

AI-Based Leukemia Screening and Blood Smear Analysis System

B G Varsha¹, Abhinaya², M M Aishwarya³, B Sushmita⁴, Vikram sandeep⁵

Department of Computer Science and Engineering — Artificial Intelligence, Ballari Institute of Technology and Management, Ballari, Karnataka, India

Email: varsha97415@gmail.com¹, abhinayagouda28@gmail.com², aishwarya.matturmth06@gmail.com³, sushmitab5929@gmail.com⁴, vikramsandeep@bitm.edu.in⁵

Abstract:

Examining a peripheral blood smear under the microscope is usually the first laboratory step when leukemia is suspected, but the examination is slow and depends on the availability of an experienced haematologist. This paper describes a web-based system that supports preliminary screening by classifying microscopic blood smear images into five categories: acute lymphoblastic leukemia (ALL), acute myeloid leukemia (AML), chronic lymphocytic leukemia (CLL), chronic myeloid leukemia (CML) and healthy. A small gatekeeper network first rejects uploads that are not blood smear images. Accepted images are classified by a MobileNetV2 network pre-trained on ImageNet and used as a frozen feature extractor, with a trained classification head and a five-way softmax output. A FastAPI back end stores every case in MongoDB so that a doctor can review the result before a report is produced. On a held-out test set of 5,000 images (1,000 per class) the classifier reached an overall accuracy of 75.9% and a macro-averaged F1-score of 75.6%, against 97.1% on the training images, which points to overfitting. Healthy smears and CML were recognised most reliably (recall of 99.2% and 94.3%), none of the 4,000 leukemia images was labelled healthy, and most errors occurred among ALL, AML and CLL. The system is meant to assist clinical review and not to replace it.

Keywords — Leukemia screening, blood smear analysis, MobileNetV2, transfer learning, convolutional neural network, medical image classification, FastAPI

I. INTRODUCTION

Leukemia is a group of cancers in which abnormal white blood cells multiply in the bone marrow and spill into the bloodstream. The types differ in their treatment and in how quickly treatment has to begin, so a screening step that can point towards a likely type is useful even before confirmatory tests are available.

In most laboratories the first look at a patient's blood is a stained peripheral smear read under a microscope. Reading a smear well takes years of practice, the result varies from one observer to another, and the work piles up when many samples arrive together. Smaller hospitals and rural laboratories often have no specialist on site at all. Confirmation still needs bone marrow examination, flow cytometry or cytogenetic tests, so the smear is best treated as a triage step: it should be quick, and it should not miss a patient who needs those further tests.

Convolutional neural networks have become a standard tool for image-based diagnostic support [1], and transfer learning from large natural-image collections [2], [3] makes it practical to train them on the few thousand images that a typical medical project can gather. Several groups have applied pre-trained networks to leukemia images [4], [5], [6], [7], but many of these studies stop at a trained model. A model alone does not help a laboratory. Images arrive in different formats and quality, some uploads are not blood smears at all, and the person who finally takes responsibility for the diagnosis is a doctor who needs to see the case, the image and the confidence of the model in one place.

This paper presents a screening-support system built around these practical points. A gatekeeper network rejects uploads that are not blood smears, a five-class MobileNetV2 classifier labels the valid images as ALL, AML, CLL, CML or healthy, and a FastAPI and MongoDB back end follows each case from upload to doctor review. The classifier is evaluated on a balanced test set of 5,000 images, and its errors are analysed class by class instead of being summed up in a single accuracy figure.

II. THEORETICAL BACKGROUND

Leukemia is classified by the cell lineage that is affected, lymphoid or myeloid, and by the pace of the disease, acute or chronic. This gives acute lymphoblastic leukemia (ALL), acute myeloid leukemia (AML), chronic lymphocytic leukemia (CLL) and chronic myeloid leukemia (CML). On a stained smear, the haematologist looks at the size, shape and nucleus of the white cells and at how many immature cells are present. Many of these cues look alike across the acute types, which is why a smear alone does not settle the diagnosis.

