Original Research Article

DIGITAL HEALTH

DIGITAL HEALTH
Volume 10: 1–11
© The Author(s) 2024
Article reuse guidelines:
sagepub.com/journals-permissions
DOI: 10.1177/20552076241258079
journals.sagepub.com/home/dhj
Sage

Deep learning model for differentiating acute myeloid and lymphoblastic leukemia in peripheral blood cell images via myeloblast and lymphoblast classification

Sholhui Park[1], Young Hoon Park[2], Jungwon Huh[1], Seung Min Baik[3,*] and Dong Jin Park[4,*]

Abstract

Objective: Acute leukemia (AL) is a life-threatening malignant disease that occurs in the bone marrow and blood, and is classified as either acute myeloid leukemia (AML) or acute lymphoblastic leukemia (ALL). Diagnosing AL warrants testing methods, such as flow cytometry, which require trained professionals, time, and money. We aimed to develop a model that can classify peripheral blood images of 12 cell types, including pathological cells associated with AL, using artificial intelligence.

Methods: We acquired 42,386 single-cell images of peripheral blood slides from 282 patients (82 with AML, 40 with ALL, and 160 with immature granulocytes).

Results: The performance of EfficientNet-V2 (B2) using the original image size exhibited the greatest accuracy (accuracy, 0.8779; precision, 0.7221; recall, 0.7225; and F1 score, 0.7210). The next-best accuracy was achieved by EfficientNet-V1 (B1), with a 256 × 256 pixels image. F1 score was the greatest for EfficientNet-V1 (B1) with the original image size. EfficientNet-V1 (B1) and EfficientNet-V2 (B2) were used to develop an ensemble model, and the accuracy (0.8858) and F1 score (0.7361) were improved. The classification performance of the developed ensemble model for the 12 cell types was good, with an area under the receiver operating characteristic curve above 0.9, and F1 scores for myeloblasts and lymphoblasts of 0.8873 and 0.8006, respectively.

Conclusions: The performance of the developed ensemble model for the 12 cell classifications was satisfactory, particularly for myeloblasts and lymphoblasts. We believe that the application of our model will benefit healthcare settings where the rapid and accurate diagnosis of AL is difficult.

Keywords

Leukemia, artificial intelligence, classification, blast, peripheral blood cells

Submission date: 6 November 2023; Acceptance date: 13 May 2024

[1]Department of Laboratory Medicine, Ewha Womans University College of Medicine, Seoul, Korea
[2]Division of Hematology-Oncology, Department of Internal Medicine, Ewha Womans University Mokdong Hospital, Seoul, Korea
[3]Division of Critical Care Medicine, Department of Surgery, College of Medicine, Ewha Womans University, Seoul, Korea
[4]Department of Laboratory Medicine, Eunpyeong St Mary's Hospital, College of Medicine, The Catholic University of Korea, Seoul, Korea
[*]Co-corresponding authors.

Corresponding authors:
Seung Min Baik, Division of Critical Care Medicine, Department of Surgery, College of Medicine, Ewha Womans University, 1071, Anyangcheon-ro, Yangcheon-gu, Seoul, Korea.
Email: baiksm@ewha.ac.kr

Dong Jin Park, Department of Laboratory Medicine, Eunpyeong St Mary's Hospital, College of Medicine, The Catholic University of Korea, 1021, Tongil-ro, Eunpyeong-gu, Seoul 03312, Korea.
Email: parkdj1280@gmail.com.

Introduction

Acute leukemia (AL) is a type of life-threatening malignant disorder of the bone marrow and blood, leading to bone marrow failure and death. AL was the 11th leading cause of cancer death worldwide in 2018, accounting for 437,033 cancer cases and 309,006 cancer-related deaths.[1] The U.S. National Institute of Health reported that the five-year relative survival rate for all types of leukemia was 66.7% from 2013 to 2019.[2] While the cause of AL is unknown, several factors are associated with an increased risk of developing it, including old age, genetics, exposure to certain chemicals, smoking, previous cancer treatment, and family history.[3] ALs commonly occur in patients across all ages and are potentially rapidly fatal if not readily treated.

ALs can be divided into acute myelogenous leukemia (AML) or acute lymphoblastic leukemia (ALL). AML is the most common AL in adults, and accounts for < 10% of AL in children aged < 10 years.[2] As each type of AL has a varying disease biology, incidence, prognosis, and treatment strategy, a timely and accurate diagnosis is crucial for appropriate clinical management.[4] Currently, AL diagnosis relies on cytomorphology, cytochemistry, immunophenotyping, and molecular genetics. Most importantly, the initial morphological assessment of peripheral blood smears plays a vital role in the detection and diagnosis of AL, leading to its classification and immediate initiation of treatment. Although it is an indispensable starting point in the diagnosis of several hematologic disorders, including AL, several limitations, such as time- and labor-intensive processes, the requirement for trained personnel, and large inter-observer variation, must be considered. Immunophenotyping by flow cytometry using bone marrow specimens plays an essential role in the detection, characterization, and classification of AL.[5,6] However, it has some disadvantages such as patient discomfort during bone marrow specimen collection, cost, the need for specialized experts in cell preprocessing, and dependance on the patient's overall health. Thus, the need to develop new tools to help overcome the various disadvantages of the abovementioned tests and to assist in a more accurate diagnosis and classification of AL is imperative.

Deep learning through convolutional neural networks (CNNs) has been actively studied in the medical field. In particular, several studies using radiologic images such as X-rays, computed tomography (CT), and magnetic resonance imaging (MRI) have shown impressive and promising results.[7–9] Advances in artificial intelligence (AI) have had a significant impact in the field of hematology, particularly in the diagnosis of leukemia from peripheral blood smear images. A comprehensive review by Fan et al. highlighted the progress in this area, detailing 75 and 25 studies focused on ALL and AML, respectively.[10] Collectively, these studies achieved an impressive average classification accuracy of 96.0%. These studies primarily focused on dichotomous classification tasks, distinguishing between normal white blood cells (WBCs) and blasts, marking an initial step in leveraging AI to rapidly identify leukemia cases. However, our literature review identified a gap in exploring multiclass classification within the same domain. A key challenge for this research is to extend the capabilities of AI to accurately classify the diverse cell types in the peripheral blood smears, rather than simply distinguishing between leukemic and normal cells. This approach is essential in mimicking the comprehensive diagnostic assessment performed in clinical hematopathology, where a nuanced understanding of different cell types can lead to more accurate treatment decisions. In this study, we aimed to develop a deep learning model with sufficient performance to simultaneously distinguish between normal and immature WBC observed in peripheral blood smears, and further differentiate myeloblasts from lymphoblasts using a large body of cell image data obtained from the Sysmex DI-60 (DI-60; Sysmex, Kobe, Japan).

