Deep Learning Methods for Peripheral Blood Cell Image Classification

BMI 707

Harshi Saha, Diana Hou, Tzu-Hsi Jen, Clementine Chen, Tina Tu

INTRODUCTION

- Differentiating peripheral blood cell types is crucial for detecting infections, immune responses, and disease progression [1].
- Traditional assessment is labor-intensive and error-prone [1, 2].

- Prior deep learning studies show strong performance:
  - **Custom CNNs** have achieved **97.98% accuracy** [4]:
    - Sequential model with convolution, normalization, dropout, and max and global average pooling layers
    - Requires extensive hyperparameter and architecture tuning
    - No extensive benefit over fine-tuned base models
  - **VGG-16** and **InceptionV3** have achieved up to **96% accuracy** [1]:
    - VGG 16: Uses multiple ReLU activated convolutional layers
      - vanishing gradient problem and shallow layer [5,7]
    - InceptionV3: Extracts information from different image scales
      - Consistent lower accuracy by previous study [8]
      - Complex architecture and more training challenges [7]
    - Computationally intensive (large epoch size and tuning effort) [1]
    - Shallow layer / vanishing gradient problem
    - Training stability / Architecture flexibility

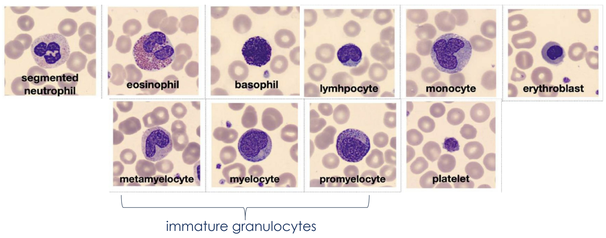
- Using **ResNet** [5, 7]:
  - Supports deeper architectures without vanishing gradient problem
  - Residual connections to skip layers when backpropagating gradient
  - Low training complexity due to simplified interface via keras
  - Compatible with interpretability tools like Grad-CAM
  - Expected to match or exceed prior accuracy

- **Objectives:**
  - Classify different types of peripheral blood cells with high accuracy using **ResNet CNN models** (ResNet-50 and ResNet-101)
  - Ensure robustness across cell types with different **image resolutions**
  - Enhance feature focus and model interpretability
  - Assess model performance by architecture and hyperparameters
  - Assess model architecture generalizability to other PBC imaging data

Dataset Information

- Original dataset [2]:
  - Sourced from Hospital Clinic of Barcelona (2015–2019).
  - Contains 17,092 normal peripheral blood cell images

  - **8** cell types: **neutrophils, eosinophils, basophils, lymphocytes, monocytes, immature granulocytes, erythroblasts, and platelets.**
  - .jpg format, 360×363 pixels, labeled by clinical pathologists.

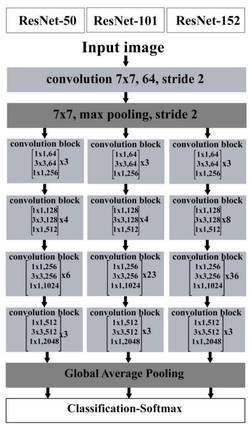


- MEDMNIST [3]:
  - Contained in the MedMNIST python package **(BloodMNIST module)**
  - Center cropped to 200×200 and resized to 28, 64, 128, and 224 pixels,
  - Split into training, validation, and test sets **(7:1:2 ratio)**
  - 8 cell types in the following proportions:
    - basophil , lymphocyte, monocyte, erythroblast **(~0.08)**
    - eosinophil, immature granulocyte, neutrophil **(~0.18)**
    - platelet **(~0.15)**

  - **Benchmarked** models with automated hyperparameter tuning:
    - ResNet-50 (28): accuracy 0.956, AUC 0.997
    - ResNet-50 (224): accuracy 0.950, AUC 0.997

METHODS

- ResNet CNNs: **ResNet-50 and ResNet-101**
  - 64×64, 128×128, and 224×224 peripheral blood cell images
  - Excluded ResNet-18 (not in Keras) and 28×28 inputs (below Keras's 32×32 minimum)
  - Not using ResNet-152 due to the computation cost



- **Total of 4 models for general performance checking**
  - ResNet-50 on 64×64 pixels
  - ResNet-101 on 64×64 pixels
  - ResNet-50 on 224×224 pixels
  - ResNet-101 on 224×224 pixels

