ĐẠI HỌC
BÁCH KHOA HÀ NỘI
HANOI UNIVERSITY
OF SCIENCE AND TECHNOLOGY

Multi-Domain Information Fusion for Plasmodium Life Cycle Development Classification

Tran Quoc Khanh 20190085
Major: Data Science

Supervisor: Dr. Nguyen Thi Oanh

ONE LOVE. ONE FUTURE.

Agenda

- Introduction
- Methodology for Domain Fusion
- Experiment Settings
- Experiment Results and Analyses
- Conclusion and Future Works

Introduction

Motivation

- Profound impact of malaria disease (bệnh sốt rét) caused by plasmodium parasite:
  - 597 thousands deaths in 83 countries in 2023 [1].
  - 1.3% lower in annual GDP growth [2].
  - Responsible for 50% missed school days of children in Africa [3].
- Early diagnosis is important for malaria disease prevention [1].
- Life cycle development is important for clinical testing of medicine.
- Manual diagnosis requires **highly-skilled experts** [4].

→ **Classification of plasmodium life cycle development.**

Context of Thesis Study

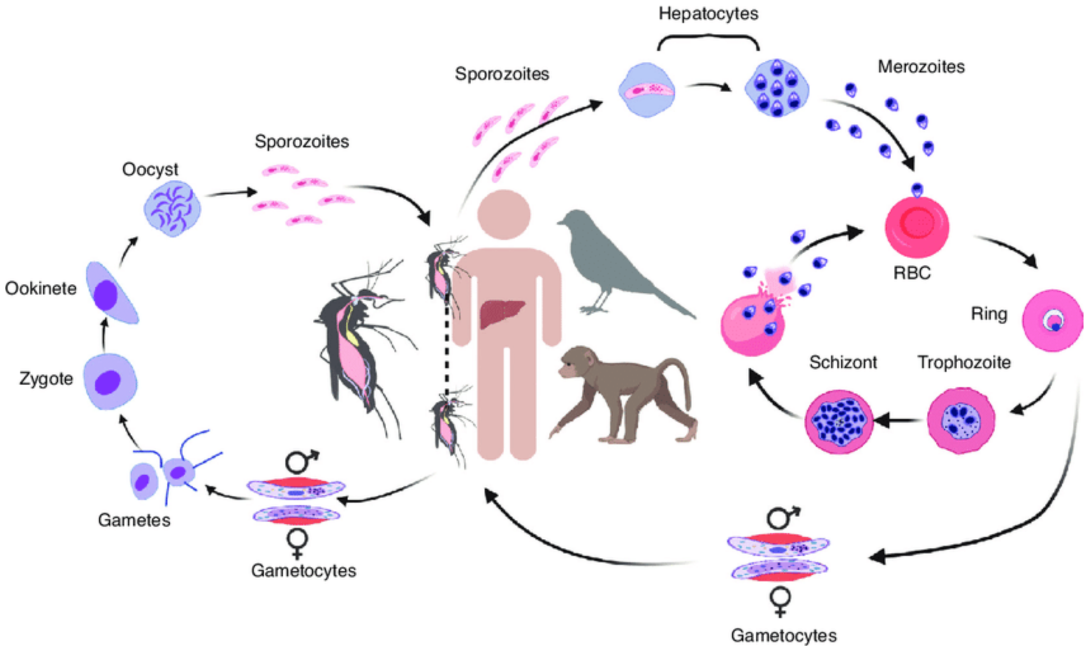
- Collaboration with French Armed Forces Center for Epidemiology and Public Health (CESPA) and French Armed Forces Biomedical Research Institute (IRBA).
- Research project goal: Development of **automatic clinical testing** for malaria medicine, i.e. to assess whether the new medicine stops the plasmodium development at a desired development stage.