Methods

Study design and data collection

From January 2012 to September 2021, 282 patients from Ewha Womans University Mokdong Hospital were included in the study: 82 with AML, 40 with ALL, and 160 with immature granulocytes (IGs) in the peripheral blood. In this study, we determined the diagnoses of AML and ALL according to the World Health Organization's Classification of Hematopoietic and Lymphoid Tumors, 4th Edition. The study included patients newly diagnosed with AL. We acquired 42,386 single-cell images of peripheral blood slides from 282 patients using Sysmex DI-60. In the preparation of peripheral blood smears, the Wright-Giemsa staining technique was employed. A schematic representation of this study is presented in Figure 1.

To accurately classify the diverse and morphologically complex cell types present in peripheral blood smears, our study employed a consensus-based annotation process involving two experienced hematology experts. This approach was designed to combine the strengths and insights of both experts to provide a comprehensive and nuanced understanding of the characteristics of each cell type. By relying on consensus between the two experts to resolve classification discrepancies, we sought to mitigate the inherent problems and potential biases associated with preclassification of Sysmex DI-60 and individual expert judgment.

This study was designed to conduct deep learning analysis on cases already diagnosed with AML and ALL via flow cytometry at the time of diagnosis, thereby assessing leukemia lineage. Flow cytometry serves as the standard

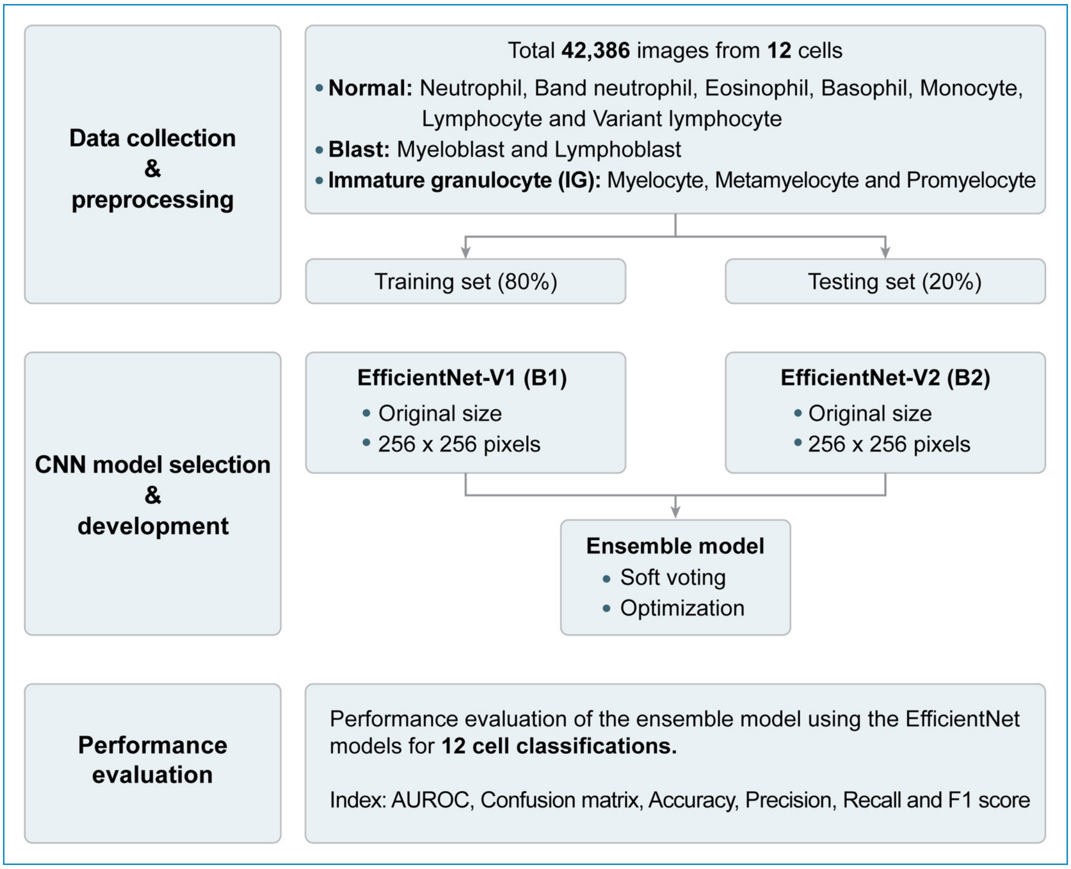
Data collection & preprocessing
Total 42,386 images from 12 cells
• Normal: Neutrophil, Band neutrophil, Eosinophil, Basophil, Monocyte, Lymphocyte and Variant lymphocyte
• Blast: Myeloblast and Lymphoblast
• Immature granulocyte (IG): Myelocyte, Metamyelocyte and Promyelocyte
Training set (80%)
Testing set (20%)
CNN model selection & development
EfficientNet-V1 (B1)
• Original size
• 256 x 256 pixels
EfficientNet-V2 (B2)
• Original size
• 256 x 256 pixels
Ensemble model
• Soft voting
• Optimization
Performance evaluation
Performance evaluation of the ensemble model using the EfficientNet models for 12 cell classifications.
Index: AUROC, Confusion matrix, Accuracy, Precision, Recall and F1 score

Figure 1. Diagram of the process of developing multiclass classification convolutional neural networks models and ensemble models.

for leukemia diagnosis, particularly in assessing the blast cells. In this study, we aimed to predict the shape of leukemia cells, a task challenging for humans to morphologically discern, utilizing artificial intelligence.

Image preprocessing

The original peripheral blood images were 352 × 355 pixels, with each image containing only one cell. For preprocessing, an image data generator resized the images to 256 × 256 pixels to facilitate more efficient processing. The dataset was divided into batches of 32 to prevent overloading the computer's memory during model training with the full set of 33,908 images. To enhance the model's ability to generalize and improve its robustness, we incorporated data augmentation techniques during the training phase.