- **Starting Hyperparameters**
  - Appended Dropout (rate = 0.3) and an 8-unit Softmax output layer
  - Compiled with Adam (learning rate = 1e-4) and categorical cross-entropy to track accuracy and AUC
  - Training augmentation: random shear (±0.2 rad), zoom (80–120%), horizontal flips
  - Uniform batch size of 16 across all experiments
- **Batch Size Comparison (batch size 16 vs. 32)**
  - Compared performance on original ResNet50 on 64x64 pixels

	Batch Size 16	Batch Size 32
Training Accuracy	0.9831	0.9856
Training AUC	0.9989	0.9993
Training Loss	0.0525	0.0412

- **Appended Attention Layer and Class Weighted Loss Function (batch size 32)**
  - No performance improvement → Kept during further hyperparameter tuning
  - Still no performance improvement → Excluded from final model

	Class Weighted	Class Weighted + Attention Layer
Training Accuracy	0.9774	0.9542
Training AUC	0.9984	0.9973
Training Loss	0.0670	0.1310

- **Optimizer Comparison (batch size 32)**
  - **Tested:**
    - SGD (learning rate = 1e-3, momentum = 0.8)
    - RMSprop (learning rate = 1e-4)
    - Adam (learning rate = 1e-4) → Best performance

Optimizer	Adam	RMSprop	SGD
Training Accuracy	0.9856	0.9839	0.9714
Training AUC	0.9993	0.9987	0.9984
Training Loss	0.0412	0.0497	0.0886

- **Dropout Rate Comparison (batch size 32, Adam at lr=1e-4))**
  - Compared dropout rates: 0.3, 0.4, 0.5

Dropout Rate	0.3	0.4	0.5
Training Accuracy	0.9856	0.9870	0.9842
Training AUC	0.9993	0.9995	0.9991
Training Loss	0.0412	0.0976	0.1154

(Note: Hyperparameter comparisons were conducted with ResNet50 on 64x64 pixels)

- **External dataset**:
  - Raabin WBC dataset with **5 cell classes**: Monocyte, Lymphocyte, Basophil, Eosinophil, Neutrophil.
  - Same model architecture
    - Base model: ResNet50
    - Input size: 64 × 64 × 3
    - Dropout (0.3) → Dense (5 softmax units)
    - Data **augmentation** (rescaling, shear, zoom, horizontal flip) with a 20% validation split.
- **Training**
  - Optimizer: Adam (learning rate 1e-4)
  - Loss: categorical cross-entropy
  - Metrics: accuracy, AUC (multi-label, 5 classes)
  - Trained for 10 epochs
- Explainability: **Grad-CAM**
  - Extracted feature maps from last ResNet block
  - Calculated gradient-based heatmaps for predicted class

RESULTS

- 224×224 pixel models exceeded **97%** test accuracy
- Loss plot result
  - 224 models showed occasional validation loss spikes
  - 64×64 models converged more gradually
  - ResNet-50's loss changes were steadier than ResNet-101's
  - Greater sensitivity of deeper networks to learning rate and regularization
- **Class-specific accuracy**
  - Eosinophils and platelets consistently achieved ~100% precision and recall on the test set
  - Monocytes and immature granulocytes had lower recall rates (< 0.9), suggesting persistent image or modeling challenges
- **Model comparison**

Model	ResNet101 (224×224)		ResNet101 (128x128)		ResNet101 (64x64)		ResNet50 (224×224)		ResNet50 (128x128)		ResNet50 (64x64)	
Batch Size	16	32	16	32	16	32	16	32	16	32	16	32
Neutrophil	0.98	1.00	0.97	0.92	0.98	0.96	0.98	0.99	0.97	0.99	0.98	0.98
Eosinophil	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
Basophil	0.99	1.00	0.99	0.99	0.98	0.98	1.00	0.99	0.94	0.98	0.98	0.98
Lymphocyte	0.99	0.96	0.97	0.93	0.96	0.92	0.98	0.94	0.97	0.97	0.96	0.95
Monocyte	0.99	0.99	0.97	0.99	0.90	0.96	0.97	0.98	0.94	0.99	0.89	0.98
Immature granulocyte	0.97	0.97	0.96	0.98	0.93	0.92	0.97	0.97	0.93	0.99	0.92	0.97
Erythroblast	0.99	0.98	0.99	0.97	0.99	0.97	0.99	0.95	0.96	0.98	0.98	0.97
Platelet	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00