IRBA

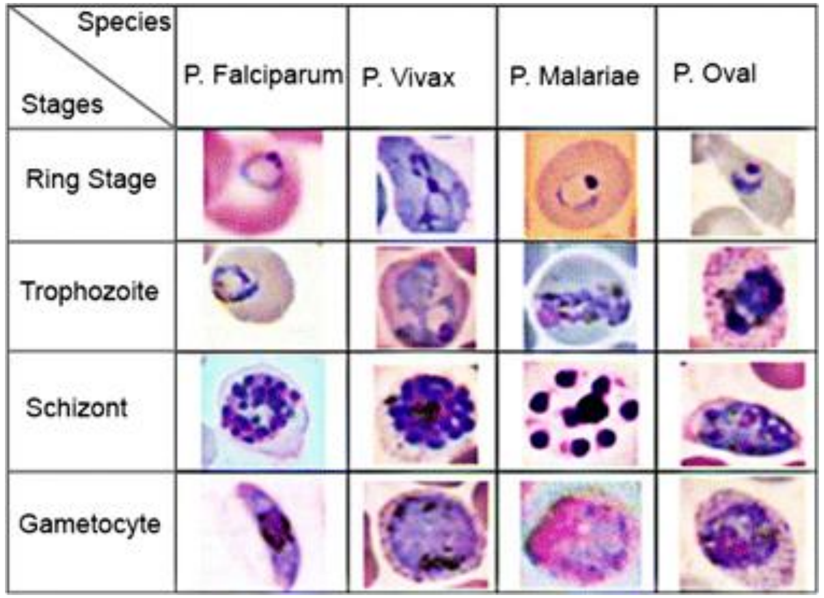
CESPA
Centre d'épidémiologie et de santé publique des armées
Observer Analyser Agir

Medical Background of Plasmodium

- Plasmodium development consists of mosquito, human liver, and human blood stages.
- Focus: classification of life cycle development in **human blood**.



Development of Plasmodium. Figure extracted from Su et al. [5]

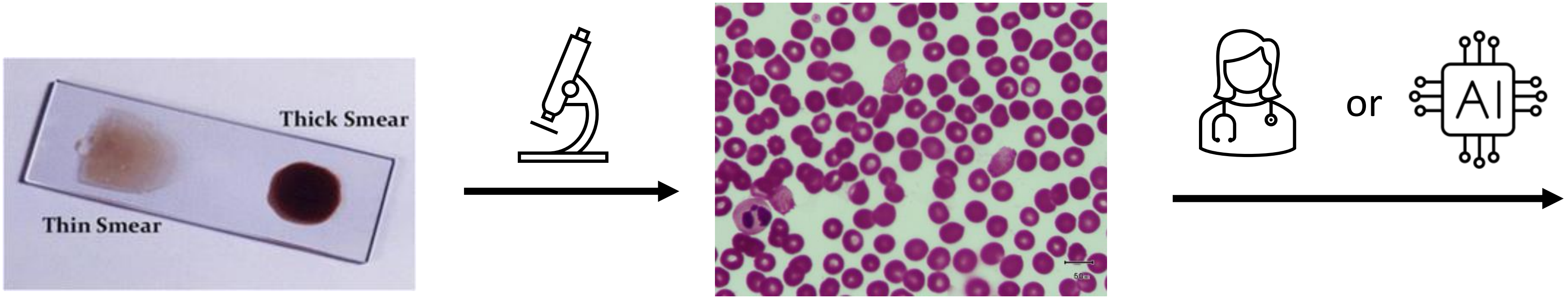


Species / Stages	P. Falciparum	P. Vivax	P. Malariae	P. Oval
Ring Stage				
Trophozoite				
Schizont				
Gametocyte				

Samples of life cycle development in blood. Figure extracted from Jan et al. [4]

Microscopy Diagnosis of Plasmodium

- Blood smear images consists of several red blood cells (RBC).
- One RBC is either healthy (uninfected) or infected (by parasite).
- Infected RBC is either in four life cycle development stages: ring, trophozoite, schizont, and gametocyte.