Specifically, we applied rotations and horizontal, and vertical flips to the images, creating varied orientations and perspectives. These augmentations were used solely for training the model, ensuring that only the original, unaugmented image data were employed for testing and validation purposes. Additionally, the image data were shuffled using the shuffle option before being fed into the model. This shuffling was crucial for preventing the model from learning in the same sequence of data across epochs, thereby promoting a more diverse and effective learning process by varying the order in which images were presented, facilitating more dynamic adjustments to the model's weights.

CNN model development

We selected EfficientNet-V1 (B1) and EfficientNet-V2 (B2) from the CNN models. Using these models, we classified the following 12 cell types: seven normal WBCs, including neutrophils, band neutrophils, eosinophils, basophils, monocytes, lymphocytes, and variant lymphocytes, and five immature WBCs, including myeloblasts, lymphoblasts, myelocytes, metamyelocytes, and promyelocytes. The EfficientNet model was selected for its excellent performance on image classification tasks across a variety of benchmarks. The architectural design of the model efficiently scales the model size through composite coefficients to achieve a harmonious balance between the depth, width, and resolution of the network. These characteristics are particularly advantageous for processing high-resolution images, a hallmark of peripheral blood cell analysis, where retention of detail is critical for accurate classification.

When the model makes a prediction, the activation function is assigned to the softmax function in the last layer to extract a probability value for each class. The patience of the early stopping option was used to prevent overfitting during learning. We also used a model checkpoint to predict optimal weights as the optimal weights for early stopping may not be the same as the weights for the last training. An Adam optimizer was used, and the initial learning rate was set to 0.0005. We employed ReduceLROnPlateau to dynamically adjust the learning rate during training, recognizing the need for fine-tuning as we approach the optimal weight.

Ensemble model development

The ensemble model approach was adopted to leverage the complementary strengths of EfficientNet-V1 (B1) and EfficientNet-V2 (B2) to increase the robustness and generalizability of the classification results. Although the two models originate from the same architectural lineage, the different versions provide different perspectives on feature extraction and classification, which enhances the predictive capabilities of the ensemble model. There are two main types of ensemble techniques: hard voting and soft voting. In this study, we used a soft voting blending technique to develop our model. The ensemble approach with a soft voting mechanism is designed to mitigate potential biases and overfitting tendencies inherent in single-model predictions. We assigned a weight of 0.4 and 0.6 to EfficientNet-V1 (B1) and EfficientNet-V2 (B2), respectively.

Table 1. Number of images acquired for each cell type.

Cell	Training set ($n = 33,908$)	Testing set ($n = 8478$)
Neutrophil	14,210	3553
Band neutrophil	720	180
Eosinophil	287	72
Basophil	213	53
Monocyte	1860	465
Lymphocyte	9474	2369
Variant lymphocyte	746	187
Myeloblast	3041	760
Lymphoblast	1310	328
Myelocyte	1084	271
Metamyelocyte	615	153
Promyelocyte	348	87

Statistical analysis

The performance of our deep learning models, including the EfficientNet-V1 (B1), EfficientNet-V2 (B2), and the developed ensemble model, was evaluated using several key statistical metrics: accuracy, precision, recall (sensitivity), F1 score, and the area under the receiver operating characteristic curve (AUROC). These metrics were selected to provide a comprehensive assessment of model performance across the multiclass classification problem. Accuracy measures the proportion of true results (both true positives and true negatives) among the total number of cases examined. Precision (positive predictive value) assesses the proportion of true positives among all positive calls made by the model. Recall (sensitivity) measures the proportion of true positives correctly identified by the model, which is critical for medical diagnostic tests wherein missing a condition could have severe consequences. The F1 score provides a harmonic mean of precision and recall, offering a balance between the two in situations of uneven class distribution. AUROC reflects the model's ability to discriminate between the classes across all possible threshold values, with values closer to 1 indicating better performance. The statistical significance of the differences in performance metrics between models was assessed using the DeLong test for AUROC comparisons. For comparisons of accuracy, precision, recall, and F1 scores, we employed McNemar's test, acknowledging the correlated nature of the data. All statistical analyses were performed using Python Version 3.7, with libraries including SciPy for statistical functions.

Results

We acquired 42,386 images of 12 cell types and categorized 80% ($n = 33,908$) into the training set and 20% ($n = 8478$) into the testing set for model development (Table 1).

Table 2. Performance of the convolutional neural networks and ensemble models.

Model	Accuracy	Precision	Recall	F1 score
EfficientNet-V1(B1)[a]	0.8751	0.7127	0.7098	0.7105
EfficientNet-V1(B1)[b]	0.8747	0.7155	0.7196	0.7163
EfficientNet-V2 (B2)[a]	0.8732	0.7213	0.7111	0.7154
EfficientNet-V2 (B2)[b]	0.8779	0.7221	0.7225	0.7210
Ensemble model[c]	0.8858	0.7423	0.7331	0.7361

aImage size of 256 × 256 pixels is used.
bOriginal image size (355 × 355 pixels) is used.
cDeveloped by blending EfficientNet-V1(B1) and EfficientNet-V2 (B2) using the original image size.

Performance of CNN models

We used EfficientNet-V1 (B1) and EfficientNet-V2 (B2) to classify the 12 cell types. First, we developed a model by downsizing the original image to 256 × 256 pixels. Moreover, we developed a model using image data without downsizing, because the original image size was 355 × 355 pixels (Table 2). The performance of EfficientNet-V2 (B2) with the original image size exhibited the greatest accuracy (accuracy, 0.8779; precision, 0.7221; recall, 0.7225; and F1 score, 0.7210). The next-best accuracy was achieved by EfficientNet-V1 (B1), with a 256 × 256 pixels image. F1 score was the best for EfficientNet-V1 (B1) with the original image size.

Performance of ensemble models including cell-by-cell classification performance

The F1 score performances of EfficientNet-V1 (B1) and EfficientNet-V2 (B2) using the original image data were excellent. Therefore, an ensemble model was developed by blending these two models. Models with high F1 scores were used to develop the ensemble models because our data had a large imbalance in the number of cells. The ensemble model demonstrated an accuracy of 0.8858, precision of 0.7423, recall of 0.7331, and F1 score of 0.7361, which were the best among the models we developed (Table 2).

