimbang.

Hal ini mengindikasikan bahwa, meskipun fitur yang digunakan sangat sederhana, model tetap mampu membedakan antara kondisi Anemia dan Non-Anemia dengan tingkat akurasi yang tinggi. Oleh karena itu, pendekatan ini berpotensi untuk diterapkan sebagai sistem skrining awal, khususnya di wilayah dengan keterbatasan fasilitas medis.