Comparative Analysis of Random Forest and Logistic Regression Methods in Predicting Leukemia Blood Cancer Using Microscopic Blood Cell Images

Galuh Wira Relungwangi[1]*, Jepri Banjarnahor[2]
Department of Information System, Faculty of Science and Technology[1] [2]
Universitas Prima Indonesia
Medan, Indonesia
galuhorlando26@gmail.com[1], jepribanjanahor@unprimdn.ac.id[2]

Abstract— **Leukemia is one of the deadliest blood cancers that urgently requires early detection for effective treatment. However, conventional diagnosis methods are often subjective, time-consuming, and expensive, posing challenges especially in resource-constrained areas. This study presents a comprehensive comparative analysis of two widely-used machine learning algorithms - Random Forest (RF) and Logistic Regression (LR) - for leukemia prediction using an open-access dataset of 10,661 preprocessed microscopic blood cell images from Kaggle. The dataset was carefully partitioned into training (80%) and testing (20%) sets, with rigorous preprocessing including image normalization and feature extraction. Our evaluation incorporated multiple performance metrics: accuracy, sensitivity, specificity, and AUC. The results show that Random Forest's performance is superior with a classification accuracy of 85.23%, specificity of 0.9351, sensitivity of 0.6774, and AUC of 0.8881, significantly outperforming LR which achieved an accuracy of 78.11%, specificity of 0.8363, sensitivity of 0.6742, and AUC of 0.8120. These findings suggest that ensemble methods like RF are particularly well-suited for detecting one of the most deadly blood cancers, leukemia, due to their ability to handle complex feature interactions in medical imaging data. While both algorithms have potential as clinical decision support, future research can test deep learning techniques and larger datasets to improve the accuracy and reliability of the model.**

Keywords— Random Forest, Logistic Regression, Prediction, Leukemia, Google Colab

I. INTRODUCTION

Leukemia is a type of blood cancer caused by uncontrolled growth of white blood cells in the bone marrow [1]. This disease can affect anyone, both children and adults, and is considered fatal if not detected early. According to the World Health Organization (WHO), leukemia is among the ten most deadly cancers in the world [2]. In Indonesia, based on Globocan 2020 data, leukemia is ranked 8th with more than 11,000 new cases each year [3]. Therefore, rapid and accurate early detection efforts are very important to reduce mortality and increase the effectiveness of treatment. Leukemia detection currently relies heavily on laboratory methods such as blood cell morphology by microscopy, cytogenetics, and immunophenotyping, which are often subjective, invasive, and require significant costs and time [4, 5]. On the other hand, the limited availability of specialists, especially in remote areas, is a major challenge in establishing an early diagnosis. In addition, the complexity of medical data involving various hematological parameters and microscopic images often complicates the manual analysis process [6].

Advances in artificial intelligence (AI), especially machine learning (ML), have opened up significant opportunities in the development of data-driven prediction systems for medical applications. Various algorithms have been applied to detect and predict diseases, including leukemia. Two algorithms that are often used in disease classification are Random Forest (RF) and Logistic Regression (LR) [7, 8]. Random Forest is known for its ability to handle large and complex datasets with high performance, while Logistic Regression offers good model interpretability, especially for linear data [9].

Although both methods have been used in various studies, direct comparison of RF and LR performance in the case of leukemia prediction is still limited. Several studies have shown that RF has high accuracy in detecting leukemia based on hematological features or microscopic images, with an accuracy reaching more than 95% [10, 11]. Narayanan et al. (2025) developed a hybrid method combining Fuzzy C-Means (FCM) for image segmentation and Random Forest (RF) for acute leukemia classification. By processing a dataset of microscopic images (~8,637 images), their model achieved 99.06% accuracy, 99.4% sensitivity, and 97.8% specificity. These results confirm the ability of RF to effectively handle image complexity and feature interactions [12]. Khan et al. (2021) conducted a study on the use of machine learning algorithms for leukemia classification using microscopic images. They found that the Random Forest-based model provided higher accuracy compared to the traditional model, showing great potential in clinical applications [13]. A study by Fernández-Delgado et al. in BMC Bioinformatics (2018) compared the performance of Random Forest (RF) and Logistic Regression (LR) in 243 real-world datasets. The benchmark results concluded that RF was superior in ~69% of cases. The average difference in accuracy was +0.029, the AUC increased by +0.041, and the Brier score decreased by -0.027—all significantly favoring the performance of RF [14]. However,

p-ISSN 2301-7988, e-ISSN 2581-0588
DOI : 10.32736/sisfokom.v14i3.2393, Copyright ©2025
Submitted : Mey 22, 2025, Revised : June 13, 2025, Accepted : June 17, 2025, Published : July 28, 2025

Logistic Regression remains relevant in certain contexts, especially when used on smaller and cleaner datasets [15]. Mahmood & Kadir (2025) applied variations of Logistic Regression with Ridge, Lasso, and especially ElasticNet regularization on gene expression data (16,383 genes, 281 samples, seven leukemia subtypes). The model with ElasticNet proved to be the most superior, with high accuracy and AUC and more efficient gene selection capabilities. These findings support the use of robust and interpretive LR especially in high-dimensional data [16]. However, in general, more in-depth comparative studies are still needed to directly evaluate the performance of both in the case of leukemia prediction based on microscopic blood cell images.