We investigated the classification performance of each of the 12 cell types in the ensemble model, AUROC (Figure 2). The AUROCs for normal cells are as follows: neutrophil (0.9949), band neutrophil (0.9616), eosinophil (0.9968), basophil (0.9845), monocyte (0.9959), lymphocyte (0.9921), and variant lymphocytes (0.9630). The AUROCs for pathologic cells are as follows: myeloblasts (0.9951), lymphoblasts (0.9935), myelocytes (0.9568), metamyelocytes (0.9593), and promyelocytes (0.9817). The results showed that the AUROC was > 0.9 for all cell types, including pathological cells. The F1 scores for the myeloblasts and lymphoblasts were 0.8873 and 0.8006, respectively.

A confusion matrix for the 12 cell types is shown in Figure 3. The developed ensemble model was successful in predicting 3398 of 3553 (95.6%) neutrophils and 2178 of 2369 (91.9%) lymphocytes. The model also predicted 697 of 760 (91.7%) myeloblasts, 265 of 328 (80.8%) lymphoblasts, 155 of 271 (57.2%) myelocytes, 57 of 153 (37.3%) metamyelocytes, and 56 of 87 (64.4%) promyelocytes.

The F1-score registered at 0.8873 for classifying AML myeloblasts and 0.8006 for classifying ALL lymphoblasts. This result is particularly significant considering that while myeloblasts can be seen in certain pathological cases with left-sided maturation, such as infections, they are not exclusive to AML. In contrast, lymphoblasts are solely indicative of ALL. For myelocytes, intermediate myelocytes, and promyelocytes, which are typically absent in standard observations, the F1 scores ranged between 0.4176 and 0.6747. Meanwhile, the AUROC remained relatively robust from 0.9586 to 0.9817 (Table 3).

Discussion

We aimed to classify cells, especially in blood samples from patients with leukemia, using AI, which has recently been actively applied in the medical field. We used 42,386 micrographs of peripheral blood cells to classify 12 cell types (normal and pathological cells). Medical AI research using image data typically focuses on radiological images.[7–11] It is relatively easy to acquire large datasets with radiologic images. Some research has delved into AI applications with large volumes of microscopic images,[10,16–18] although not to the extent observed in radiology. Therefore, we believe our study contributes to the research on AI using microscopic images. Our study was validated by two hematology specialists, employing more than 40,000 slide images. The performance of the developed AI model (AUROC) was satisfactory despite the data imbalance, in which the number of images of pathological cells such as myeloblasts, lymphoblasts, myelocytes, promyelocytes, and metamyelocytes was markedly smaller than the number of normal cell image data. The developed ensemble model showed an accuracy of 0.8858 and an F1 score of 0.7361 (Table 2). Moreover, it demonstrated an AUROC of > 0.99 and an F1 score of > 0.8 for myeloblasts and lymphoblasts, which are considered important in AL (Table 3). When developing a classification prediction model for imbalance, the F1 score and the harmonic mean of precision and recall are as important as the AUROC and accuracy.[19] However, with multiclass classification, as in our study, the AUROC value is important. Concerning improving the model's ability to predict blast cells (myeloblast, lymphoblast), it may be helpful to build a secondary model to distinguish between myeloblasts and lymphoblasts associated with AL. Nevertheless, the primary goal of this study was to develop a leukemia screening model capable of classifying all blood cells detected in peripheral blood. The data used in most medical AI studies are imbalanced; therefore, the AI model development method implemented in our study is valuable.

This study aimed to construct a model capable of not only detecting pathological cells indicative of diseases such as AML and ALL but also identifying all cell types commonly encountered in peripheral blood smear analysis. The inclusion of normal cells in the classification framework stemmed from several key considerations. First, our model's ability to distinguish between normal and immature cell types is intended to augment the diagnostic process by providing a secondary layer of validation to the existing preclassification system, Sysmex DI-60, which, albeit effective, may not capture all of the cellular variability. Second, by including normal cells, the model provides a comprehensive overview of peripheral blood

