Detection and Classification of Immature Leukocytes for Diagnosis of Acute Myeloid Leukemia

Satvik Dasariraju, The Lawrenceville School

Introduction

Acute Myeloid Leukemia (AML) is the deadliest type of leukemia, accounting for 11,000 deaths annually in the US with an average five year survival rate of 28.7% [1]]. AML progresses quickly and is fatal within weeks if not detected and treated immediately.

The current diagnostic method is microscopic examination and complete blood count to classify leukocytes. Just an initial assessment of each blood smear takes over 3 minutes; the manual examination is inefficient and inaccurate as there is a 30-40% error rate [2,3]. Especially in developing countries, where diagnosis of AML takes multiple weeks, an automated approach would aid greatly.

Previous studies aiming to detect and classify AML were limited by small or unbalanced data sets, low accuracy, a low amount of cell features used, or low classification performance despite high detection accuracy [3,4,5,6,7]

Purpose and Hypothesis

Different types of leukocytes vary in cytomorphology, so detection and classification of immature leukocytes can be formulated as a machine learning classification task based on cytomorphological features

To overcome the limitations of the manual diagnostic method and the shortcomings of previous research, this project in biomedical image analysis aimed to:

1) Develop a model capable of accurate detection and classification of 4 types of immature leukocytes in AML cells
2) Calculate and identify the most important features for classification of leukocytes

The hypothesis is that the superior performance of a Random Forest classifier with unbalanced data will lead to quick and accurate detection and classification of immature leukocytes.

Materials

Images of leukocytes in patients with AML and healthy controls were obtained from a publicly available data set in The Cancer Imaging Archive [8,9]. 1,070 images (457 mature and 613 immature leukocytes) were used for the detection of immature leukocytes, while 613 images of immature leukocytes were used for classification into 4 types.

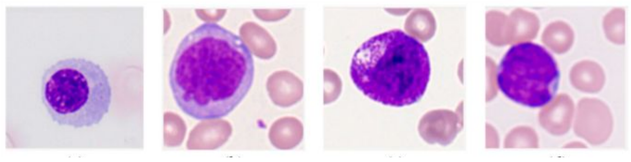

Figure 1. Samples of the four types of immature leukocytes used in the study (erythroblast, monoblast, promyelocyte, myeloblast). Images obtained from [8,9]. *All other figures and tables created by student researcher.*

Methods

The methodology of this project consisted of 4 main components: image segmentation, feature extraction, classification through machine learning, and calculation of feature importance.

Image Segmentation

The goal of segmentation was to produce binary masks of the whole cell and the isolated nucleus for each leukocyte. Image format conversion, erosion after dilation, multi-Otsu thresholding, and smoothing were used to deliver two final results for each image.

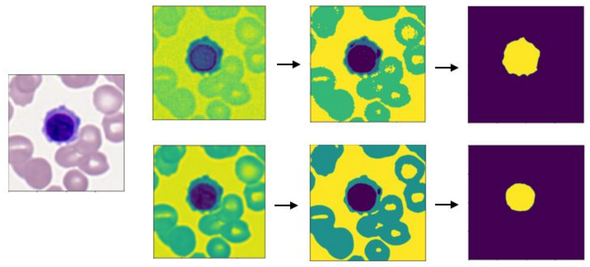

Figure 2. Sample of the segmentation process on an image of an erythroblast. Final results are binary mask of cell and isolated nucleus.

Feature Extraction

From each image, 16 cytomorphological features were extracted and inputted to a features matrix.

- 12 shape features (eg. N:C ratio)
- 4 color features, 2 of which are new color features proposed in this study:
  1) Average of B Channel Intensity of Nucleus in LAB Color Space
  2) Standard Deviation of B Channel Intensity of Nucleus in LAB Color Space

Classification with Machine Learning

Random Forest was chosen as the machine learning classifier because of its superior performance compared to k-Nearest Neighbor and Support Vector Machine for unbalanced data. Data for binary classification (detecting immature leukocytes) was split 80% and 20% into training and test sets, respectively, while a 70%-30% split was used for classification of immature leukocytes. After the initial Random Forest algorithm was trained and evaluated with the test set, it was optimized based on average precision across classes and retested.