A convolutional neural network learns image features in layers, from edges and colours in the early layers to more complex shapes in the later ones. Training such a network from scratch needs far more labelled images than most medical projects can collect. Transfer learning avoids this by starting from weights learned on a large natural-image collection such as ImageNet [3] and adapting only part of the network to the new task [2].

MobileNetV2 [8] is a compact network made of depthwise separable convolutions arranged in inverted residual blocks

with linear bottlenecks. These choices reduce the number of parameters and the amount of computation, so the network stays small. Deeper residual networks such as ResNet50 [9] use skip connections to make very deep stacks of layers trainable and usually give richer features, but at a higher computational cost. This work uses MobileNetV2.

III. SYSTEM ARCHITECTURE

The system is organised into four layers: the presentation layer, the application layer, the model layer and the data layer. Keeping the layers separate makes it easier to replace a model or the database without rewriting the rest of the system.

A. Presentation Layer

This layer is the web interface, written in HTML5, CSS3 and JavaScript with Tailwind CSS. The laboratory technician logs in, enters the patient details and uploads a smear image either from a file (JPEG, PNG or WebP, up to 10 MB) or by capturing it with a webcam. The interface also shows the dashboard, the case history and the review panel in which the doctor sees the result.

B. Application Layer

A FastAPI service provides the endpoints for upload, prediction and case retrieval. It checks the file type and size of every upload, passes the image to the gatekeeper network and, if the image is accepted, to the classifier.

C. Model Layer

This layer holds the two networks built with TensorFlow and Keras [10]: the binary gatekeeper network and the five-class MobileNetV2 classifier. Both are described in Section VI.

D. Data Layer

MongoDB stores the patient details, the uploaded image, the predicted class with its probabilities and the status of each case. A new case waits in an “awaiting review” state, and an upload that fails the gatekeeper is stored as “rejected” without a class label.

IV. LITERATURE REVIEW

The studies reviewed here come from research on deep learning for medical images and on leukemia detection from blood smear images. Early methods relied on hand-crafted features and cell segmentation, while recent work learns

features directly from the image [1]. TABLE I summarises the studies that are most relevant to this work.

TABLE I
LITERATURE REVIEW SUMMARY

No	Study	Method	Key findings	Limitations
01	Pan and Yang [2], 2010. <i>A Survey on Transfer Learning</i>	Transfer learning (survey)	Explains when knowledge from a source task can help a target task	Gain depends on how close the two domains are
02	He et al. [9], 2016. <i>Deep Residual Learning for Image Recognition</i>	Residual networks (ResNet)	Skip connections make very deep networks trainable	Heavier computation than light networks
03	Litjens et al. [1], 2017. <i>A Survey on Deep Learning in Medical Image Analysis</i>	Survey of CNN methods	Summarises deep learning across medical imaging tasks	General survey, not specific to blood smears
04	Sandler et al. [8], 2018. <i>MobileNetV2: Inverted Residuals and Linear Bottlenecks</i>	Depthwise separable convolution	Compact network with low computation	Smaller capacity than deeper networks
05	Shafique and Tehsin [4], 2018. <i>ALL Detection and Classification of Its Subtypes Using Pretrained Deep CNNs</i>	Fine-tuned AlexNet	Detects ALL and separates its subtypes	Limited to ALL
06	Rehman et al. [5], 2018. <i>Classification of Acute Lymphoblastic Leukemia Using Deep Learning</i>	Deep CNN	Classifies ALL from microscopic images	Limited to ALL
07	Kassani et al. [7], 2019. <i>A Hybrid Deep Learning Architecture for Leukemic B-lymphoblast Classification</i>	VGG16 + MobileNet features	Separates leukemic B-lymphoblasts from normal cells	Two-class task
08	Loey et al. [6], 2020. <i>Deep Transfer Learning in Diagnosing Leukemia in Blood Cells</i>	Transfer learning with AlexNet	Diagnoses leukemia in blood-cell images	Older, heavier AlexNet backbone

In the studies we reviewed, the task is most often the binary question of whether leukemic cells are present, or the detection of a single disease such as ALL. Fewer studies consider the four main types together with healthy samples, and fewer still describe how a model is placed inside a working application with input validation, record keeping and clinician review. The present work addresses that gap.