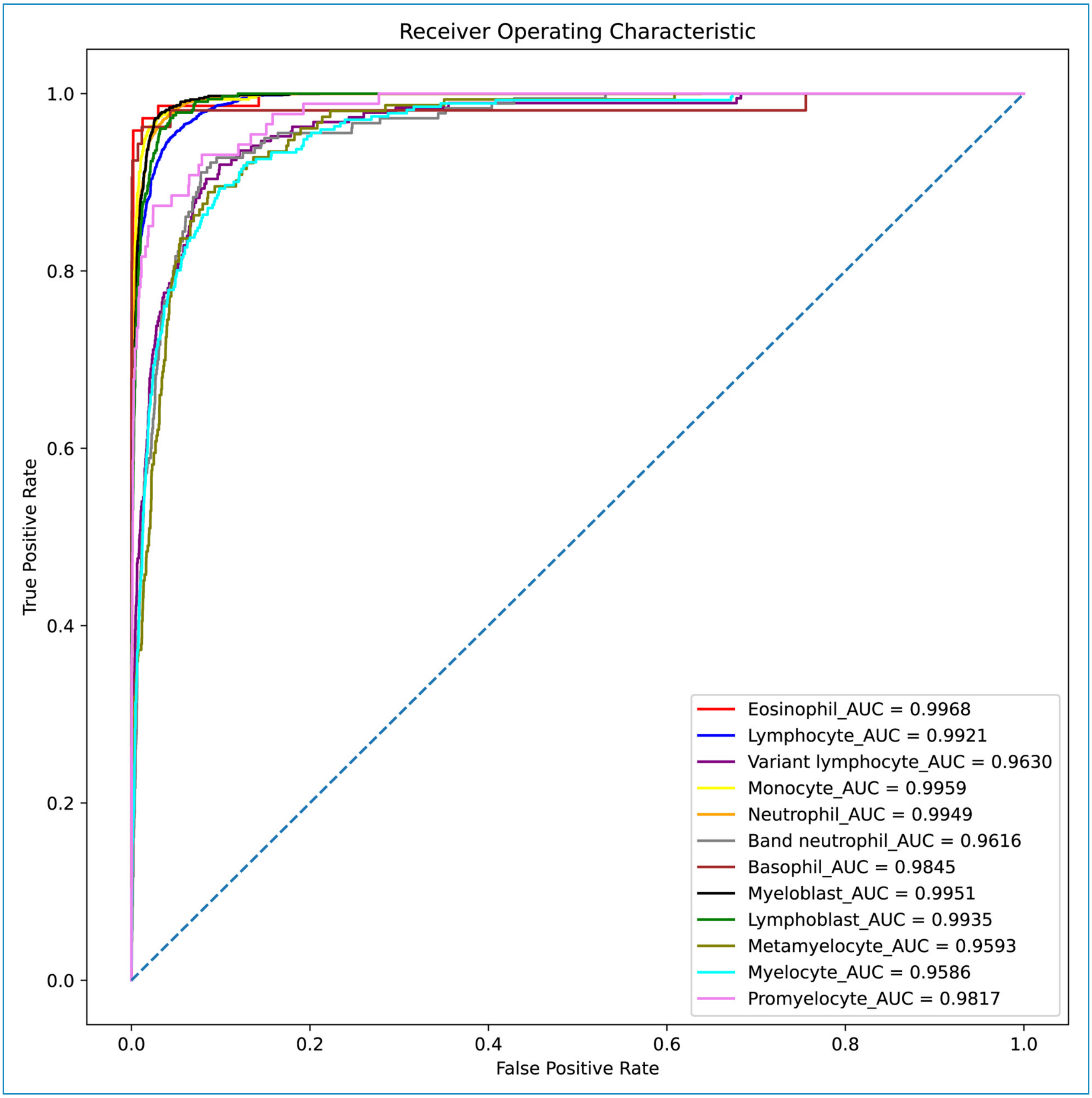

Figure 2. Area under the receiver operating characteristic of the ensemble model for classifying 12 cell types.

smears, reflecting the natural heterogeneity observed in clinical samples, thereby supporting a more nuanced interpretation of blood smears. Last, the ability to accurately classify normal and pathologic cell types increases the model's applicability beyond simple diagnostic support, as a tool for learning that promotes a deeper understanding of hematologic morphology for healthcare professionals. While we acknowledge that incorporating normal cells in the classification task adds complexity, the benefits in diagnostic precision and clinical applicability are significant. Experimental results demonstrate that this inclusive approach does not compromise the model's ability to identify pathologic cells; rather, it enriches the usefulness of the model by providing a more holistic tool for analyzing blood cells and thereby likely supporting clinical diagnosis.

We recognize the essential role of immunophenotyping in the final diagnosis of AML and ALL, which goes beyond morphological assessment to include determinations of cloning and differentiation, T-cell versus B-cell lineage in ALL, and other clinically relevant prognostic features. Our AI model is not designed to replace these essential diagnostic processes, but rather to complement them by providing a rapid and efficient preliminary screening tool for peripheral blood cell images. The tool was developed to improve the speed and

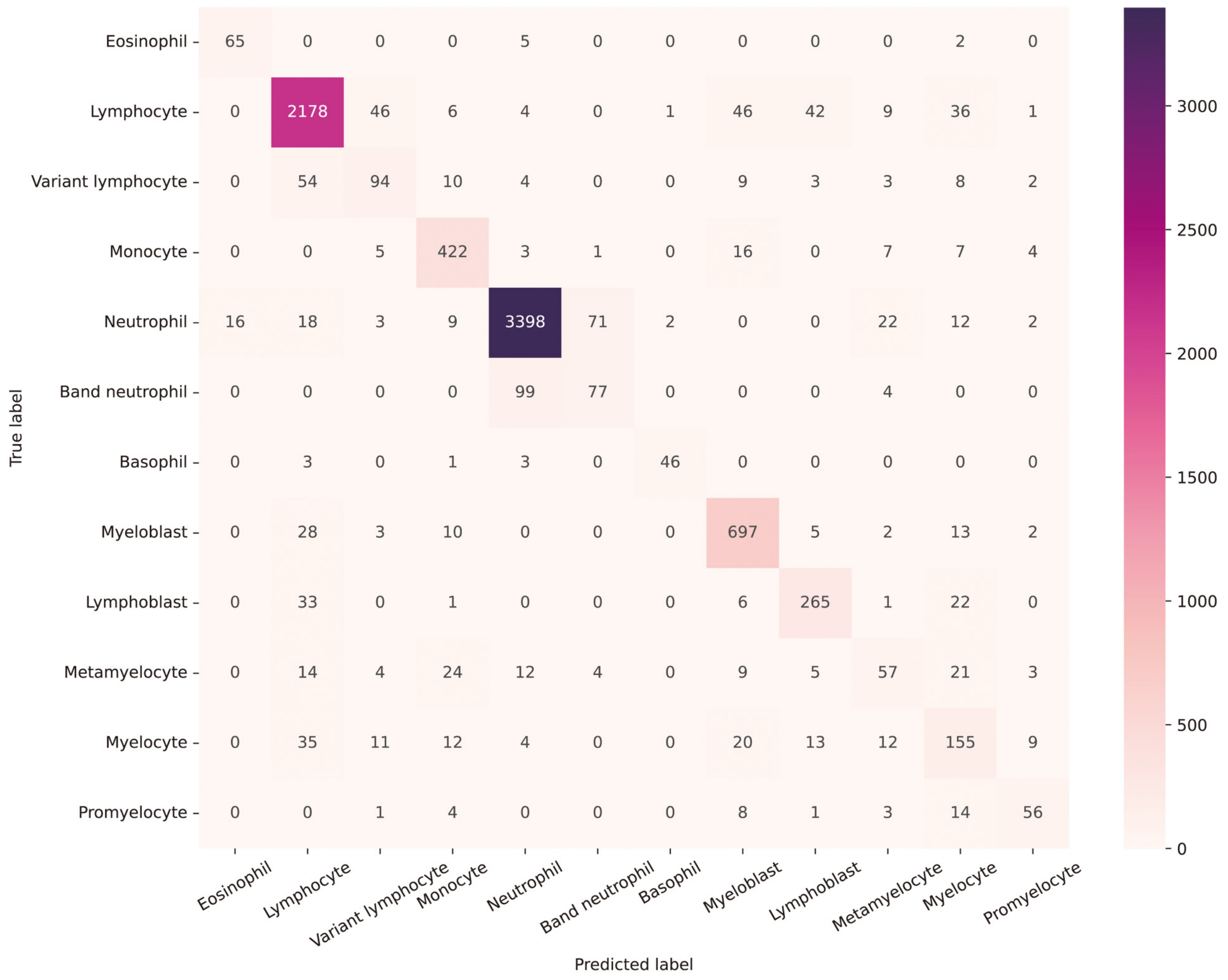

Figure 3. Confusion matrix of the ensemble model for classifying the 12 cell types.