Calculation of Feature Importance

The Gini Importance of a Random Forest Algorithm is defined as the mean reduction in the probability of a wrong classification over all nodes. In this study, Gini Importance was used to determine the five most important features for both detection and classification of immature leukocytes. This is the first study to calculate the importance of a varied collection of cell features for the classification of immature leukocytes.

Results: Detection of Immature Leukocytes

Table 1. Performance metrics of the optimized model for binary classification between immature and mature leukocytes on training and testing sets.

Set	Accuracy	Precision	Recall (Sensitivity)	Specificity
Training Set	100%	100%	100%	100%
Testing Set	92.99%	91.23%	95.41%	90.48%

The optimized model was able to differentiate immature and mature leukocytes with 92.99% accuracy, detecting 95.41% of immature leukocytes. The area under the curve of the receiver operating characteristic curve is 0.9803, displaying the model's effectiveness. The results are on par with the current state of art models for detection of immature leukocytes.

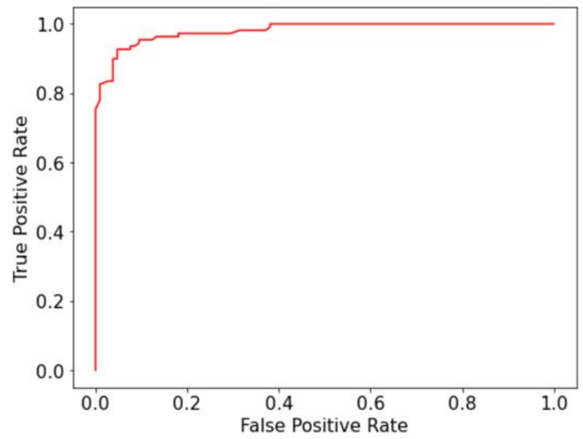

Figure 3. Receiver Operating Characteristic Curve for Binary Classification (Immature or Mature). Graph created with Matplotlib.

Results: Classification of Immature Leukocytes

Table 2. Performance metrics of the optimized model for classification of immature leukocytes.

Metric	Erythroblasts	Monoblasts	Promyeloctes	Myeloblasts
Precision	100.00%	77.78%	69.23%	97.56%
Recall	91.30%	100.00%	75.00%	96.77%

The optimized Random Forest algorithm classified the 4 types of immature leukocytes in an unbalanced dataset with 93.45% overall accuracy. Precision values for each class were above 65% and recall values for each class were at least 75%, exhibiting an improvement over the current state of art for multiclass classification of immature leukocytes, since previous models had precision values below 65% for the majority of immature leukocyte types.

Results: Most Important Features

Table 3. The five most important features for detection of immature leukocytes based on Gini Importance.

Feature	Gini Importance
N:C Ratio	0.2801
Area to Perimeter Ratio	0.1076
Nucleus Minor Axis Length	0.0829
Nucleus Major Axis Length	0.0803
Area	0.0627

Table 4. The five most important features for classification of immature leukocytes based on Gini Importance. Color intensity is calculated in B Channel of LAB color space. Asterisk indicates new color features proposed in this study.

Feature	Gini Importance
Average Nucleus Color Intensity*	0.2532
Standard Deviation of Nucleus Color Intensity*	0.1853
N:C Ratio	0.1765
Standard Deviation of Cytoplasm Color Intensity	0.0618
Average Cytoplasm Color Intensity	0.0571

N:C Ratio was calculated to be a significant discriminator for both detection and classification of immature leukocytes. For classification, the Gini Importance of the two new nucleus color features were the highest of all 16 features, while cytoplasm color features were also shown to be discriminative.

