Semi-Supervised Active Learning for the Classification of Blood Cells

Thanushika,T., Manivannan, S., and Nirthika, R.
Department of Computer Science, University of Jaffna, Jaffna, Sri Lanka
{2019sp255, siyam, nirthika}@univ.jfn.ac.lk

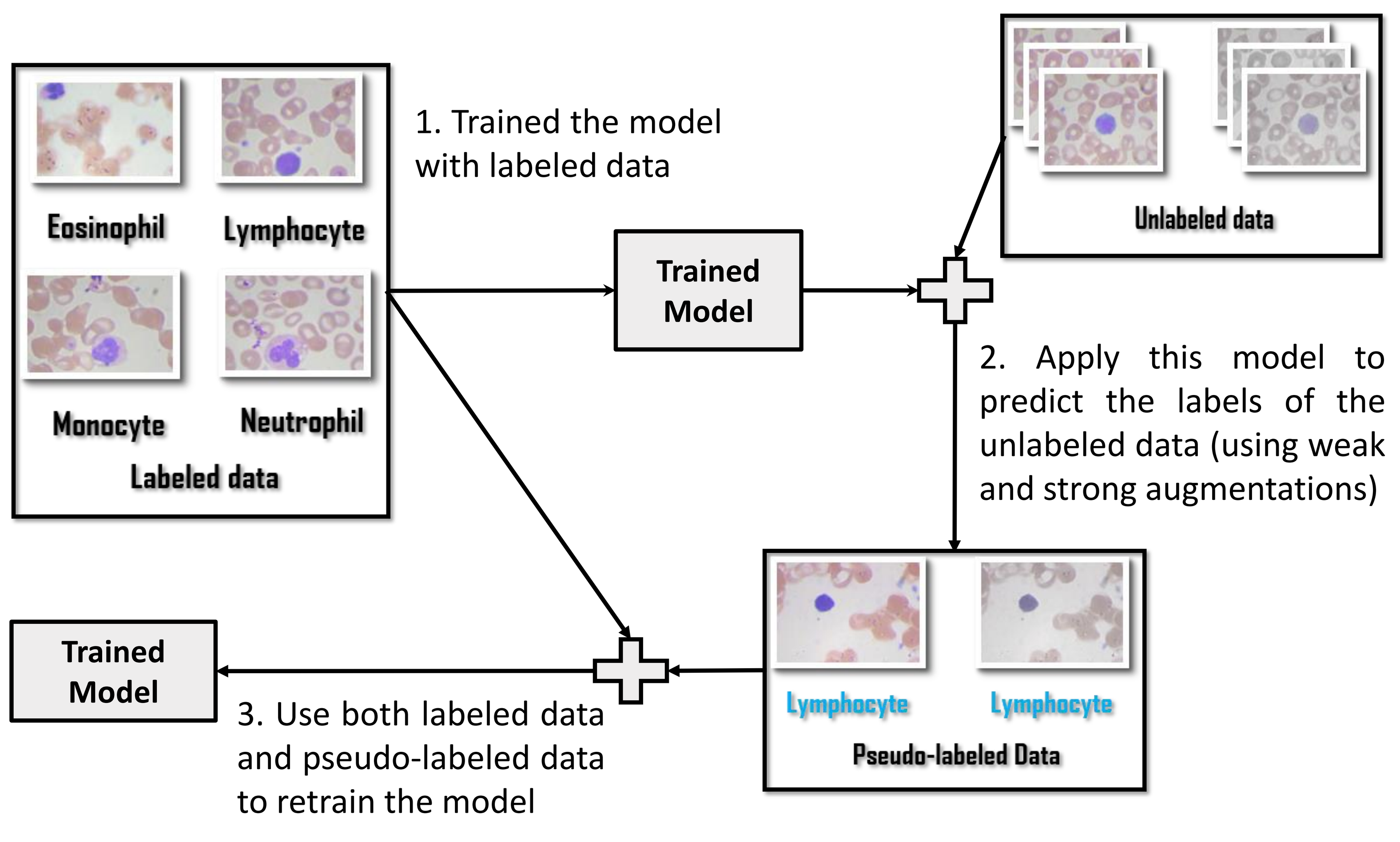
DCS UNIVERSITY OF JAFFNA

Introduction and Motivation

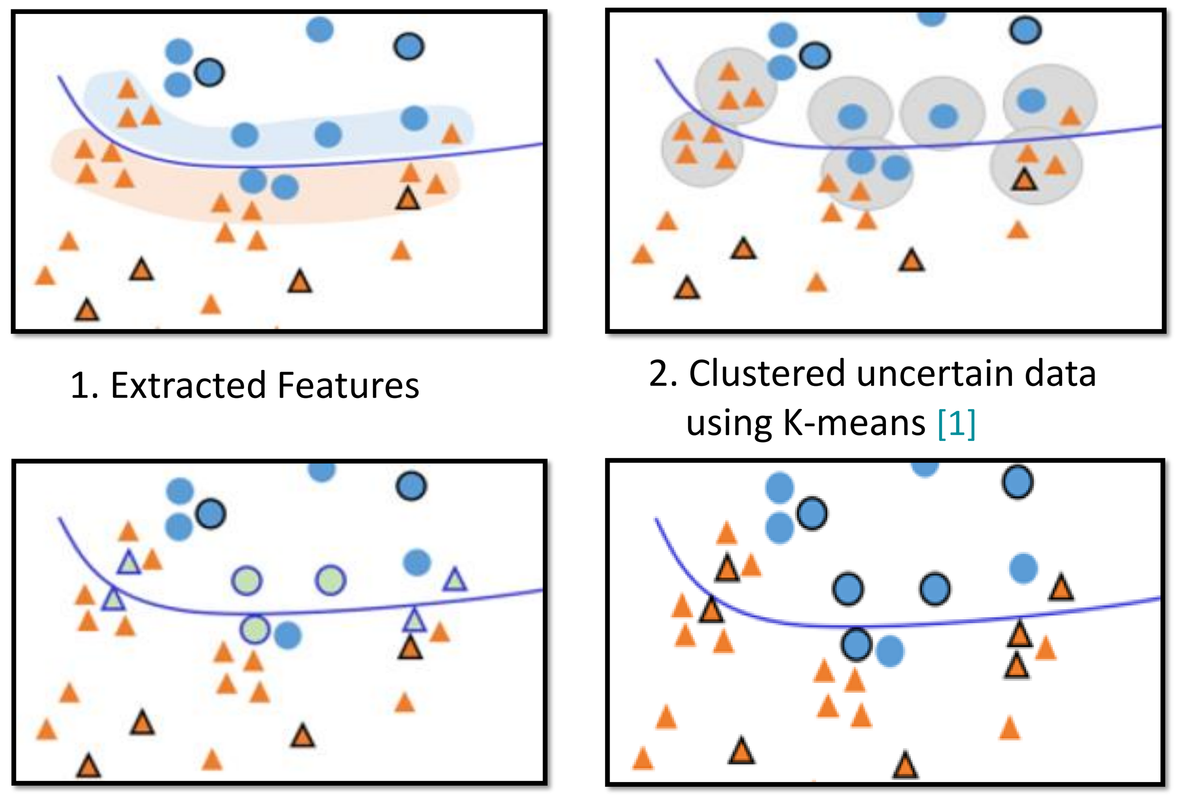
Blood cell classification plays a critical role in medical diagnostics, particularly for diseases like anemia and leukemia. Traditional automated methods require large amounts of labeled data, which is costly and time-consuming to obtain. In this study, we aim to address these challenges by using a Semi-Supervised Learning (SSL) approach combined with Active Learning (AL). SSL leverages both labeled and unlabeled data, reducing the need for extensive manual labeling, while AL focuses on identifying the most informative samples for labeling. Experimental evaluations on the widely-used Blood Cell Count and Detection (BCCD) dataset demonstrate that our approach achieves superior results in terms of accuracy and F1-scores, outperforming state-of-the-art methods, all while using only a small fraction of labeled data. This combination aims to improve the efficiency and accuracy of blood cell classification, making it more practical for real-world applications.