This study aims to fill this gap by comparing the performance of Random Forest and Logistic Regression methods in detecting leukemia based on available medical datasets. Evaluation will be conducted using various performance metrics such as accuracy, sensitivity, specificity, and AUC (Area Under Curve). The results of this study are expected to be the basis for the development of a medical decision support system that can assist in early, rapid, and accurate leukemia diagnosis, especially in areas with limited resources.

II. RESEARCH METHODOLOGY

This research falls under the category of a quantitative comparative study designed to compare the performance of two machine learning models, namely Random Forest and Logistic Regression, in predicting the likelihood of leukemia based on classification data. The research method employed is a quantitative experiment utilizing secondary data, which is subsequently analyzed using statistical and computational techniques via open-source software.

This study uses a population consisting of all medical records of patients with clinical characteristics related to leukemia. The population data includes various diagnostic parameters such as blood test results, microscopic images of white blood cells, and various other supporting laboratory examinations that serve as indicators of leukemia diagnosis.

The samples used in this study are taken from a public dataset titled “Leukemia Classification” which is available on the Kaggle platform and can be accessed via the link: *https://www.kaggle.com/datasets/andrewmvd/leukemia-classification*. This dataset has gone through a preprocessing stage, where images are converted into feature vectors, and consists of two main categories: leukemia and non-leukemia. The total data used is 212, each consisting of 106 leukemia data and 106 non-leukemia data, so this dataset is balanced.

Key instruments include: (1) Programming tools (Google Colab, Python, scikit-learn for algorithm implementation); (2) Visualization libraries (Matplotlib for ROC curves); and (3) Performance metrics (accuracy, sensitivity, specificity, AUC) for model evaluation. The scikit-learn library specifically facilitates both Logistic Regression and Random Forest model development and testing.

The general steps of the Logistic Regression (LR) process begin with modeling the linear relationship between input and output variables, then calculating the output probability using the sigmoid/logistic function. This model estimates the weight parameters using optimization, such as gradient descent, to minimize a loss function, such as log-loss.

While in Random Forest (RF), the process begins by building many decision trees randomly from a subset of data and features. Each tree provides a prediction, and the final result is obtained based on a majority vote. The advantage of RF lies in its ability to reduce overfitting and handle complex feature interactions through an ensemble approach of various decision tree models.

The scikit-learn library specifically facilitates the development and testing of both models efficiently in a Python-based computing environment.

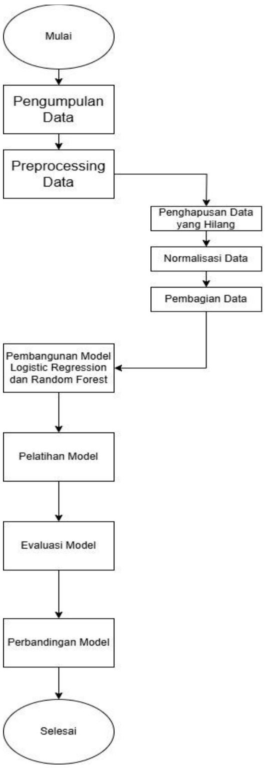


Fig 1. Leukemia Blood Cancer Prediction Flowchart

The flowchart illustrates the research process starting from collecting microscopic blood cell image data, followed by data preprocessing which includes cleaning missing data, normalization, and data sharing. Furthermore, model development is carried out using the Random Forest and Logistic Regression algorithms, followed by model training and evaluation. The final stage is the calculation of the performance results of the two models to determine the most effective method in predicting leukemia.

A. Data Collection

Data on patients with leukemia diagnosis were collected from public health data centers that provide complete and valid datasets related to patients' medical conditions. The data used includes blood test results and relevant patient medical records as indicators of leukemia diagnosis.

B. Data Preprocessing

Before training the model, the data will undergo a preprocessing stage that includes:

1. Removal of Incomplete Data: Entries with missing or incomplete values will be removed or filled using the average value. Entries with empty or incomplete values are deleted or imputed using the average value. This step is important because missing data can distort the feature distribution, disrupt the model learning process, and reduce prediction accuracy. For example, if an important feature such as white blood cell count is unavailable, the algorithm cannot accurately recognize important patterns associated with leukemia.
2. Data Normalization: Numerical features will be normalized to have a uniform scale, to prevent features with large value ranges from dominating the model. Numeric features are normalized to be on a uniform scale using the StandardScaler method. Normalization is necessary to prevent the dominance of features with large values over the model, especially in algorithms sensitive to data scale, such as Logistic Regression. In addition, normalization helps speed up convergence during training, since all features are in a similar range of values, thus reducing gradient variance and speeding up the optimization process.
3. Dataset Split: The dataset is divided into two parts, namely 80% as training data and 20% as test data using the train_test_split function from the scikit-learn library. This division aims to separate the training and testing processes to avoid overfitting, as well as to evaluate the model's generalization ability to new data that has never been seen before.