V. WORKFLOW OF THE SYSTEM

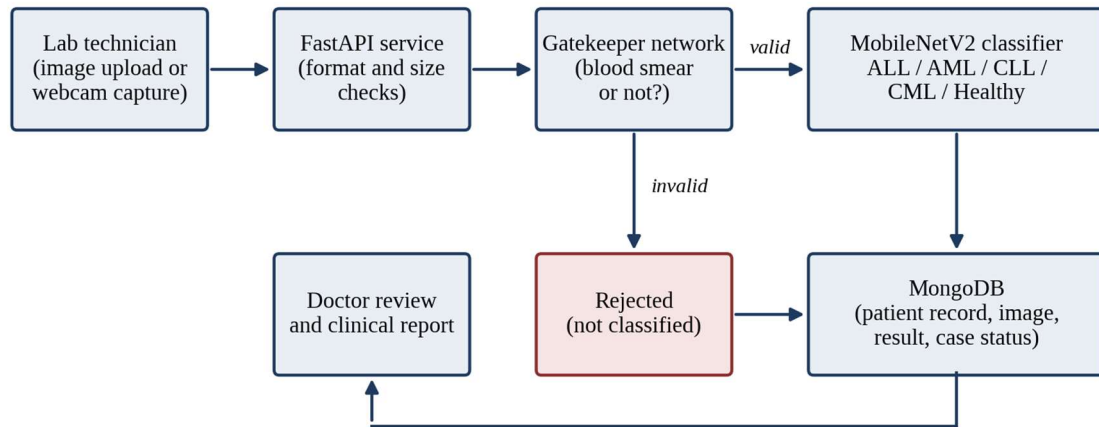


Fig. 1. Workflow diagram

A laboratory technician registers a patient case and uploads a smear image. Every upload first passes basic checks on file type and size in the FastAPI service. The gatekeeper network then decides whether the image looks like a blood smear. Images that fail are marked as rejected and are not classified, which prevents an unrelated photograph from receiving a leukemia label (Fig. 1).

Valid images go to the MobileNetV2 classifier, and the prediction, the class probabilities and the image are stored together with the patient details. The doctor opens the stored case, looks at the image and the probability given to each class, and prepares the clinical report. The final decision always stays with the doctor.

VI. PROPOSED METHODOLOGY

Step 1: Collect labelled blood smear images of ALL, AML, CLL, CML and healthy samples

Step 2: Resize all images to 128×128 pixels and scale the pixel values

Step 3: Augment the training images with flips, a small rotation and zoom

Step 4: Train the binary gatekeeper network on smear and non-smear images

Step 5: Train the five-class MobileNetV2 classifier on the training folder

Step 6: Evaluate the classifier on the 5,000-image test set

Step 7: Deploy both networks behind the FastAPI service

Step 8: Check every upload with the gatekeeper and reject images that are not blood smears

Step 9: Classify the accepted images and store the result with the patient record

Step 10: Send the case to the doctor for review and the clinical report

A. Dataset and Preprocessing

The classifier was trained and tested on a publicly available collection of leukemia blood smear images hosted on Kaggle [11]. It has five classes: ALL, AML, CLL, CML and healthy.

The collection holds about 20,000 images, of which about 15,000 form the training folder and 5,000 the test folder, a split of roughly 75% and 25%. The test folder contains 1,000 images per class and is kept apart from the training folder, and no image was copied from one to the other.

