BENCHMARKING FEW-SHOT GPT-4O AGAINST SUPERVISED FINE-TUNING FOR MORPHOLOGIC IDENTIFICATION OF PERIPHERAL BLOOD SMEARS

MOBINA SHRESTHA, MD Sentara Albemarle Medical Center, USACS

M Shrestha[1], B Mandal[2] and V Mandal[3]

1. Sentara Albemarle Medical Center, Elizabeth City, NC United States
2. Kansas State University, Manhattan, KS United States
3. University of Missouri-Columbia, United States

US Acute Care Solutions

Introduction

Peripheral blood smear review remains crucial for hematologic diagnosis, as automated analyzers often miss key morphologic abnormalities. Manual interpretation by experts ensures accuracy but is time-consuming and subject to variability. These limitations highlight the need for supportive technologies that combine efficiency with consistency. Artificial intelligence (AI) offers a promising solution by learning from large image datasets to detect subtle morphologic changes, potentially improving diagnostic accuracy and accessibility.

Aim

To explore the potential of GPT-4o, a vision-language foundation model, in classifying peripheral blood smears using few-shot prompting. The study specifically aimed to assess its ability to distinguish eight morphologic cell types: neutrophil, eosinophil, basophil, lymphocyte, monocyte, immature granulocyte, erythroblast, and platelet, and evaluate its usefulness as a supportive tool in hematology, particularly where expert review and labeled datasets are limited.

Objectives

To benchmark GPT-4o with few-shot prompting against supervised models ResNet-50 and ViT trained on the BloodMNIST dataset (16,990 images). Testing was done on 192 images (24 per class) The objective was to compare their accuracies and see how close can GPT-4o get to fully fine-tuned models in peripheral blood smear classification.

Schematic Workflow

Dataset Preparation

- **Dataset**: BloodMNIST
- **Classes**: 8 blood cell types
- **Size**: Resized to 224×224
- **Modality**: RGB
- **Test**: 192 images (24 per class) (Supervised models and GPT-4o)

Learning Strategies

Supervised

- **Models**: ResNet-50, ViT-B/16
- **Train**: 10 epochs
- **Hardware**: NVIDIA A100 GPU

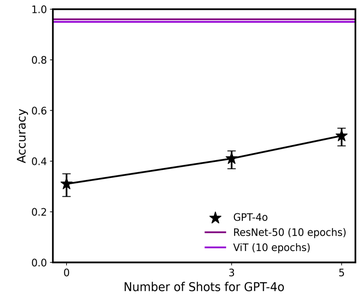
ICL (GPT-4o)

- **Shots**: 0, 3, 5
- **Support**: image + label pairs
- **Query**: image + prompt
- **Parameter Updates**: none

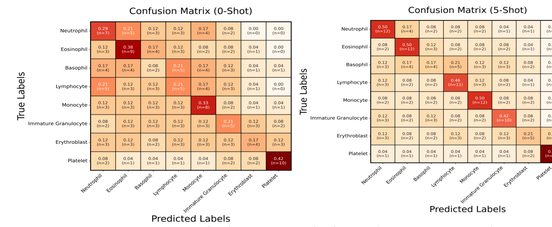
Evaluation

- **Metrics**: **Accuracy, 95% CI**
- **Confusion matrix per-class**
- **Same test set across all models**
- **No overlap with training data**

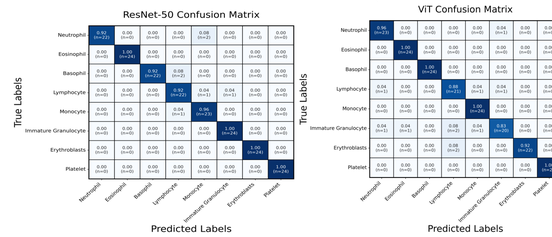
Results



Performance comparison of GPT-4o few shot vs. supervised methods



Confusion matrices for GPT-4o (zero-shot and 5-shot)



Confusion matrices for ResNet-50 and ViT

Summary

GPT-4o reached 50% accuracy in 5-shot setting for classifying 8 blood cell types, improving with more examples. Supervised models, ResNet-50 and ViT exceeded 95% accuracy, highlighting both the difficulty of this multi-class task and the promise of few-shot learning in data-limited settings.

Discussion

- GPT-4o showed progressive improvement with few-shot prompting:
  - 0-shot: 0.31 (95% CI: 0.26-0.35)
  - 3-shot: 0.41 (CI: 0.37-0.44)
  - 5-shot: 0.50 (CI: 0.46-0.53)
- Supervised models outperformed few-shot prompting:
  - ResNet-50: 0.96 (CI: 0.94-0.98)
  - Vision Transformer (ViT): 0.95 (CI: 0.92-0.98)
- Improvements in recall for several classes (e.g., Platelets up to 0.67, Monocytes up to 0.50) demonstrate potential adaptability of GPT-4o with limited examples.

References

1. Arber, D. A., Orazi, A., Hasserjian, R., Thiele, J., Borowitz, M. J., Le Beau, M. M., ... & Vardiman, J. W. (2016). The 2016 revision to the World Health Organization classification of myeloid neoplasms and acute leukemia. *Blood, the journal of the american society of hematology*, *127*(20), 2391-2405.
2. Matek, C., Schwarz, S., Spiekermann, K., & Marr, C. (2019). Human-level recognition of blast cells in acute myeloid leukaemia with convolutional neural networks. *Nature Machine Intelligence*, *1*(11), 538-544.
3. Ferber, D., Wölflein, G., Wiest, I. C., Ligero, M., Sainath, S., Ghaffari Laleh, N., ... & Kather, J. N. (2024). In-context learning enables multimodal large language models to classify cancer pathology images. *Nature Communications*, *15*(1), 10104.
4. Shrestha, M., Mandal, B., Mandal, V., & Shrestha, A. (2025). In-Context Learning for Label-Efficient Cancer Image Classification in Oncology. *Informatics in Medicine Unlocked*.

SOHO2025