C. Model Development

The Logistic Regression model was built using the scikit-learn library through the LogisticRegression function, where the default parameters of the function were used for model building. Meanwhile, the Random Forest model is built by utilizing the RandomForestClassifier() function from the same library. The number of decision trees in the Random Forest model will be set based on experimental results and initial evaluation to obtain optimal performance.

D. Model Training

The developed model will be trained using training data in order to learn the patterns contained in the data. This training process allows the model to develop a prediction function that is used to compare the level of accuracy and performance between the Random Forest and Logistic Regression methods in predicting leukemia blood cancer diseases.

E. Model Evaluation

Once the training process is complete, model evaluation is performed using test data that was not previously used during training. The model will be assessed based on various performance metrics, including: accuracy, sensitivity, specificity and AUC (Area Under Curve).

F. Model Comparison

After both models have been tested and evaluated, the results of each model will be compared to determine which algorithm is most effective in predicting leukemia-type blood cancers.

Data analysis in this study was conducted using various statistical techniques and machine learning model evaluation. Accuracy measures the frequency with which the model provides correct predictions. Sensitivity (recall) assesses the ability of the model to detect leukemia cases (positive), while specificity measures the ability of the model to recognize non-leukemia data (negative). In addition, AUC (Area Under Curve) is used to assess the overall performance of the model based on the ROC curve, which illustrates the balance between sensitivity and specificity [17][18][19].

a). Accuracy: Measures how often the model makes correct predictions.

$$Accuracy = \frac{(TP + TN)}{(TP + TN + FP + FN)} \quad (1)$$

- TP = True Positive (the number of correct positive predictions)
- TN = True Negative (the number of correct negative predictions)
- FP = False Positive (the number of incorrect positive predictions)
- FN = False Negative (the number of incorrect negative predictions)

b). Sensitivity (Recall): Measures how well the model identifies leukemia cases (positive).

$$Sensitivity\ (Recall) = \frac{TP}{(TP + FN)} \quad (2)$$

- TP = True Positive (the number of correctly predicted positive instances)
- FN = False Negative (the number of incorrectly predicted negative instances; the model predicted negative, but the actual class was positive)

c). Specificity: Measures how well the model identifies non-leukemia data (negative).

$$Specificity = \frac{TN}{TN + FP} \quad (3)$$

- TN = True Negative (the number of correctly predicted negative instances; the model predicted negative, and the actual class was negative)
- FP = False Positive (the number of incorrectly predicted positive instances; the model predicted positive, but the actual class was negative)

d). AUC (Area Under Curve): Measures the overall performance of the model based on the ROC curve.

p-ISSN 2301-7988, e-ISSN 2581-0588
DOI : 10.32736/sisfokom.v14i3.2393, Copyright ©2025
Submitted : Mey 22, 2025, Revised : June 13, 2025, Accepted : June 17, 2025, Published : July 28, 2025

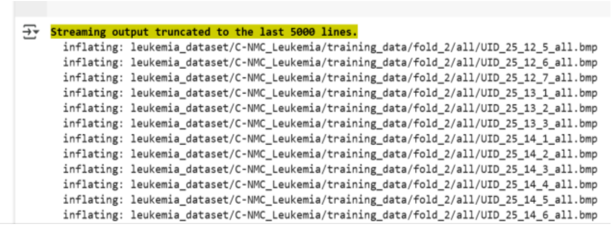
- TPR (True Positive Rate): The ratio of true positives detected out of all data that are actually positive
- FPR (False Positive Rate): The ratio of false positives out of all data that are actually negative.
- AUC (Area Under the Curve): The area under the ROC curve, which plots TPR vs FPR at various classification thresholds.

III. RESULT AND DISCUSSION

At this stage, an evaluation is conducted on the performance of two machine learning algorithms, namely Random Forest (RF) and Logistic Regression (LR), in predicting leukemia. The evaluation is performed using the metrics of accuracy, sensitivity (recall), specificity, and AUC (Area Under Curve), based on the confusion matrix obtained from the model prediction results.

A. Data Collection Results

The dataset used in this study was obtained from the Kaggle website under the name "Leukemia Classification," compiled by Kaggle user andrewmvd. This dataset contains microscopic images of blood cells that have been classified to detect types of leukemia.