efficiency of the initial diagnostic process by prioritizing cases for more detailed analysis. By identifying cells with a high degree of precision that are likely to be leukemic, the AI model acts as an additional screening layer to rapidly identify cases that require urgent attention. Moreover, it provides a standardized approach to cell image analysis, which can be particularly useful in settings where access to highly skilled morphologists is limited. Integrating AI technology into the leukemia diagnostic workflow therefore has the potential to improve the sensitivity of initial screening, support clinical decision-making, and rapidly guide patients to the most appropriate diagnostic and treatment pathway. These improvements in the diagnostic process highlight the complementary role of AI in modern hematology, and align with the goal of facilitating timely and accurate patient care.

In general, the application of deep learning AI in medical research is to differentiate data into two classes, such as survival and death, or normal and immature.[13,20] However, we used EfficientNet for multi-class classification. EfficientNet models have achieved state-of-the-art results in various computer vision tasks, including image classification, object detection, and segmentation.[19] They are typically pretrained on large datasets such as ImageNet and can be fine-tuned for specific tasks or used as feature extractors. Therefore, they have been used in many recent medical image AI studies.[21–23] Our model using EfficientNet showed a good classification performance for microscopic images, and we look forward to future research in this area.

Augmentation is a common technique used to develop AI models using image data.[24,25] When augmenting an original medical data image, the following methods should be used with caution: horizontal flipping, vertical flipping, and rotational range. This is because for chest radiographs, for example, changing the orientation and position of the image may affect the answer value. In our study, we used pictures of individual cells in the peripheral blood taken using Sysmex DI-60. Changes in cell position and orientation do not affect the shape of the cell, which is appropriate. Therefore, we used the horizontal, vertical, and rotation range options that did not affect the original image data.

Table 3. Classification performance of each of the 12 cell types in the ensemble model.

Cell	AUROC	Precision	Recall	F1 score
Neutrophil	0.9949	0.9621	0.9564	0.9592
Band neutrophil	0.9616	0.5033	0.4278	0.4625
Eosinophil	0.9968	0.8025	0.9028	0.8497
Basophil	0.9845	0.9388	0.8679	0.9020
Monocyte	0.9959	0.8457	0.9075	0.8755
Lymphocyte	0.9921	0.9217	0.9194	0.9205
Variant lymphocyte	0.9630	0.5629	0.5027	0.5311
Myeloblast	0.9951	0.8594	0.9171	0.8873
Lymphoblast	0.9935	0.7934	0.8079	0.8006
Myelocyte	0.9586	0.5345	0.5720	0.5526
Metamyelocyte	0.9593	0.4750	0.3725	0.4176
Promyelocyte	0.9817	0.7089	0.6437	0.6747

AUROC: area under the receiver operating characteristic.

AL diagnosis typically begins with an initial complete blood count screening and a morphological evaluation of peripheral blood. AL is suspected when severe pancytopenia or the presence of leukemic blasts coincides with leukocytosis. The detection rate and time for identifying leukemic cells in the peripheral blood can vary among laboratories owing to biological differences and observer expertise. Previous automated diagnostic machine learning models have predominantly focused on distinguishing between normal and immature WBCs in a binary system, yielding favorable performances.[17,26,27] The accuracy of discriminating ALL cells from normal cells ranges from 89.8% to 98%.[28] In our study, we developed a comprehensive cell-discrimination model encompassing 12 types of leukocytes in the peripheral blood of patients with AL. The model exhibited excellent performance, with an AUROC of 0.9951 and 0.9935 for distinguishing myeloblasts and lymphoblasts from other cells, respectively, and demonstrated an excellent ability to identify blast cells among various cell types in the peripheral blood.

AL encompasses AML and ALL, which present challenges in differentiation through peripheral blood analysis. Immunophenotyping using flow cytometry is the gold standard for determining AL lineage. However, access to flow cytometry equipment and expertise may be limited in certain laboratories. Laboratory detection of blasts is crucial for identifying AL. AML exhibits a heterogeneous appearance with diverse differentiation among bone marrow-derived myeloid cells, which varies by subtype. Our model demonstrated excellent performance with 91.7% sensitivity (recall) in classifying myeloblasts by simultaneously learning diverse shapes found in peripheral blood samples from patients with AML, except for acute promyelocytic leukemia (APL). In ALL, visually distinguishing normal lymphocytes, reactive lymphocytes, lymphoma cells, and hematogones is challenging. The model successfully classified a significant number of lymphoblasts (42 of 328), achieving a sensitivity of 80.8%. This sensitivity is comparable to that reported by Boldu et al., with 82% sensitivity for single-cell classification of B lymphoblasts.[17]

Peripheral blood analysis reveals important information about the myeloid lineage, particularly when observing IGs such as promyelocytes, myelocytes, and metamyelocytes, which indicate a left shift. This left shift is commonly observed in reactive neutrophilia and chronic myelogenous leukemia (CML).[29] CML often exhibits a myeloid left shift with an increase in the number of basophils or eosinophils. Reactive neutrophilia can occur in response to inflammatory stimuli or granulocytic cytokines such as granulocyte colony-stimulating factor (G-CSF).

Determination of the myeloid left shift relies on nuclear morphology, distinguishing it from IGs associated with bacterial infections. Myelocytes possess round nuclei, metamyelocytes display kidney bean-shaped nuclei, and promyelocytes are larger and contain azurophilic granules. Notably, myelocytes and metamyelocytes exhibited lower sensitivity and precision than other cell types; however their AUROC of 0.9586 and 0.9593, respectively, allowed differentiation from other cell types. We acknowledge the limitation that this study was conducted at a single institution. AI research using peripheral blood slides is difficult to perform in large-scale, multicenter studies because it requires simultaneous labeling by hematology experts. Nevertheless, based on our study, it is possible to develop an AI model for leukemia diagnosis with better performance if data collection from multiple centers could be performed in the future.