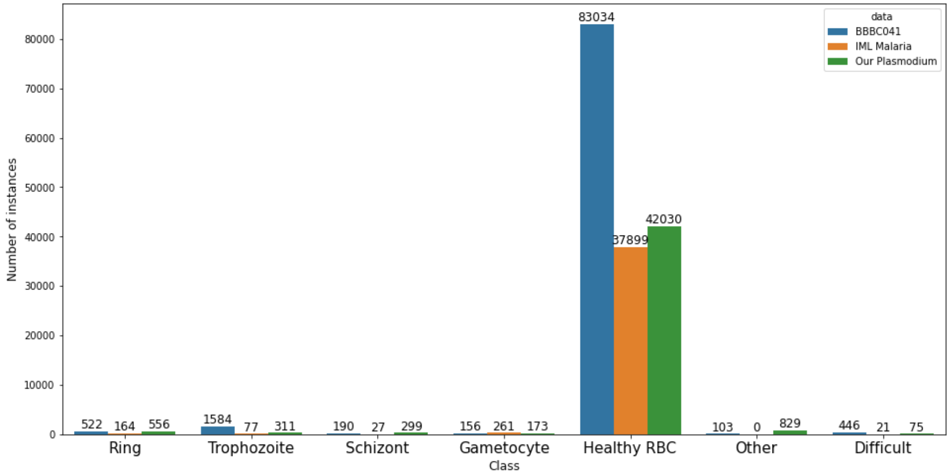


- Is the blood sample infected?
- In which development stage?
- Caused by which species?

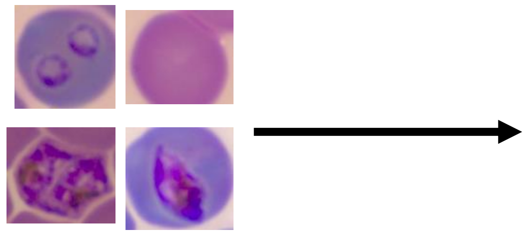
Pipeline for plasmodium diagnosis, carried out either manually by experts or semi-automatically by deep learning model.

Research Problem

- Research problem: Data imbalance of life cycle development **classification** from RBCs.
- State of Research:
  - Good results on detection of RBCs from blood smear images [6], [7].
  - Addressing data imbalance issue in life cycle classification:
    - Employing additional model for infection classification of RBCs [11], [12] → computational demand.
    - Utilizing additional unlabeled data [10] → ineffective use of data.



Number of instances by class in three datasets. All are dominated by healthy red blood cells (healthy RBC).



1. Healthy RBC

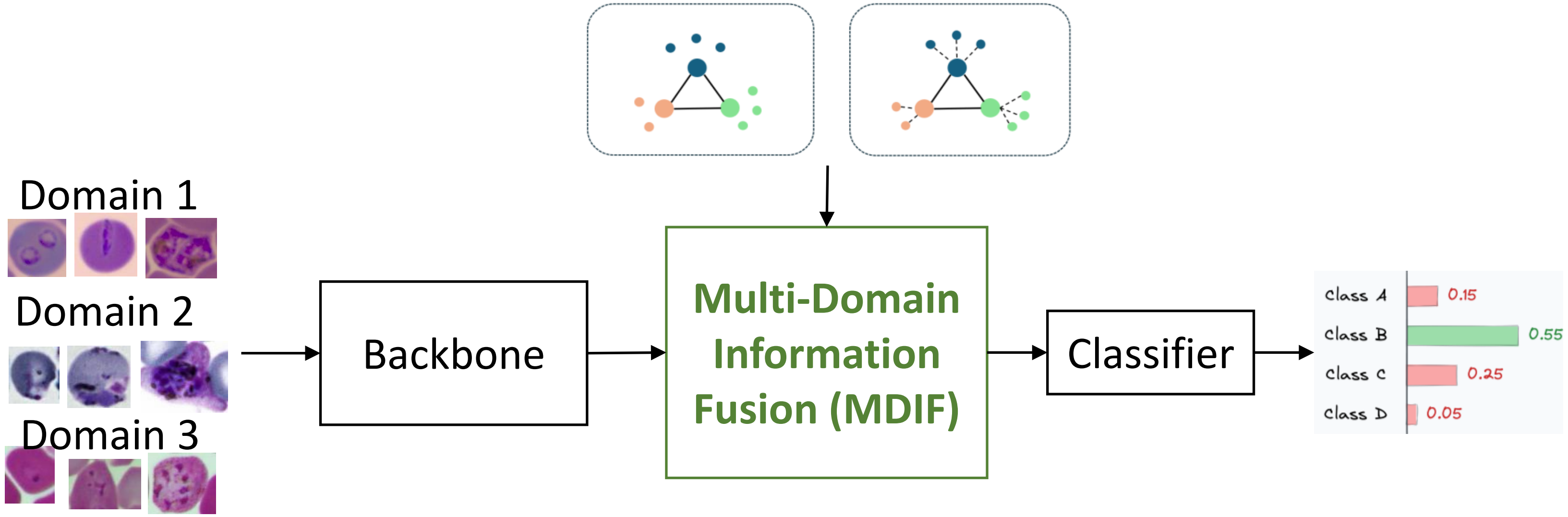
Life cycle stage:

2. Ring
3. Trophozoite
4. Schizont
5. Gametocyte

Illustration of our classification pipeline.

Objective and Contributions

- Address severe data imbalance issue in life cycle development **classification** of RBCs by **enriching minor classes** with **data from multiple domains**.
- Contributions:
  1. Introduce multi-domain learning concept to the plasmodium life cycle classification task.
  2. Propose **Multi-Domain Information Fusion** (MDIF) to bridge domain gap.



Life cycle development classification framework with MDIF.

Methodology for Domain Fusion

Why We Need Multi-Domain Information Fusion (MDIF)?

- Different data acquisition methods of different datasets, e.g. staining agents, microscope devices → different distribution in feature space.
- MDIF to **bridge the distributions across domain**.

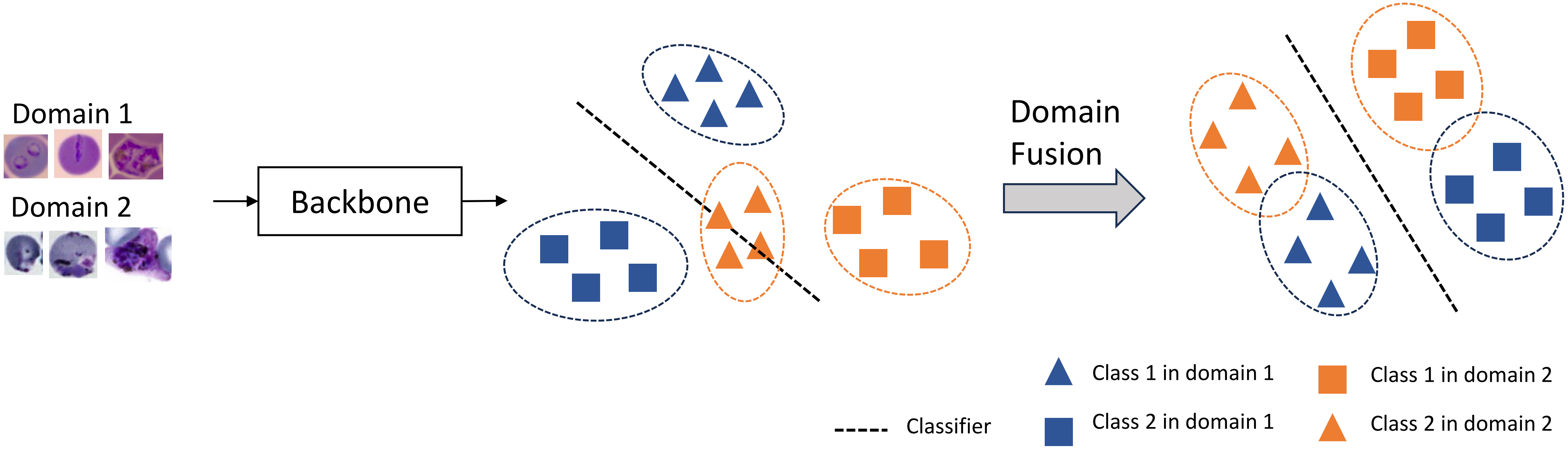
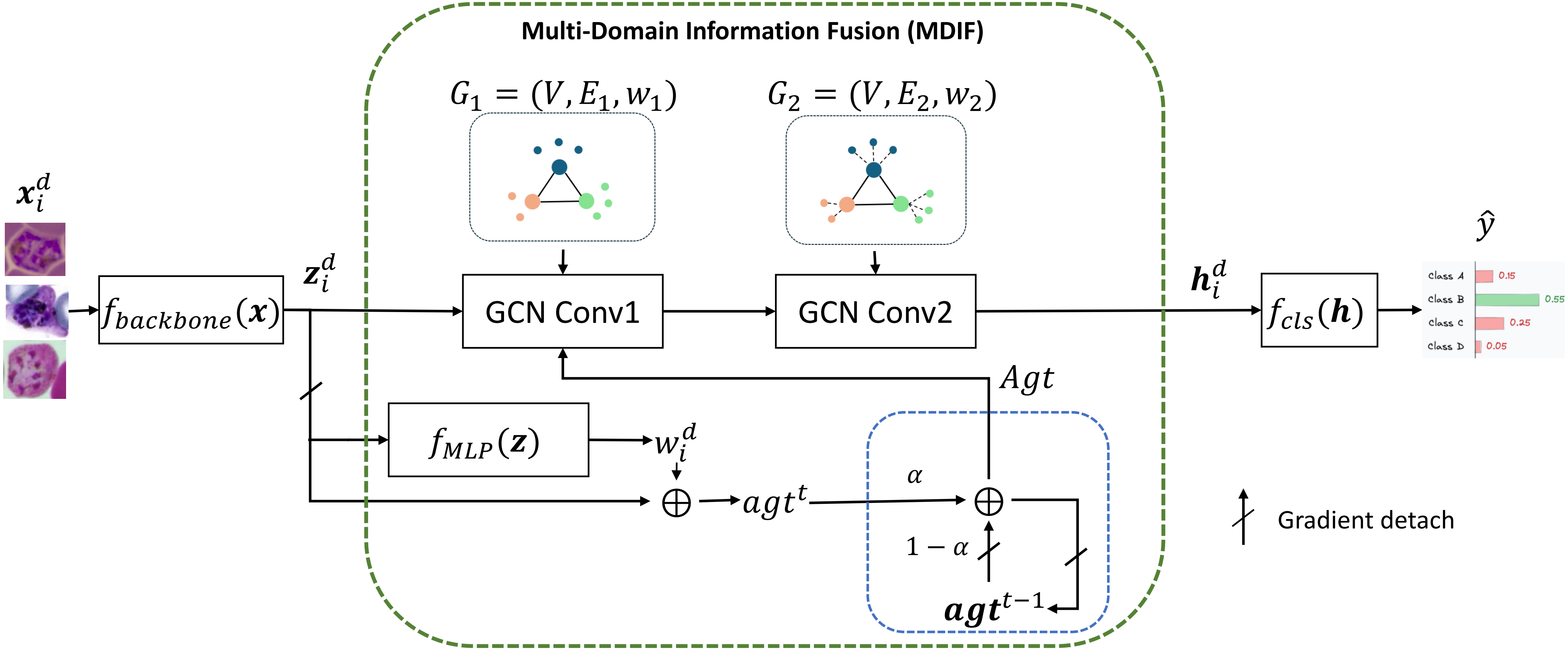


Illustration of utilizing MDIF to bridge distributions across domain. Colors represent domains.

Multi-Domain Information Fusion (MDIF)

- A module placed on top of backbone (feature extractor) to integrate information across domains at feature level.
- Components of MDIF:
  - **Two GCN Layers**:
    - First layer: integrate information globally.
    - Second layer: all instance nodes receive global information.
  - **Agent node**: noise-tolerant global representation.
  - **Knowledge graph** (adjacency matrix) defines how information flows.
- Two versions: MDIF Domain-level and MDIF Class-level.