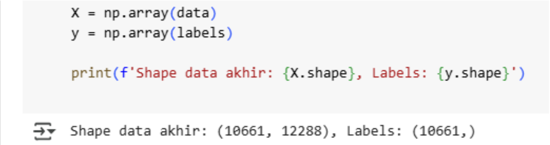


```
Streaming output truncated to the last 5000 lines.
  inflating: leukemia_dataset/C-NMC_Leukemia/training_data/fold_2/all/UID_25_12_5_all.bmp
  inflating: leukemia_dataset/C-NMC_Leukemia/training_data/fold_2/all/UID_25_12_6_all.bmp
  inflating: leukemia_dataset/C-NMC_Leukemia/training_data/fold_2/all/UID_25_12_7_all.bmp
  inflating: leukemia_dataset/C-NMC_Leukemia/training_data/fold_2/all/UID_25_13_1_all.bmp
  inflating: leukemia_dataset/C-NMC_Leukemia/training_data/fold_2/all/UID_25_13_2_all.bmp
  inflating: leukemia_dataset/C-NMC_Leukemia/training_data/fold_2/all/UID_25_13_3_all.bmp
  inflating: leukemia_dataset/C-NMC_Leukemia/training_data/fold_2/all/UID_25_14_1_all.bmp
  inflating: leukemia_dataset/C-NMC_Leukemia/training_data/fold_2/all/UID_25_14_2_all.bmp
  inflating: leukemia_dataset/C-NMC_Leukemia/training_data/fold_2/all/UID_25_14_3_all.bmp
  inflating: leukemia_dataset/C-NMC_Leukemia/training_data/fold_2/all/UID_25_14_4_all.bmp
  inflating: leukemia_dataset/C-NMC_Leukemia/training_data/fold_2/all/UID_25_14_5_all.bmp
  inflating: leukemia_dataset/C-NMC_Leukemia/training_data/fold_2/all/UID_25_14_6_all.bmp
```

Fig 2. Data Collection Results

Each fold (fold) in the training data folder is further divided into three main subfolders, namely all, hem, and all. Each of these subfolders represents a specific category or condition of blood cells in microscopic images. The all subfolder generally contains all cell images without any specific classification, while hem contains images of cells that indicate leukemia. The image format available in this dataset is bitmap (.bmp), which maintains high image quality and is suitable for pixel-based analysis in machine learning.

B. Data Preprocessing Results



```
X = np.array(data)
y = np.array(labels)

print(f'Shape data akhir: {X.shape}, Labels: {y.shape}')
```

```
Shape data akhir: (10661, 12288), Labels: (10661,)
```

Fig 3. Data Preprocessing Results

In the data preprocessing stage, images from the previously collected Leukemia Classification dataset are prepared through a series of processes to make them suitable and optimal for use as input in machine learning and deep learning algorithms. This process includes transforming raw images into a numeric format that can be understood by the model, such as feature vectors, so that the model can recognize visual patterns related to the classification of normal and leukemia blood cells. This process involves the following steps:

- Defining Folder Paths

  The training data used is located in the directory /content/leukemia_dataset/C-NMC_Leukemia/training_data, which consists of three folds: fold_0, fold_1, and fold_2. Each fold has two classes that is all (label 0), hem (label 1)

- Reading and Processing Images

  Each image within the folders is read using the load_img() function from TensorFlow and resized to 64x64 pixels. The images are then converted into arrays using img_to_array() and flattened into a 1-dimensional vector using the flatten() function for further processing by the model.

- Storing in Numpy Array Format

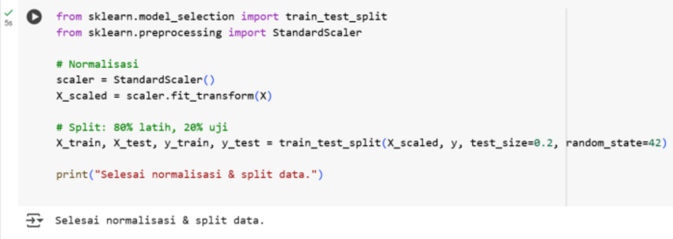
  All image data and labels are stored into two lists: data and labels, and then converted into numpy arrays (x and y), which is the standard input format for ML/DL models.

- Output of Preprocessing

  Based on the final results displayed that is total number of image data: 10,661, Dimensions of each image after flattening: 12,288 pixels (result of 64 x 64 x 3) and number of labels: 10,661.

C. Normalization and Data Splitting

Normalization in this study was performed using the StandardScaler method from the scikit-learn library, which transforms the data distribution to have a mean of 0 and a standard deviation of 1. According to Roscher et al. (2023) [20], this type of normalization is crucial, especially in the context of image-based machine learning, because it can enhance the stability and accuracy of classification models.



```
from sklearn.model_selection import train_test_split
from sklearn.preprocessing import StandardScaler

# Normalisasi
scaler = StandardScaler()
X_scaled = scaler.fit_transform(X)

# Split: 80% latih, 20% uji
X_train, X_test, y_train, y_test = train_test_split(X_scaled, y, test_size=0.2, random_state=42)

print("Selesai normalisasi & split data.")
```

```
Selesai normalisasi & split data.
```

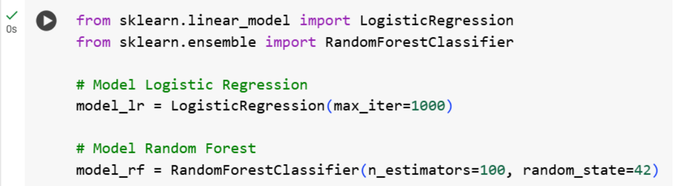
Fig 4. *Normalization and Data Splitting Result*

The figure shows the stages of data normalization and dataset splitting using the scikit-learn library. The normalization process is carried out with StandardScaler() to change the feature distribution to have a mean of 0 and a standard deviation of 1, thereby accelerating model convergence and preventing the dominance of large-scale features. After normalization, the data is split into two parts: 80% for training and 20% for testing, using the train_test_split function with the random_state=42 parameter to ensure consistent results.