Conclusions

To overcome the limitations of the manual diagnosis methodology for AML, a Random Forest model for automatic detection and classification of immature leukocytes was presented. The model achieved accurate detection of immature leukocytes on par with the current state of art and precise classification results that were superior to previous studies. Moreover, for the first time, the importance of several cytomorphological features for the classification of leukocytes was mathematically calculated using Gini Importance.

Applications of this study are twofold:

1) The presented algorithm can be used as an effective support tool in the clinical diagnosis of AML, especially in developing countries where diagnosis takes many weeks.
2) The features calculated to be most important can serve as a basis for future researchers aiming to classify leukocytes.

Future studies can expand on this work by identifying more cytomorphological features and determining their importance. Improving the discrimination between similar cell types, such as myeloblasts and promyelocytes, is also an avenue for future work. Future research will also develop systems that can be completely integrated into the clinical diagnosis method for automatic detection of potentially cancerous cells.

References

1. "Acute Myeloid Leukemia - Cancer Stat Facts," SEER. https://seer.cancer.gov/statfacts/html/amyl.html (accessed Jul. 22, 2020).
2. A. Adewoyin and B. Nwogoh, "PERIPHERAL BLOOD FILM - A REVIEW," Ann Ib Postgrad Med, vol. 12, no. 2, pp. 71–79, Dec. 2014.
3. F. Kazemi, T. A. Najafabadi, and B. N. Araabi, "Automatic Recognition of Acute Myelogenous Leukemia in Blood Microscopic Images Using K-means Clustering and Support Vector Machine," J Med Signals Sens, vol. 6, no. 3, pp. 183–193, 2016.
4. C. Matek, S. Schwarz, K. Spiekermann, and C. Marr, "Human-level recognition of blast cells in acute myeloid leukaemia with convolutional neural networks," Nature Machine Intelligence, vol. 1, no. 11, pp. 538–544, Nov. 2019, doi: 10.1038/s42256-019-0101-9.
5. L. Bigorra, A. Merino, S. Alférez, and J. Rodellar, "Feature Analysis and Automatic Identification of Leukemic Lineage Blast Cells and Reactive Lymphoid Cells from Peripheral Blood Cell Images," J. Clin. Lab. Anal., vol. 31, no. 2, Mar. 2017, doi: 10.1002/jcla.22024.
6. E. S. Wiharto, S. Palgunadi, Y. R. Putra, and E. Suryani, "Cells identification of acute myeloid leukemia AML M0 and AML M1 using K-nearest neighbour based on morphological images," in 2017 International Conference on Data and Software Engineering (ICoDSE), Nov. 2017, pp. 1–6, doi: 10.1109/ICODSE.2017.8285851.
7. W. Wiharto, E. Suryani, and Y. R. Putra, "Classification of blast cell type on acute myeloid leukemia (AML) based on image morphology of white blood cells," TELKOMNIKA, vol. 17, no. 2, p. 645, Aug. 2018, doi: 10.12928/telkomnika.v17i2.8666.
8. C. Matek, S. Schwarz, C. Marr, and K. Spiekermann, "A Single-cell Morphological Dataset of Leukocytes from AML Patients and Non-malignant Controls [Data Set]," in The Cancer Imaging Archive. doi: 10.7937/tcia.2019.36f5o9ld.
9. K. Clark et al., "The Cancer Imaging Archive (TCIA): Maintaining and Operating a Public Information Repository," J Digit Imaging, vol. 26, no. 6, pp. 1045–1057, Dec. 2013, doi: 10.1007/s10278-013-9622-7.

Acknowledgments and Notes

This project was completed under the mentorship and guidance of Marc Huo (Stanford University) and Dr. Serena McCalla (Jericho High School) at iResearch Institute.

This research project has been peer reviewed and published in *Bioengineering* Journal:
https://doi.org/10.3390/bioengineering7040120