Classification Framework with MDIF



Multi-Domain Information Fusion (MDIF)

$\boldsymbol{x}_i^d$

$f_{backbone}(\boldsymbol{x})$

$\boldsymbol{z}_i^d$

$G_1 = (V, E_1, w_1)$

$G_2 = (V, E_2, w_2)$

GCN Conv1

GCN Conv2

$\boldsymbol{h}_i^d$

$f_{cls}(\boldsymbol{h})$

$\hat{y}$

Class A 0.15

Class B 0.55

Class C 0.25

Class D 0.05

$f_{MLP}(\boldsymbol{z})$

w_i^d

$\oplus$

agt^t

α

$1 - \alpha$

Agt

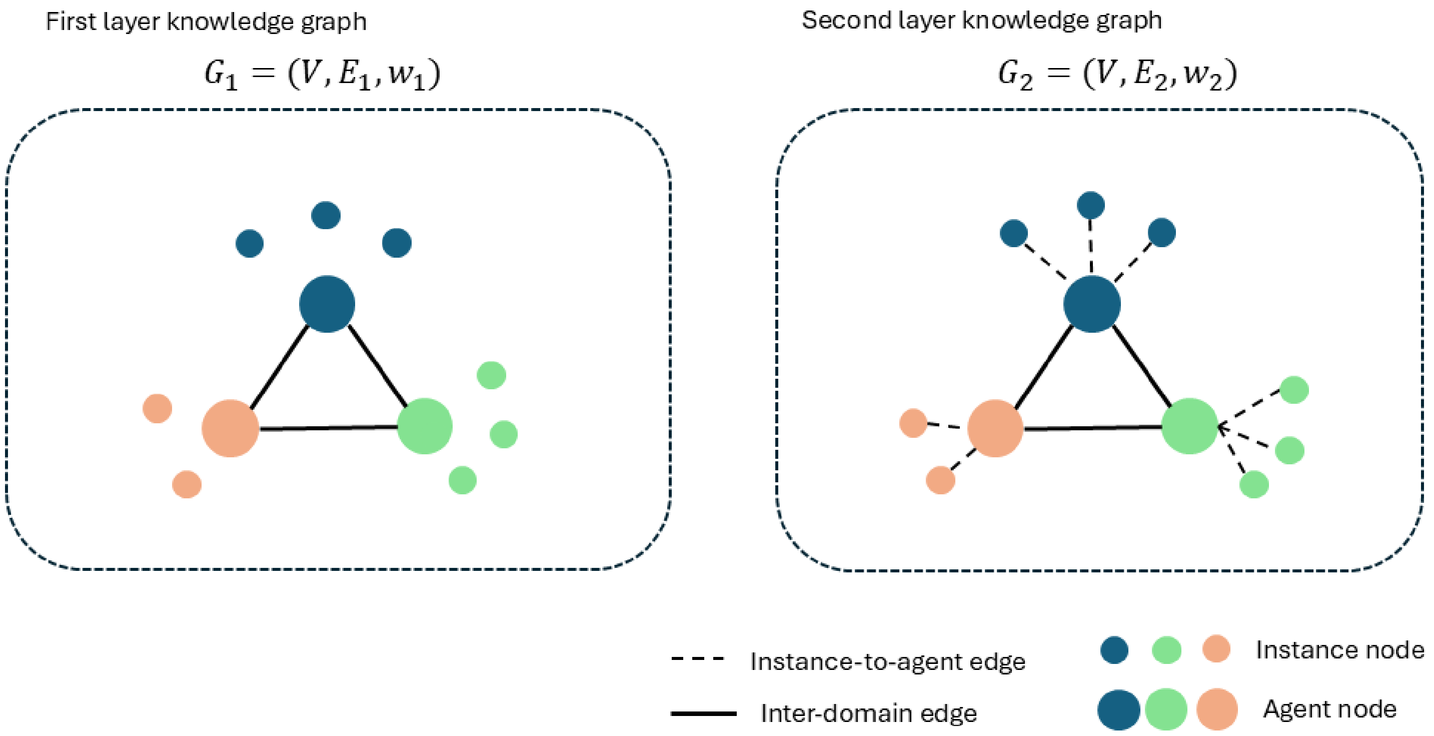
$\boldsymbol{agt}^{t-1}$

Gradient detach

Classification framework with MDIF.