Submitted : Mey 22, 2025, Revised : June 13, 2025, Accepted : June 17, 2025, Published : July 28, 2025

D. Development of Random Forest and Logistic Regression Models

The model was built using the LogisticRegression function from the scikit-learn library, with the parameter max_iter=1000 to ensure the training process reaches convergence. According to Raschka et al. (2023) [17], Logistic Regression remains a strong and reliable baseline, especially when model interpretability is a priority.



```
from sklearn.linear_model import LogisticRegression
from sklearn.ensemble import RandomForestClassifier

# Model Logistic Regression
model_lr = LogisticRegression(max_iter=1000)

# Model Random Forest
model_rf = RandomForestClassifier(n_estimators=100, random_state=42)
```

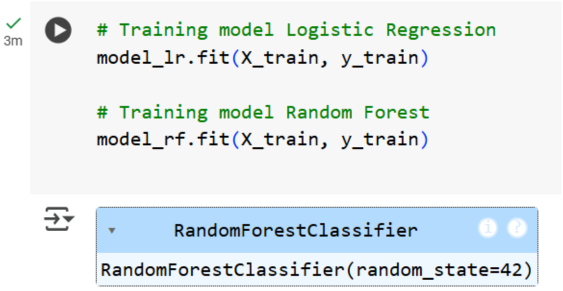
Fig 5. Development of Random Forest and Logistic Regression Models

The figure above shows the development process of two machine learning models using the scikit-learn library. The first model is Logistic Regression, which is initialized with the parameter max_iter=1000 to ensure the training process reaches convergence. The second model is Random Forest, which is built using 100 decision trees (n_estimators=100) and random_state=42 to ensure consistent and reproducible results. Both models are used as classification tools in leukemia prediction based on microscopic blood cell images. According to Liu et al. (2023) [21], Random Forest is highly effective in handling datasets with a large number of features and complex inter-feature correlations, such as in image data resulting from pixel extraction.

E. Model Training

The Logistic Regression model is trained to find the relationship between the features extracted from the image and the class label (0 for "all" and 1 for "hem"). It learns weight parameters to maximize classification accuracy based on a logistic (sigmoid) function.

The Random Forest model is trained to build a number of decision trees, where each tree learns from a randomly drawn subset of the training data. The final prediction result is a combination of the majority votes from all trees.



```
# Training model Logistic Regression
model_lr.fit(X_train, y_train)

# Training model Random Forest
model_rf.fit(X_train, y_train)
```

```
RandomForestClassifier
RandomForestClassifier(random_state=42)
```

Fig 6. Model Training

The figure above shows the model training process for two machine learning algorithms, namely Logistic Regression and Random Forest. The models are trained using training data (X_train, y_train) through the .fit() method, which allows each model to learn patterns from the data and build predictive functions. This training is an important stage before evaluating and comparing the performance between models in detecting leukemia based on microscopic blood cell images.

According to Géron (2023) [22], the training of supervised learning models is an iterative process that requires careful attention to overfitting and underfitting to ensure the model achieves optimal performance.

F. Model Evaluation

Model evaluation was performed by comparing the performance of two classification algorithms that are Logistic Regression and Random Forest.

TABLE I. CONFUSION MATRIX

Logistic Regression	1121 (True Negative))	237 (False Positive)
	230 (False Negative)	455 (True Positive)
Random Forest	1354 (True Negative)	94 94 (False Positive)
	221 ((False Negative)	464 464 (True Positive)

Table I shows the confusion matrix of the classification results by two models, namely Logistic Regression and Random Forest, on leukemia test data. In the Logistic Regression model, it appears that the number of False Positives (237) and False Negatives (230) is quite high, indicating that this model still often misclassifies data, both in detecting positive and negative cases.

In contrast, the Random Forest model shows a significant increase in correctly identifying classes, especially in the negative class, as evidenced by the higher True Negatives (1354) and much lower False Positives (94). This reflects the high specificity (0.9351) of the model.

In a clinical context, high specificity is crucial, because it means that the model has a good ability to recognize patients who do not have leukemia. This will reduce the number of false positive cases, thus avoiding misdiagnosis, unnecessary further tests, and psychological anxiety in patients. On the other hand, good enough sensitivity is also important to ensure that as many positive cases as possible are identified, although in this model the value can still be improved.

Thus, Random Forest can be considered more reliable to be applied as an early diagnostic tool, especially in clinical settings that require high accuracy in excluding non-leukemia cases.