In our endeavor to harness AI for the advancement of AL diagnosis through peripheral blood smear analysis, our study focused on classifying 12 specific cell types associated with AML, ALL, and the presence of immature granulocytes. This approach has demonstrated considerable promise in enhancing the accuracy of cellular classification, contributing valuable insights into the application of AI within hematology diagnostics. However, we recognize key limitations that frame the scope of our current work and outline essential directions for future research to extend the clinical utility and precision of our model. One principal limitation lies in our study's emphasis on cellular-level analysis without evaluating its implications in a broader, patient-specific diagnostic context. Moreover, our methodology concentrated on the model's proficiency in classifying individual cell images without synthesizing these classifications to infer overarching patient-level diagnoses. This gap underscores an opportunity for subsequent research to bridge cell-level classification with patient-level diagnostic accuracy, thereby elucidating the full potential of AI models in aiding the comprehensive diagnosis of hematological malignancies. Furthermore, while our model proficiently identifies the cell types at the core of our study, the realm of peripheral blood testing in clinical practice encompasses a broader spectrum of WBCs reflecting a variety of diseases and malignancies, each distinguished by unique morphological features. Notably, cell types representing specific diseases, such as hairy cell leukemia, Burkitt's lymphoma, and chronic lymphocytic leukemia, as well as other less common but clinically relevant cell types, such as plasma cells, were not included in the analysis. The omission of these cell types indicates a limitation in the model's ability to represent all of the different WBC variants encountered in the clinical setting. Enhancing the dataset to include these additional cell types in the future will be essential to broaden the diagnostic applicability of the AI tool. Additionally, the framework of the current model does not specifically accommodate the diagnosis of complex diseases such as dual phenotype or mixed phenotype acute leukemia (MPAL). These rare diseases, characterized by the simultaneous expression of myeloid and lymphoid markers, require accurate identification because they have significant implications for patient management and treatment. Accurate classification of MPAL, which depends on nuanced morphological and molecular analysis, is essential for determining the appropriate treatment approach according to AML or ALL protocols. To address these limitations, datasets need to be enriched with a variety of MPAL cases to empower models to accurately recognize and classify the features of these challenging cases. In addition to the aforementioned limitations, another important aspect that needs to be discussed relates to the specificity of the cell types used in our study, specifically the differentiation and expression of promyelocytes. Our dataset included promyelocytes commonly found in APL and promyelocytes seen in other clinical situations, such as severe infections or after G-CSF treatment. Although we trained our model to recognize and classify these cell types, we did not provide a detailed analysis of the number of each specific type of promyelocyte in our training set. This omission limits the granularity of our analysis and may affect the effectiveness of our model in clinical scenarios where it is important to accurately distinguish APL promyelocytes from other variants. Misclassification of these cell types can lead to serious diagnostic errors, especially in the case of APL, which requires specific therapeutic interventions to distinguish it from other diseases characterized by promyelocytes. Future work will aim to unambiguously quantify and distinguish promyelocytes in different pathological contexts within the training dataset to improve the diagnostic precision and clinical utility of the model.

Conclusions

To be clinically deployed, an AI model for diagnosing AL must meet several key performance criteria, including high accuracy, precision, reproducibility, and specificity. Moreover, the model must also demonstrate robustness in a diverse patient population, maintain performance amid challenges in sample preparation and imaging techniques, and integrate seamlessly into clinical workflows without placing a significant burden on healthcare professionals. While the specific thresholds for laying the foundation for these benchmarks vary depending on regulatory standards and intended use, models targeted for clinical deployment should aim to minimize the risk of false positives and false negatives with an accuracy of at least 95%, and relatively high precision and recall. These performance metrics ensure that the model can reliably identify AL cases, enabling timely and accurate diagnostic and treatment decisions. The AI model developed has shown promising results in classifying both normal and pathological cells, most notably in identifying myeloblasts and lymphoblasts, which are important in the diagnosis of AL. However, despite the demonstrated potential of the model, concerns about accuracy and false positives in classifying IGs such as intermediate myeloid cells, myeloblasts, and promyelocytes have been noted as areas for improvement. Considering these aspects, we believe that our model represents a step forward in the use of AI to improve peripheral blood smear slide analysis. The model will serve as a foundation for future work to improve accuracy and reliability, particularly in classifying more challenging IGs. Optimism about the current capabilities of the model and its potential contribution to hematology must be balanced with a rigorous evaluation of its performance for all cell types under consideration. Future efforts will therefore focus on addressing the limitations identified with the goal of achieving a truly comprehensive tool that meets the high standards required for clinical application.

Authors' Notes: Seung Min Baik and Dong Jin Park contributed equally as the first authors.

Contributorship: SMB and DJP conceived and designed the study. SMB and DJP developed the methodology. SP, YHP, and DJP acquired the data. All authors analyzed and interpreted data. SP, YHP, SMB, and DJP wrote the first draft of the manuscript. All authors reviewed and edited the manuscript and approved the final version of the manuscript.

Data availability: The data used in this work will be available upon request.

Declaration of conflicting interests: The authors declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Ethical approval: This study was approved by the Institutional Review Board of the Ewha Womans University Mokdong Hospital (Approval No: EUMC 2021-10-046). The patient records were reviewed and published in accordance with the Declaration of Helsinki. The requirement for informed consent was waived owing to the retrospective nature of the study.

Funding: The authors disclosed receipt of the following financial support for the research, authorship, and/or publication of this article: This work was supported by the National Research Foundation of Korea (Grant No. 2021R1I1A1A01061072).

Guarantor: S.M.B. and D.J.P.