MDIF Domain-Level

- Integrate information at domain level.
- An agent node represents a domain.



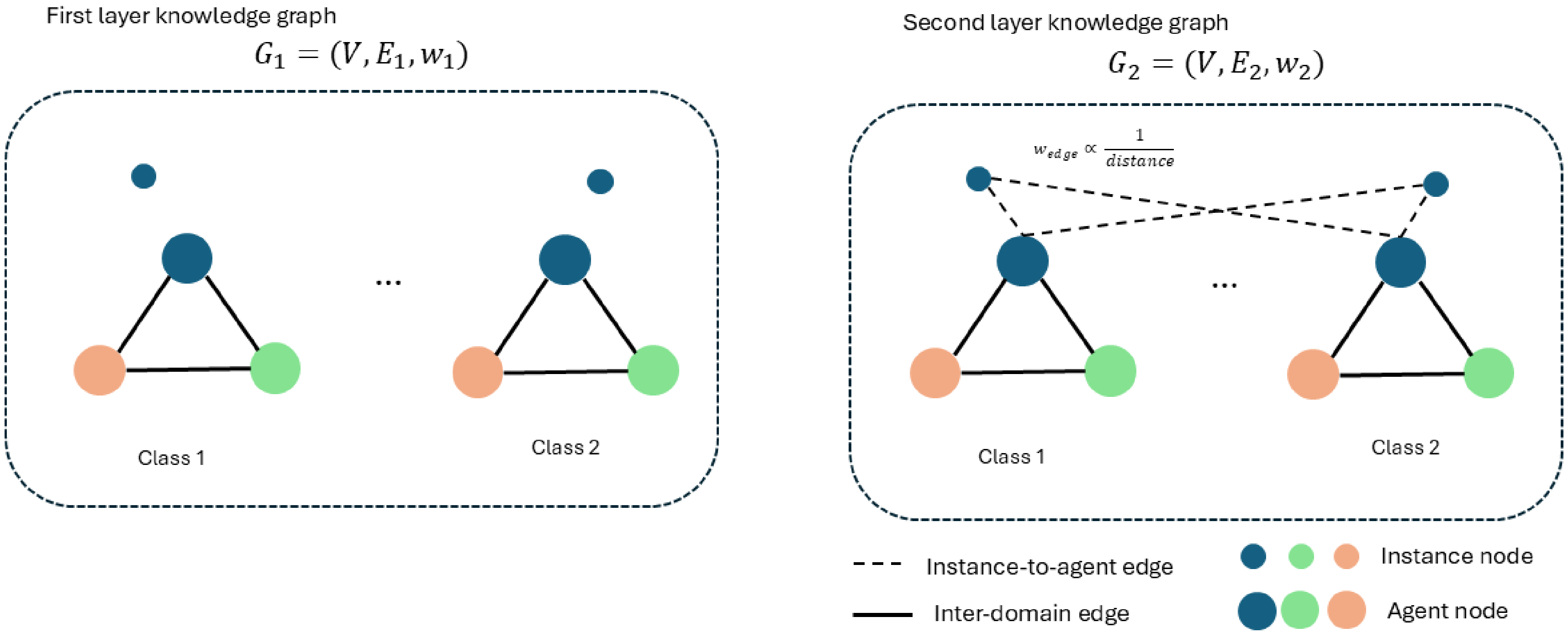
Knowledge graph (adjacency matrix) construction in MDIF Domain-level. Colors represent domains.

Domain-level agent node:

$$\boldsymbol{agt}^d = \sum_{i=1}^{N_d} \boldsymbol{w}_i^d \cdot \boldsymbol{z}_i^d$$

MDIF Class-Level

- Refinement of MDIF Domain-level with information integration at class level.
- An agent node represents a class in one domain.



Knowledge graph (adjacency matrix) construction in MDIF Class-level. Colors represent domains.