1)Logistic Regression

a) $$Accuracy = \frac{(TP + TN)}{(TP + TN + FP + FN)}$$
$$= \frac{(455 + 1211))}{((455 + 1211 + 237 + 230))}$$
$$= 0.7811$$

b). $$Sensitivity\ (Recall) = \frac{TP}{(TP + FN)}$$
$$= \frac{455}{(455 + 230)}$$
$$= 0.6642$$

c). $Specificity = \frac{TN}{TN + FP}$

$$= \frac{1211}{1211 + 237}$$

$$= 0.8363$$

d). *AUC (Area Under Curve)* = 0.8120

2) *Random Forest*

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

$$= \frac{(464 + 1354)}{(464 + 1354 + 94 + 221)}$$

$$= 0.8523$$

b). *Sensitivity (Recall)* $= \frac{TP}{(TP + FN)}$

$$= \frac{464}{(464 + 221)}$$

$$= 0.6674$$

c). $Specificity = \frac{TN}{TN + FP}$

$$= \frac{1354}{1354 + 94}$$

$$= 0.9351$$

d). *AUC (Area Under Curve)* = 0.8881

TABLE II. MODEL EVALUATION RESULT

Logistic Regression	*Accuracy*	*Sensitivity (Recall)*	*Specificity*	*AUC (Area Under Curve)*
	0.7811	0.6642	0.8363	0.8120
Random Forest	*Accuracy*	*Sensitivity (Recall)*	*Specificity*	*AUC (Area Under Curve)*
	0.8523	0.6674	0.9351	0.8881

Based on the values in Table II, the evaluation results demonstrate that the Random Forest (RF) model exhibits superior performance compared to the Logistic Regression (LR) model across most of the evaluation metrics.

Logistic Regression achieved an accuracy of 78.11%, with a sensitivity (recall) of 66.42% and a specificity of 83.63%. The AUC value of 0.8120 indicates that this model has reasonably good classification ability; however, there is room for improvement, particularly in detecting positive cases (as shown by the relatively low sensitivity value).

On the other hand, Random Forest significantly improved the performance. This model achieved a higher accuracy of 85.23% compared to Logistic Regression. Its sensitivity is slightly better at 66.74%, while its specificity increased substantially to 93.51%. The AUC value of 0.8881 indicates a stronger overall classification capability.

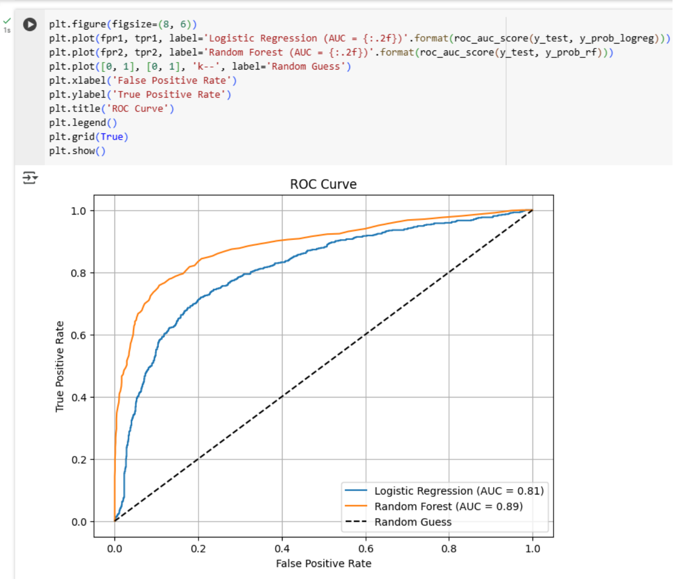
The technical advantages of Random Forest over Logistic Regression can be explained as follows:

- Random Forest works as an ensemble method that builds multiple decision trees and combines their results through majority voting. This approach makes the model more resistant to overfitting and is able to capture non-linear interactions between features, which are common in medical image data such as blood cells.

- Logistic Regression, on the other hand, works with the assumption of linearity between input and output variables. This becomes a limitation when the patterns in the data are non-linear, as is the case in the pixel representation of microscopic images.

G. AUC (Area Under Curve)

AUC-ROC curves of leukemia prediction results using two models, namely Logistic Regression and Random Forest. The X-axis shows the False Positive Rate (1 – Specificity), while the Y-axis shows the True Positive Rate (Sensitivity). In the graph, the orange colored curve represents the Random Forest model, while the blue colored curve represents the Logistic Regression model.