ORCID iDs: Seung Min Baik https://orcid.org/0000-0003-1051-6775
Dong Jin Park https://orcid.org/0000-0002-2412-5292

References

1. Bray F, Ferlay J, Soerjomataram I, et al. Global cancer statistics 2018: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. *CA Cancer J Clin* 2018; 68: 394–424. DOI: 10.3322/caac.21492.
2. National Cancer Institute, The Surveillance, Epidemiology, and End Results, Cancer Stat Facts: Leukemia, https://seer.cancer.gov/statfacts/html/leuks.html (2020, accessed June 12 2023).
3. Tebbi CK. Etiology of Acute Leukemia: A Review. *Cancers (Basel)* 2021; 13. DOI: 10.3390/cancers13092256.
4. Estey E. Why Is Progress in Acute Myeloid Leukemia So Slow? *Semin Hematol* 2015; 52: 243–248. DOI: 10.1053/j.seminhematol.2015.03.007.
5. Weir EG and Borowitz MJ. Flow cytometry in the diagnosis of acute leukemia. *Semin Hematol* 2001; 38: 124–138. DOI: 10.1016/s0037-1963(01)90046-0.
6. Del Principe MI, De Bellis E, Gurnari C, et al. Applications and efficiency of flow cytometry for leukemia diagnostics. *Expert Rev Mol Diagn* 2019; 19: 1089–1097. DOI: 10.1080/14737159.2019.1691918.
7. Laino ME, Ammirabile A, Posa A, et al. The Applications of Artificial Intelligence in Chest Imaging of COVID-19 Patients: A Literature Review. *Diagnostics (Basel)* 2021; 11. DOI: 10.3390/diagnostics11081317.
8. Zhang W, Li R, Deng H, et al. Deep convolutional neural networks for multi-modality isointense infant brain image segmentation. *Neuroimage* 2015; 108: 214–224. DOI: 10.1016/j.neuroimage.2014.12.061.
9. Zhang F, Breger A, Cho KIK, et al. Deep learning based segmentation of brain tissue from diffusion MRI. *Neuroimage* 2021; 233: 117934. DOI: 10.1016/j.neuroimage.2021.117934.
10. Fan BE, Yong BSJ, Li R, et al. From microscope to micropixels: A rapid review of artificial intelligence for the peripheral blood film. *Blood Rev* 2024; 64: 101144. DOI: 10.1016/j.blre.2023.101144.
11. Hosny A, Parmar C, Quackenbush J, et al. Artificial intelligence in radiology. *Nat Rev Cancer* 2018; 18: 500–510. DOI: 10.1038/s41568-018-0016-5.
12. Gore JC. Artificial intelligence in medical imaging. *Magn Reson Imaging* 2020; 68: A1–a4. DOI: 10.1016/j.mri.2019.12.006.
13. Baik SM, Hong KS and Park DJ. Deep learning approach for early prediction of COVID-19 mortality using chest X-ray and electronic health records. *BMC Bioinformatics* 2023; 24: 190. DOI: 10.1186/s12859-023-05321-0.
14. Zammit J, Fung DLX, Liu Q, et al. Semi-supervised COVID-19 CT image segmentation using deep generative models. *BMC Bioinformatics* 2022; 23: 343. DOI: 10.1186/s12859-022-04878-6.
15. Dong H, Zhu B, Zhang X and Kong X. Use data augmentation for a deep learning classification model with chest X-ray clinical imaging featuring coal workers' pneumoconiosis. *BMC Pulm Med* 2022; 22: 271. DOI: 10.1186/s12890-022-02068-x.
16. Atteia G, Alhussan AA and Samee NA. BO-ALLCNN: Bayesian-Based Optimized CNN for Acute Lymphoblastic Leukemia Detection in Microscopic Blood Smear Images. *Sensors (Basel)* 2022; 22. DOI: 10.3390/s22155520.
17. Boldú L, Merino A, Acevedo A, et al. A deep learning model (ALNet) for the diagnosis of acute leukaemia lineage using peripheral blood cell images. *Comput Methods Programs Biomed* 2021; 202: 105999. DOI: 10.1016/j.cmpb.2021.105999.
18. Saleem S, Amin J, Sharif M, et al. A deep network designed for segmentation and classification of leukemia using fusion of the transfer learning models. *Complex & Intelligent Systems* 2022; 8: 3105–3120. DOI: 10.1007/s40747-021-00473-z.
19. Goutte C and Gaussier E. A Probabilistic Interpretation of Precision, Recall and F-Score, with Implication for Evaluation. In: Berlin, Heidelberg, Springer Berlin Heidelberg, Santiago de Compostela, Spain, March 21-23, 2005, pp.345-359.
20. Kurt Z, Işık Ş, Kaya Z, et al. Evaluation of EfficientNet models for COVID-19 detection using lung parenchyma. *Neural Comput Appl* 2023; 35: 12121–12132. DOI: 10.1007/s00521-023-08344-z.
21. Booysens A and Viriri S. Exploration of Ear Biometrics Using EfficientNet. *Comput Intell Neurosci* 2022; 2022: 3514807. DOI: 10.1155/2022/3514807.
22. Huang C, Wang W, Zhang X, et al. Tuberculosis Diagnosis using Deep Transferred EfficientNet. *IEEE/ACM Trans Comput Biol Bioinform* 2022; Pp. DOI: 10.1109/tcbb.2022.3199572.
23. Rahhal MMA, Bazi Y, Jomaa RM, et al. Contrasting EfficientNet, ViT, and gMLP for COVID-19 Detection in Ultrasound Imagery. *J Pers Med* 2022; 12: 1707. DOI: 10.3390/jpm12101707.
24. Chlap P, Min H, Vandenberg N, et al. A review of medical image data augmentation techniques for deep learning applications. *J Med Imaging Radiat Oncol* 2021; 65: 545–563. DOI: 10.1111/1754-9485.13261.

25. Chen Y, Yang XH, Wei Z, et al. Generative Adversarial Networks in Medical Image augmentation: A review. *Comput Biol Med* 2022; 144: 105382. DOI: 10.1016/j.compbiomed.2022.105382.
26. Hegde RB, Prasad K, Hebbar H, et al. Automated Decision Support System for Detection of Leukemia from Peripheral Blood Smear Images. *J Digit Imaging* 2020; 33: 361–374. DOI: 10.1007/s10278-019-00288-y.
27. Elhassan TA, Mohd Rahim MS, Siti Zaiton MH, et al. Classification of Atypical White Blood Cells in Acute Myeloid Leukemia Using a Two-Stage Hybrid Model Based on Deep Convolutional Autoencoder and Deep Convolutional Neural Network. *Diagnostics (Basel)* 2023; 13: 196. DOI: 10.3390/diagnostics13020196.
28. Rehman A, Abbas N, Saba T, et al. Classification of acute lymphoblastic leukemia using deep learning. *Microsc Res Tech* 2018; 81: 1310–1317. DOI: 10.1002/jemt.23139.
29. Arber DA, Orazi A, Hasserjian R, et al. The 2016 revision to the World Health Organization classification of myeloid neoplasms and acute leukemia. *Blood* 2016; 127: 2391–2405. DOI: 10.1182/blood-2016-03-643544.