No separate validation set was created. The 5,000-image test set served as the validation set during training, so its score was observed after every epoch. No early stopping or best-checkpoint selection was used, and no hyperparameters were tuned on it; the final-epoch model was retained. All images were resized to 128×128 pixels. For the classifier, pixel values were divided by 255, and the training images were augmented with horizontal and vertical flips, a rotation factor of 0.05 and a zoom factor of 0.1. The test images were not augmented.

B. Gatekeeper Network

A classifier trained on five blood-related classes will still return a confident answer when it is shown a photograph of a face or a screenshot, because a softmax output always sums to one. To prevent this, a binary gatekeeper network separates blood smear images from everything else, and only images that pass are sent forward.

The gatekeeper uses a frozen MobileNetV2 base with ImageNet weights, followed by global average pooling [12], a dense layer of 64 ReLU units, a dropout layer with a rate of 0.4 [13] and a single sigmoid unit. It has 2,340,033 parameters, of which 82,049 are trainable. It was trained on 6,000 images: 3,000 blood smear images, 600 from each of the five classes, and 3,000 images from CIFAR-10 [14], which served as the “not a blood smear” class and show everyday objects such as aeroplanes, cats and trucks. The images were split 80:20 into 4,800 for training and 1,200 for validation. They were resized to 128×128 pixels and scaled to the range $[-1, 1]$ with the MobileNetV2 preprocessing function, and the training images were augmented with horizontal flips and a rotation factor of 0.05. Training ran for 5 epochs with the Adam optimizer [15], a learning rate of 0.001, a batch size of 64 and binary cross-entropy loss. On the validation split the gatekeeper reached an accuracy of 87%.

C. MobileNetV2 Classifier

The classifier is built on MobileNetV2 with weights pre-trained on ImageNet. The backbone was frozen, so only the newly added classification head was trained, and the backbone was not fine-tuned. The head ends in a fully connected layer with a

softmax function that returns one probability for each of the five classes (Fig. 2). The network was trained for 10 epochs with the Adam optimizer, a learning rate of 0.001, a batch size of 128 and categorical cross-entropy loss.

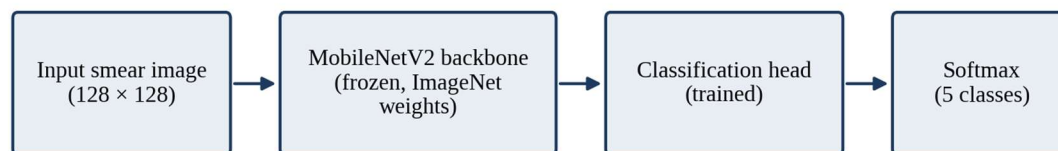


Fig. 2. Architecture of the classification model

VII. RESULTS AND DISCUSSION

A. Evaluation Metrics

The classifier was evaluated on the held-out test set of 5,000 images described in Section VI-A. Accuracy, and precision, recall and F1-score for each class, were computed from the confusion matrix. For a given class, TP, FP and FN denote the numbers of true positives, false positives and false negatives:

$$\text{Precision} = TP / (TP + FP) \quad (1)$$

$$\text{Recall} = TP / (TP + FN) \quad (2)$$

$$F1 = 2 \times \text{Precision} \times \text{Recall} / (\text{Precision} + \text{Recall}) \quad (3)$$

B. Classification Performance

Fig. 3 shows the confusion matrix and TABLE II lists the per-class scores. The classifier assigned the correct label to 3,793 of the 5,000 test images, an overall accuracy of 75.9%. The macro-averaged precision, recall and F1-score were 76.3%, 75.9% and 75.6%. On the training images the same model reached 97.1%, a gap of about 21 percentage points, which shows that the classification head overfits the training data.