Methodology

SSL – leverages unlabeled data to improve model performance

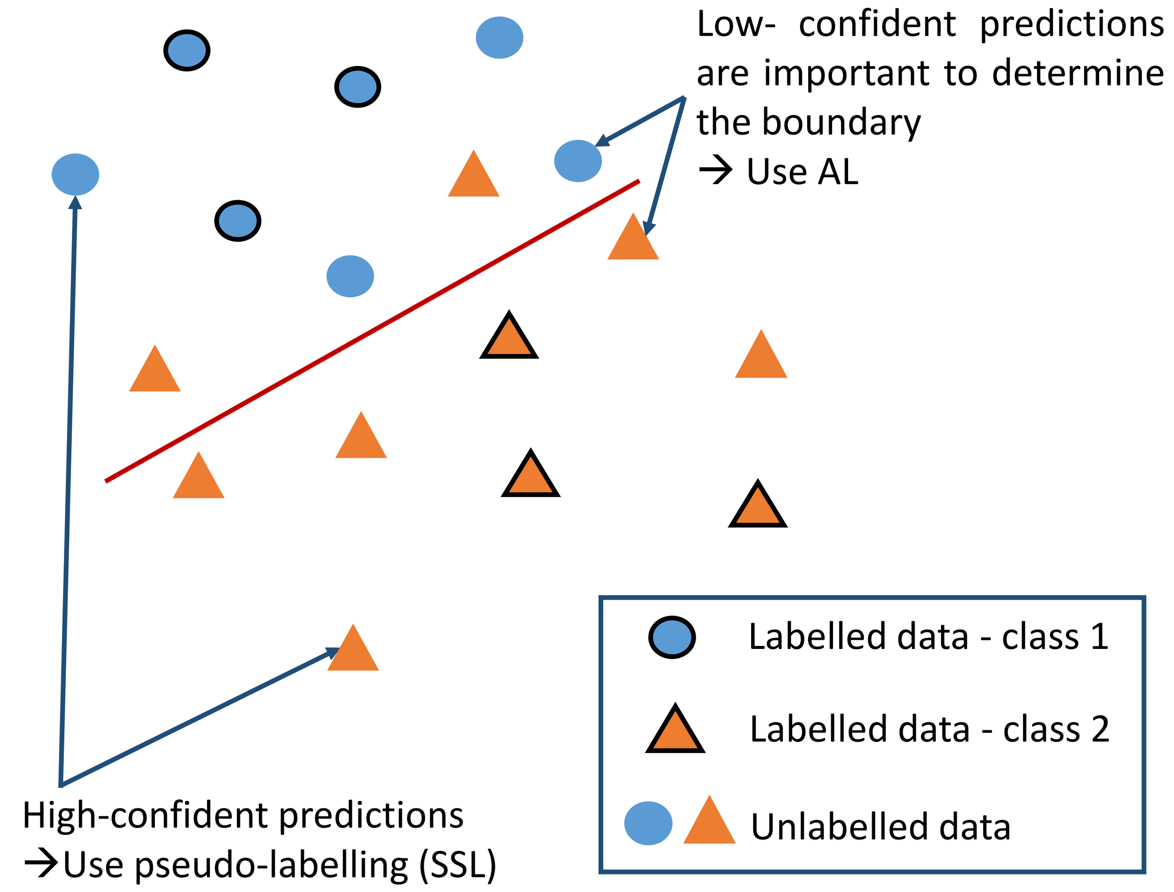


AL – identifies the most informative samples for human annotations.



1. Extracted Features
2. Clustered uncertain data using K-means [1]
3. Identified the representative samples from cluster
4. Annotations were obtained for identified representative samples

SSL + AL – optimize the use of both labeled and unlabeled data.



Overall Algorithm

- f_1 – model training using labelled data D_l
- For i = 1 to n:
  - // SSL
    - Select high-confident data (H) using SSL based on f_i
  - $D_l = D_l \cup H$
  - // AL
    - Select low-confident data (L) using AL based on f_i
  - $D_l = D_l \cup L$
- f_{i+1} - Retrain using the augmented labelled set D_l

Loss Function

No. of samples   No. of classes

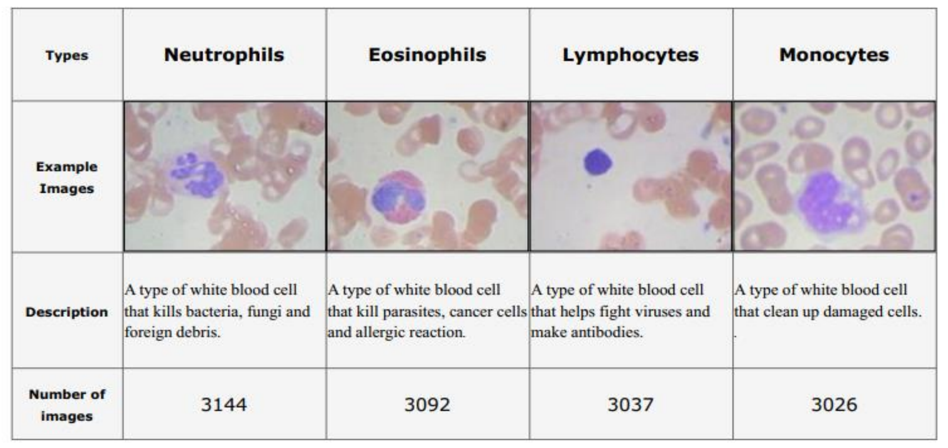
$$\mathcal{L}_{CEw} = -\sum_{i=1}^{n}\sum_{l=1}^{c} w_l \mathbb{1}_{(y_i=l)} \log \mathbf{P}_{il}$$

Predicted probability for class l of i^{th} sample.