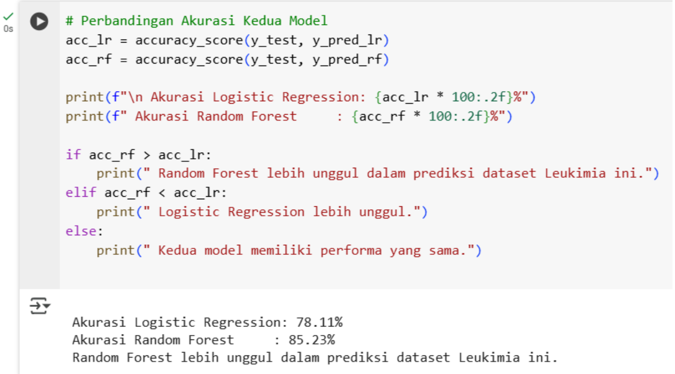


```
plt.figure(figsize=(8, 6))
plt.plot(fpr1, tpr1, label='Logistic Regression (AUC = {:.2f})'.format(roc_auc_score(y_test, y_prob_logreg)))
plt.plot(fpr2, tpr2, label='Random Forest (AUC = {:.2f})'.format(roc_auc_score(y_test, y_prob_rf)))
plt.plot([0, 1], [0, 1], 'k--', label='Random Guess')
plt.xlabel('False Positive Rate')
plt.ylabel('True Positive Rate')
plt.title('ROC Curve')
plt.legend()
plt.grid(True)
plt.show()
```

Fig 7. AUC-ROC Curve

Based on the ROC curve, it was found that the Random Forest model obtained an AUC value of 0.89, which means that this model is able to distinguish between positive and negative samples with an accuracy of 89%. Meanwhile, the Logistic Regression model obtained an AUC of 0.81, showing a lower performance than Random Forest in leukemia classification.

H. Comparison of Random Forest and Logistic Regression Models



```
# Perbandingan Akurasi Kedua Model
acc_lr = accuracy_score(y_test, y_pred_lr)
acc_rf = accuracy_score(y_test, y_pred_rf)

print(f"\n Akurasi Logistic Regression: {acc_lr * 100:.2f}%")
print(f" Akurasi Random Forest     : {acc_rf * 100:.2f}%")

if acc_rf > acc_lr:
    print(" Random Forest lebih unggul dalam prediksi dataset Leukimia ini.")
elif acc_rf < acc_lr:
    print(" Logistic Regression lebih unggul.")
else:
    print(" Kedua model memiliki performa yang sama.")
```

```
Akurasi Logistic Regression: 78.11%
Akurasi Random Forest     : 85.23%
Random Forest lebih unggul dalam prediksi dataset Leukimia ini.
```

Fig 8. Comparison Results of Random Forest and Logistic Regression Models

Based on these accuracy values, it can be concluded that the Random Forest model demonstrates better performance in making predictions on the leukemia dataset compared to Logistic Regression. The figure shows the accuracy between two model classifications, namely Logistic Regression and Random Forest, based on the prediction results on the test data. The Random Forest model shows a higher accuracy (85.23%) compared to Logistic Regression (78.11%). Based on this comparison, Random Forest is declared superior in making predictions on the leukemia dataset, indicating that this model has better classification performance on microscopic blood cell image data.

Although the results obtained show superior performance of Random Forest, this study has several limitations:

- The dataset used only includes two classes (leukemia and non-leukemia), without considering more complex leukemia subtypes.
- Feature extraction was performed manually and based on raw pixels, rather than using high-level feature-based techniques such as CNN or PCA.
- Computational time or efficiency analysis has not been performed, which may affect the choice of algorithm in clinical practice.

For further research, it is recommended to use deep learning techniques such as Convolutional Neural Networks (CNN) to enable automatic feature extraction from more complex cell images. In addition, the classification can be extended to the level of leukemia subtypes such as Acute Lymphoblastic Leukemia (ALL) and Acute Myeloid Leukemia (AML) to improve diagnostic specificity. Cross-validation and testing on external datasets are also important to test the generalization ability of the model. Finally, comparing additional algorithms such as SVM, XGBoost, and deep neural networks can enrich the analysis of classification performance.

IV. CONCLUSION

The conclusion is that both algorithms, Random Forest (RF) and Logistic Regression (LR), are able to predict leukemia based on microscopic images of blood cells, but RF shows superior performance. With an accuracy of 85.23%, sensitivity of 67.74%, specificity of 93.51%, and AUC of 0.8881, RF is proven to be more reliable than LR which only recorded an accuracy of 78.11%, sensitivity of 66.42%, specificity of 83.63%, and AUC of 0.8120. The main advantage of Random Forest (RF) lies in its ability to handle complex feature interactions and minimize false negative errors, which are crucial in medical diagnosis. These findings support the research objective to compare two classification methods in leukemia prediction, and provide empirical evidence that ensemble algorithms such as RF are more effective for medical image classification. Practically, RF is worthy of consideration as a component of clinical decision support systems, especially in areas with limited resources. With its ability to provide fast, accurate, and affordable automated diagnosis, the RF model can help accelerate early detection and improve the chances of successful treatment in leukemia patients.

REFERENCES

[1] A. E. Aby, S. Salaji, K. K. Anilkumar, and T. Rajan, “A review on leukemia detection and classification using Artificial Intelligence-based techniques,” Comput. Biol. Med., vol. 169, p. 107630, 2024.