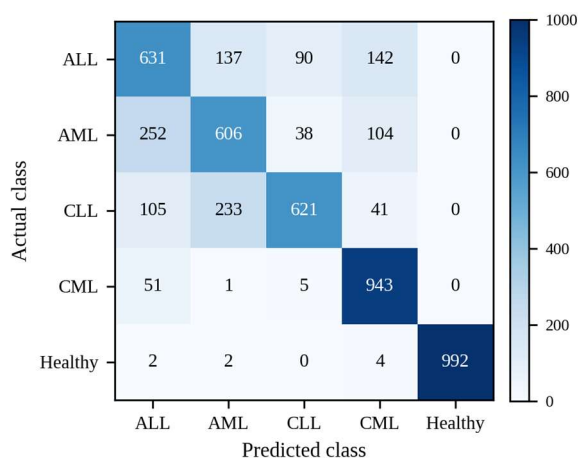


Fig. 3. Confusion matrix on the 5,000-image test set (rows: actual class, columns: predicted class)

TABLE II
PER-CLASS RESULTS ON THE TEST SET

Class	Precision (%)	Recall (%)	F1-score (%)	Images
ALL	60.6	63.1	61.8	1,000
AML	61.9	60.6	61.2	1,000
CLL	82.4	62.1	70.8	1,000
CML	76.4	94.3	84.4	1,000
Healthy	100.0	99.2	99.6	1,000
Macro average	76.3	75.9	75.6	5,000

C. Error Analysis

Healthy smears were almost never confused with disease: 992 of 1,000 were classified correctly, and the eight errors were spread over ALL, AML and CML. The reverse mistake is the more serious one in screening, and it did not occur. None of the 4,000 leukemia images was labelled healthy. Treated as a two-way question of leukemia against healthy, the classifier therefore missed no leukemia case and raised eight false alarms among 1,000 healthy images on this test set.

CML had the highest recall among the leukemia classes (94.3%), but its precision was lower (76.4%) because 142 ALL, 104 AML and 41 CLL images were also predicted as CML. The main weakness lies in telling ALL, AML and CLL apart. Of the 1,000 AML images, 252 were labelled ALL; of the 1,000 CLL images, 233 were labelled AML; and 137 ALL images were labelled AML. If only the four leukemia classes are considered, the accuracy is 70.0%.

Some of this confusion is plausible. Blast cells in acute leukemias can look alike on a smear, and the features that separate lymphoid from myeloid cells are subtle. Differences in staining and image capture in a public dataset add further variation, and a frozen backbone that was trained on everyday photographs may not capture fine cell detail well. We did not review the misclassified images cell by cell, so these explanations remain hypotheses. The overall accuracy is moderate, and the result is best read as showing that the pipeline works end to end and separates healthy from leukemic smears well, not as a replacement for confirmatory testing.

VIII. LIMITATIONS AND FUTURE SCOPE

Several limitations should be stated plainly. The classifier overfits: its accuracy is 97.1% on the training images but 75.9% on the test images. The test images also served as the validation set during training. No early stopping, checkpoint selection or hyperparameter tuning was based on them, but there was no separate validation set, and the test score was visible after every epoch. The classifier was developed and tested on one public dataset, so its behaviour on smears from other laboratories, stains or cameras is unknown. The confusion among ALL, AML and CLL holds the five-class accuracy at 75.9%, which is not adequate for any use beyond screening support, and the frozen backbone was not fine-tuned. The gatekeeper was trained with CIFAR-10 images as its only negative class, so it has not been validated on phone photographs, documents or other kinds of microscopy images. The system has not been validated on prospective clinical cases or tested with clinicians. A smear image also cannot replace bone marrow examination, immunophenotyping or cytogenetics, which remain necessary for a diagnosis.

Future work will concentrate on closing the gap between training and test accuracy. We plan to fine-tune the upper layers of the backbone, compare other architectures such as ResNet50 [9], and add stain normalisation and stronger augmentation to help separate ALL, AML and CLL. A separate validation set will be created so that the test set stays untouched. Images from local laboratories will be collected for external validation, the gatekeeper will be trained with negative examples closer to real laboratory uploads, and the confidence scores will be calibrated so that uncertain cases can be sent to a specialist.