Class weights   Indicator function

Dataset

BCCD[1] containing total of 12,444 images and four classes.
Sample images:



Types	Neutrophils	Eosinophils	Lymphocytes	Monocytes
Example Images				
Description	A type of white blood cell that kills bacteria, fungi and foreign debris.	A type of white blood cell that kill parasites, cancer cells and allergic reaction.	A type of white blood cell that helps fight viruses and make antibodies.	A type of white blood cell that clean up damaged cells.
Number of images	3144	3092	3037	3026

[1] https://www.kaggle.com/datasets/paultimothymooney/blood-cells

Implementation Details

- Backbone: ResNet-18 (pre-trained on ImageNet)
- Data: training - 50%, testing - 50%
- From training set: labeled - 2%, unlabeled - 98%
- Learning rate: 0.01, Optimizer: SGD
- Batch size: 64, No. of epochs: 40
- Budget: 100
- Evaluation Criteria
  - Accuracy
  - F1 - Score

  $\Big\}$ Mean ± standard deviation over 3 experiments

Results & Discussion

Threshold - 0.97, Budget Size - 100

Iteration	Accuracy (%)		
	SSL	AL	SSL+AL
Baseline	81.50 ± 0.42		
1st	85.73 ± 0.53	89.95 ± 0.80	89.90 ± 1.28
2nd	87.51 ± 0.69	71.63 ± 5.89	93.29 ± 1.56
3rd	88.04 ± 0.71	94.46 ± 0.16	95.28 ± 1.36

Iteration	F1-Score (%)		
	SSL	AL	SSL+AL
Baseline	81.66 ± 0.39		
1st	85.73 ± 0.55	89.97 ± 0.81	89.90 ± 1.27
2nd	87.47 ± 0.70	72.17 ± 5.45	93.29 ± 1.56
3rd	88.00 ± 0.71	94.47 ± 0.16	95.27 ± 1.37

Note: Baseline refers to the supervised learning on 2% labeled data.
Threshold is the confidence level for accepting a model's prediction as a label in SSL.
Budget Size refers to the number of samples selected for labeling during AL.

Ablation Study

Selection of the best budget size for clustering in AL and threshold value for pseudo-labeling in SSL.

Iteration	Accuracy (%)	
	Budget - 50	Budget - 100
1st	87.59 ± 0.06	89.95 ± 0.80
2nd	90.15 ± 0.37	71.63 ± 5.89
3rd	89.64 ± 0.95	94.46 ± 0.16

Threshold	Accuracy (%)
0.5	83.96 ± 0.83
0.7	85.09 ± 0.33
0.9	86.66 ± 0.44
0.97	88.04 ± 0.71
0.99	87.65 ± 0.29

Comparison with State-of-the-art

Method	Labeled	Unlabeled	Accuracy (%)
MT† [3]	20%	80%	94.42
SRC-MT† [4]	20%	80%	94.57
FixMatch† [5]	20%	80%	94.24
FixMatch+DARP† [6]	20%	80%	94.56

Method	Labeled	Unlabeled	Accuracy (%)
FlexMatch† [7]	20%	80%	94.50
FullMatch† [8]	20%	80%	94.88
SPLAL [2]	20%	80%	95.13
Ours (Baseline)	2%	98%	81.66 ± 0.39
Ours (SSL + AL)	2%	98%	95.28 ± 1.36

† represents that the results are taken from SPLAL [2]

Conclusion

The proposed semi-supervised active learning approach has shown promising results in classifying blood cells with high accuracy using limited labeled data. By combining the strengths of semi-supervised learning and active learning, we can achieve competitive performance while significantly reducing the need for labeled datasets. This technique can be highly beneficial in medical diagnostics and other domains where data labeling is resource-intensive. Future work can focus on refining the model and applying it to other medical datasets to validate its broader applicability.

References

[1] Manivannan, S. (2024). Pseudo-labeling and clustering-based active learning for imbalanced classification of wafer bin map defects. Signal, Image and Video Processing, vol.18(3), pp.2391-2401.
[2] Mahmood, M. J., *et al.* (2024). SPLAL: Similarity-based pseudo-labeling with alignment loss for semi-supervised medical image classification. Biomedical Signal Processing and Control, vol.89, pp.105665.
[3] Tarvainen, A., *et al.* (2017). Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results. Advances in neural information processing systems, vol.30.
[4] Liu, Q., *et al.* (2020). Semi-supervised medical image classification with relation-driven self-ensembling model. IEEE transactions on medical imaging, vol.39(11), pp.3429-3440.
[5] Sohn, K., *et al.* (2020). Fixmatch: Simplifying semi-supervised learning with consistency and confidence. Advances in neural information processing systems, vol.33, pp.596-608.
[6] Kim, J., *et al.* (2020). Distribution aligning refinery of pseudo-label for imbalanced semi-supervised learning. Advances in neural information processing systems, vol.33, pp.14567-14579.
[7] Zhang, B., *et al.* (2021). Flexmatch: Boosting semi-supervised learning with curriculum pseudo labeling. Advances in Neural Information Processing Systems, vol.34, pp.18408-18419.
[8] Peng, Z., *et al.* (2023). Semi-supervised medical image classification with adaptive threshold pseudo-labeling and unreliable sample contrastive loss. Biomedical Signal Processing and Control, vol.79, pp.104142.