Table 1: Class-level F1 score performance for dropout rate 0.3, Adam optimizer (1e-4), comparing batch sizes 16 vs 32

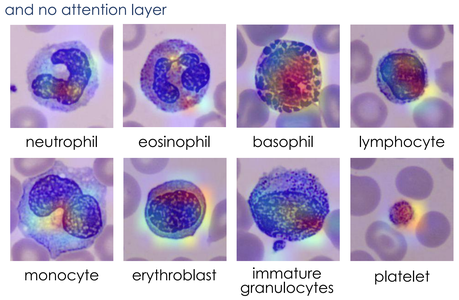
	Learning Rate →					
Optimizer	**1e-2**	**1e-3**	**5e-4**	**3e-4**	**1e-4**	**1e-5**
Adam	0.8173	0.3072	0.8997	0.8834	0.9553	0.9445
RMSprop	0.6144	0.8974	–	–	**0.9737**	0.9638
SGD (m=0.9)	–	0.8652	–	–	–	–
SGD (m=0.8)	0.9275	0.9617	–	–	0.9547	–
SGD (m=0.7)	–	0.9407	–	–	–	–
SGD (m=0.5)	–	0.9676	–	–	–	–

Table 2: Test accuracy of different optimizers and learning rates for ResNet50 (64×64 pixels, batch size 32, with attention layer and class-weighted loss)

	batch size = 16						batch size = 32					
	224×224		128×128		64×64		224×224		128×128		64×64	
	ResNet101	ResNet50	ResNet101	ResNet50	ResNet101	ResNet50	ResNet101	ResNet50	ResNet101	ResNet50	ResNet101	ResNet50
Test Accuracy	**0.9880**	0.9857	0.9807	0.9687	0.9693	0.9676	0.9848	0.9734	0.9728	0.9860	0.9667	0.9775
Test AUC	0.9989	0.9985	0.999	0.9984	0.9978	0.9956	0.9983	0.9984	0.9975	0.9989	0.9965	0.9966

Table 3: Test set performance of all models (dropout rate = 0.3, Adam 1e-4, batch size 16 vs. 32, without attention layer or class-weighted loss)

- **Hyperparameter tuning** in ResNet50, 64×64 model with fixed batch size 32:
  - With attention layer and class-weighted loss function: highest test accuracy 0.974 with RMSProp optimizer at lr=1e-4
  - Without attention layer or class-weighted loss function: highest test accuracy 0.978 with Adam optimizer at lr=1e-4
- **Attention layer and class-weighted loss function** : test accuracy 0.955 from 0.978; AUC 0.996 from 0.997
- **Optimal overall configuration by accuracy:** ResNet-101(224), batch size 16, Adam optimizer at lr=1e-4, dropout rate = 0.3, and no attention layer

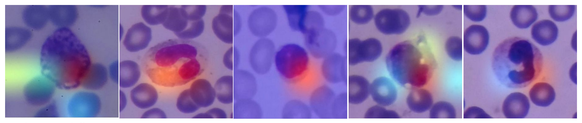


- Grad-CAM warmer areas are those more influential in final prediction
- Depicted ResNet-101 (224) detects subtleties in biological indicators
- Highest influence from intra-cell regions reflecting known morphology: monocyte u-shaped nucleus, basophilic granules, erythroblast and lymphocyte chromatin, eosinophil and neutrophil nucleus lobes

```
                precision    recall  f1-score   support

     basophil       1.00      0.94      0.97        17
   eosinophil       0.83      0.89      0.86        64
   lymphocyte       0.94      0.98      0.96       206
     monocyte       0.54      0.80      0.64        46
   neutrophil       0.99      0.92      0.95       532

     accuracy                           0.93       865
    macro avg       0.86      0.91      0.88       865
 weighted avg       0.94      0.93      0.93       865

basophil        accuracy: 0.9412
eosinophil      accuracy: 0.8906
lymphocyte      accuracy: 0.9757
monocyte        accuracy: 0.8043
neutrophil      accuracy: 0.9211
```

Table 4: Class-level performance on the Raabin dataset using ResNet-50 (64×64), dropout rate 0.3, Adam optimizer (1e-4), batch size 32