IX. CONCLUSION

This paper described a web-based system that screens peripheral blood smear images for four types of leukemia and for healthy samples. A gatekeeper network filters invalid uploads, a MobileNetV2 classifier assigns one of five labels, and a FastAPI and MongoDB back end carries each case to doctor review. On a balanced test set of 5,000 images the classifier reached an accuracy of 75.9%, against 97.1% on the training images. It separated healthy from leukemic smears reliably, with no leukemia image labelled healthy, but it often confused ALL, AML and CLL.

The results suggest that the approach is promising as a triage aid, and the overfitting and the weak separation among the leukemia types show where further work is needed: more varied training data, fine-tuning, external validation and a closer study of the misclassified cases. In its present form the system is intended to support the judgement of a doctor and not to replace it.

REFERENCES

- [1] G. Litjens et al., "A survey on deep learning in medical image analysis," *Medical Image Analysis*, vol. 42, pp. 60–88, 2017.
- [2] S. J. Pan and Q. Yang, "A survey on transfer learning," *IEEE Trans. Knowledge and Data Engineering*, vol. 22, no. 10, pp. 1345–1359, 2010.

- [3] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li and L. Fei-Fei, "ImageNet: A large-scale hierarchical image database," in *Proc. IEEE Conf. Computer Vision and Pattern Recognition (CVPR)*, 2009, pp. 248–255.
- [4] S. Shafique and S. Tehsin, "Acute lymphoblastic leukemia detection and classification of its subtypes using pretrained deep convolutional neural networks," *Technology in Cancer Research & Treatment*, vol. 17, Art. no. 1533033818802789, 2018.
- [5] A. Rehman, N. Abbas, T. Saba, S. I. ur Rahman, Z. Mehmood and H. Kolivand, "Classification of acute lymphoblastic leukemia using deep learning," *Microscopy Research and Technique*, vol. 81, no. 11, pp. 1310–1317, 2018.
- [6] M. Loey, M. Naman and H. Zayed, "Deep transfer learning in diagnosing leukemia in blood cells," *Computers*, vol. 9, no. 2, Art. no. 29, 2020.
- [7] S. H. Kassani, P. H. Kassani, M. J. Wesolowski, K. A. Schneider and R. Deters, "A hybrid deep learning architecture for leukemic B-lymphoblast classification," in *Proc. Int. Conf. Information and Communication Technology Convergence (ICTC)*, 2019, pp. 271–276.
- [8] M. Sandler, A. Howard, M. Zhu, A. Zhmoginov and L.-C. Chen, "MobileNetV2: Inverted residuals and linear bottlenecks," in *Proc. IEEE Conf. Computer Vision and Pattern Recognition (CVPR)*, 2018, pp. 4510–4520.
- [9] K. He, X. Zhang, S. Ren and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Computer Vision and Pattern Recognition (CVPR)*, 2016, pp. 770–778.
- [10] M. Abadi et al., "TensorFlow: Large-scale machine learning on heterogeneous distributed systems," arXiv:1603.04467, 2016.
- [11] PriyaadharshiniVS, "Leukemia Dataset," Kaggle. [Online]. Available: <https://www.kaggle.com/datasets/priyaadharshinivs062/leukemia-dataset>
- [12] M. Lin, Q. Chen and S. Yan, "Network in network," arXiv:1312.4400, 2013.
- [13] N. Srivastava, G. Hinton, A. Krizhevsky, I. Sutskever and R. Salakhutdinov, "Dropout: A simple way to prevent neural networks from overfitting," *J. Machine Learning Research*, vol. 15, no. 56, pp. 1929–1958, 2014.
- [14] A. Krizhevsky, "Learning multiple layers of features from tiny images," Univ. Toronto, Toronto, ON, Canada, Tech. Rep., 2009.
- [15] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," in *Proc. Int. Conf. Learning Representations (ICLR)*, 2015.