[2] J. Ferlay, M. Colombet, I. Soerjomataram, D. M. Parkin, M. Piñeros, A. Znaor, and F. Bray, "Cancer statistics for the year 2020: An overview," Int. J. Cancer, vol. 149, no. 4, pp. 778–789, Aug. 2021, doi: 10.1002/ijc.33588.

[3] Z. Cheng et al.,“Artificial intelligence reveals the predictions of hematological indexes in children with acute leukemia,” BMC Cancer, vol. 24, p. 993, 2024.

[4] K. Kou et al., “Comprehensive Sepsis Risk Prediction in Leukemia Using a Random Forest Model and Restricted Cubic Spline Analysis” Front. Immunol., vol. 16, p. 1514273, 2025.

[5] H. Liao, F. Zhang, F. Chen, Y. Li, Y. Sun, D. D. Sloboda, Q. Zheng, B. Ying, and T. Hu, "Application of artificial intelligence in laboratory hematology: Advances, challenges, and prospects," Clinica Chimica Acta, vol. 550, pp. 1–10, 2024, doi: 10.1016/j.cca.2023.117180..

[6] L. Sari et al., “Penerapan Random Forest untuk Prediksi Penyakit Berdasarkan Data Hematologi,” J. Infotekmesin, vol. 14, no. 1, pp. 45–52, 2024.

[7] H. Li et al., “Integrating Random Forest and Logistic Regression for Disease Classification: A Hybrid Approach,” Heliyon, vol. 10, no. 3, p. e25369, 2024.

[8] K. Kashef et al., “Comparative analysis of ML algorithms for leukemia detection,” BMC Med. Inform. Decis. Mak., vol. 24, p. 122, 2024.

[9] S. Triglycerides, “Leukemia Detection Using CNN and ML Approaches,” Indones. J. Comput. Sci., vol. 13, no. 3, pp. 4115–4125, 2024.

[10] X. Fu et al., “ML models for acute leukemia detection: Performance comparison,” BMC Cancer, vol. 24, p. 993, 2024.

[11] L. Narayanan, K. Santhana, R. Harold, and M. A. Banu, “Enhancing Acute Leukemia Classification Through Hybrid Fuzzy C-Means and Random Forest Methods,” Meas. Sens., vol. 39, p. 101876, 2025.

[12] M. A. Khan, M. Sharif, M. Raza, and T. Saba, “A Machine Learning Framework for the Classification of Leukemia Using Microscopic Blood Images,” Microsc. Res. Tech., vol. 84, no. 12, pp. 2917–2928, 2021.

[13] M. Fernández-Delgado, E. Cernadas, S. Barro, and D. Amorim, “Random Forest versus Logistic Regression: A Large-Scale Benchmark Experiment,” BMC Bioinform., vol. 19, p. 270, 2018.

[14] L. Liu, Y. Zhang, and H. Li, “Application of logistic regression in hematological diagnosis,” BMC Cancer, vol. 24, p. 993, 2024.

[15] N. H. Mahmood and D. H. Kadir, “Sparsity Regularization Enhances Gene Selection and Leukemia Subtype Classification via Logistic Regression,” Leuk. Res., vol. 150, p. 107663, 2025.

[16] R. Roscher, B. Bohn, P. Feth, and B. Waske, “Explainable Machine Learning for Remote Sensing Applications,” IEEE Trans. Geosci. Remote Sens., vol. 61, pp. 1–20, 2023.

p-ISSN 2301-7988, e-ISSN 2581-0588
DOI : 10.32736/sisfokom.v14i3.2393, Copyright ©2025
Submitted : Mey 22, 2025, Revised : June 13, 2025, Accepted : June 17, 2025, Published : July 28, 2025

[17] S. Raschka, V. Mirjalili, and M. Khan, Python Machine Learning, 4th ed. Birmingham, UK: Packt Publishing, 2023.

[18] Y. Liu, H. Yin, and Y. Zhang, “Comparative Study of Machine Learning Algorithms for Medical Image Classification,” J. Healthc. Eng., vol. 2023, pp. 1–10, 2023.

[19] A. Géron, Hands-On Machine Learning with Scikit-Learn, Keras, and TensorFlow, 3rd ed. Sebastopol, CA: O’Reilly Media, 2023.

[20] A. M. Najar, I. W. Sudarsana, M. U. Albab, and S. Andhika, “Machine Learning untuk Identifikasi Jenis Kanker Darah (Leukemia),” Vygotsky: J. Pendidik. Mat. dan Matematika, vol. 4, no. 1, pp. 47–56, 2022.

[21] Y. Maulana, A. P. Nugroho, and D. D. Prasetyo, “Evaluasi Performa Machine Learning untuk Analisis Leukosit Abnormal Darah Tepi pada Penderita Acute Lymphoblastic Leukemia,” Laporan Penelitian, Universitas Gadjah Mada, 2022

p-ISSN 2301-7988, e-ISSN 2581-0588
DOI : 10.32736/sisfokom.v14i3.2393, Copyright ©2025
Submitted : Mey 22, 2025, Revised : June 13, 2025, Accepted : June 17, 2025, Published : July 28, 2025