- Class-specific performance
  - Excellent performance with **Neutrophils** and **Basophils**
  - Low precision with **Monocytes**



CONCLUSION

- **ResNet-50 and ResNet-101** models were fit for resolutions **64**, **128**, and **224**
  - Unfreeze base layers and removed ImageNet classifier
  - Added GlobalAverage2D pooling layer
  - Used **Adam** optimizer, learning rate **1e-4**, dropout rate **0.3**
  - Used categorical cross-entropy and softmax output layer
- Overall high performance by accuracy and AUC but comparatively lower predictive power on the monocyte and immature granulocytes classes
- The best model prior to tuning had **accuracy 0.988** with **AUC 0.999** ResNet-101 (224), batch size 16, Adam lr1e-4, dropout=0.3, no attention layer

- Tested hyperparameters (**batch size** 16 and 32, **dropout rate** 0.3-0.5, **lr** 1e-2 and 1e-4), **optimizers** (Adam, SGD, RMSProp), **attention layers**, and **class-weighted loss function**
- Best model by overall accuracy and AUC is still the one above

- Higher **batch size** improved **F-1 scores** for lower performance classes
- ResNet-50 (64) **0.3 dropout rate** had highest test AUC and accuracy mitigating overfitting and preserving model capacity
- **Adam** with **lr=1e-4** outperformed higher and lower rates across optimizers with stable validation metrics and smooth convergence
- **Attention layers and class-weighted loss function** reduced test accuracy by up to 2%, suggesting increased complexity may have introduced noise or overfitting

- **Grad-CAM** results showed that the models captured biological patterns
- Extension of model architecture and hyperparameter choices to other data showed good performance and similar interpretable model feature focus

- Future directions include further **hyperparameter** tuning, model **architecture** changes, and testing **generalizability** to other PBC imaging datasets

SELECTED REFERENCES

1. Acevedo A;Alférez S;Merino A;Puigví L;Rodellar J; (n.d.). Recognition of peripheral blood cell images using convolutional neural networks. Computer methods and programs in biomedicine. https://doi.org/10.1016/j.cmpb.2019.105020
2. Acevedo A;Merino A;Alférez S;Molina Á;Boldú L;Rodellar J; (n.d.). A dataset of microscopic peripheral blood cell images for development of Automatic Recognition Systems. Data in brief. https://doi.org/10.1016/j.dib.2020.105474
3. Yang, J., Shi, R., Wei, D., Liu, Z., Zhao, L., Ke, B., Pfister, H., & Ni, B. (2023, January 19). MedMNIST v2 - a large-scale lightweight benchmark for 2D and 3D Biomedical Image Classification. Nature News. https://doi.org/10.1016/j.dib.2020.105474
4. Anand, V., Gupta, S., Koundal, D., Alghamdi, & Alsharbi, B. M. (2024). Deep learning-based image annotation for leukocyte segmentation and classification of blood cell morphology. BMC Medical Imaging, 24, 83. https://doi.org/10.1186/s12880-024-01254-z
5. Yuan Yang, Lin Zhang, Mingyu Du, Jingyu Bo, Haolei Liu, Lei Ren, Xiaohe Li, M. Jamal Deen, A comparative analysis of eleven neural networks architectures for small datasets of lung images of COVID-19 patients toward improved clinical decisions, Computers in Biology and Medicine, https://doi.org/10.1016/j.compbiomed.2021.104887.
6. Jung, C., Abuhamad, M., Mohaisen, D., Han, K., & Nyang, D. (2022). WBC image classification and generative models based on convolutional neural network. BMC Medical Imaging, 22(1). https://doi.org/10.1186/s12880-022-00818-1
7. Saidani O, Umer M, Alturki N, Alshardan A, Kiran M, Alsubai S, Kim TH, Ashraf I. White blood cells classification using multi-fold pre-processing and optimized CNN model. Sci Rep. 2024 Feb 12;14(1):3570. doi: 10.1038/s41598-024-52880-0. PMID: 38347011; PMCID: PMC10861568.
8. Habibzadeh M, Jannesari M, Rezaei Z, Baharvand H, Totonchi M. Automatic white blood cell classification using pre-trained deep learning models: resnet and inception. In: International conference on machine vision (ICMV), 2018. p. 274–81.