Class-level agent node:

$$\boldsymbol{agt}^{d,c} = \sum_{i=1}^{N_d} (y_i^d == c) \cdot \boldsymbol{w}_i^d \cdot \boldsymbol{z}_i^d$$

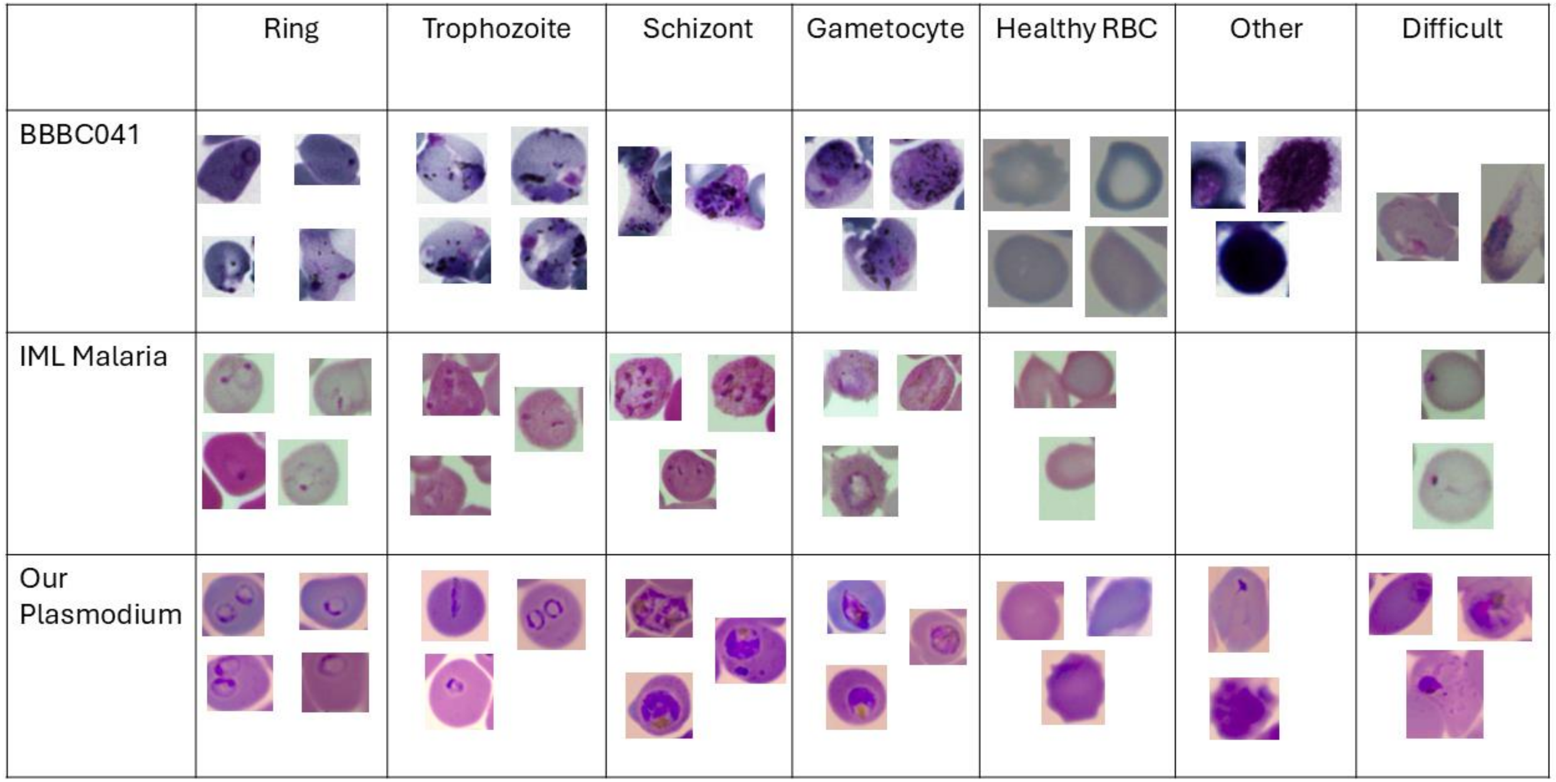
HUST

Experiment Settings

hust.edu.vn fb.com/dhbkhn

Datasets

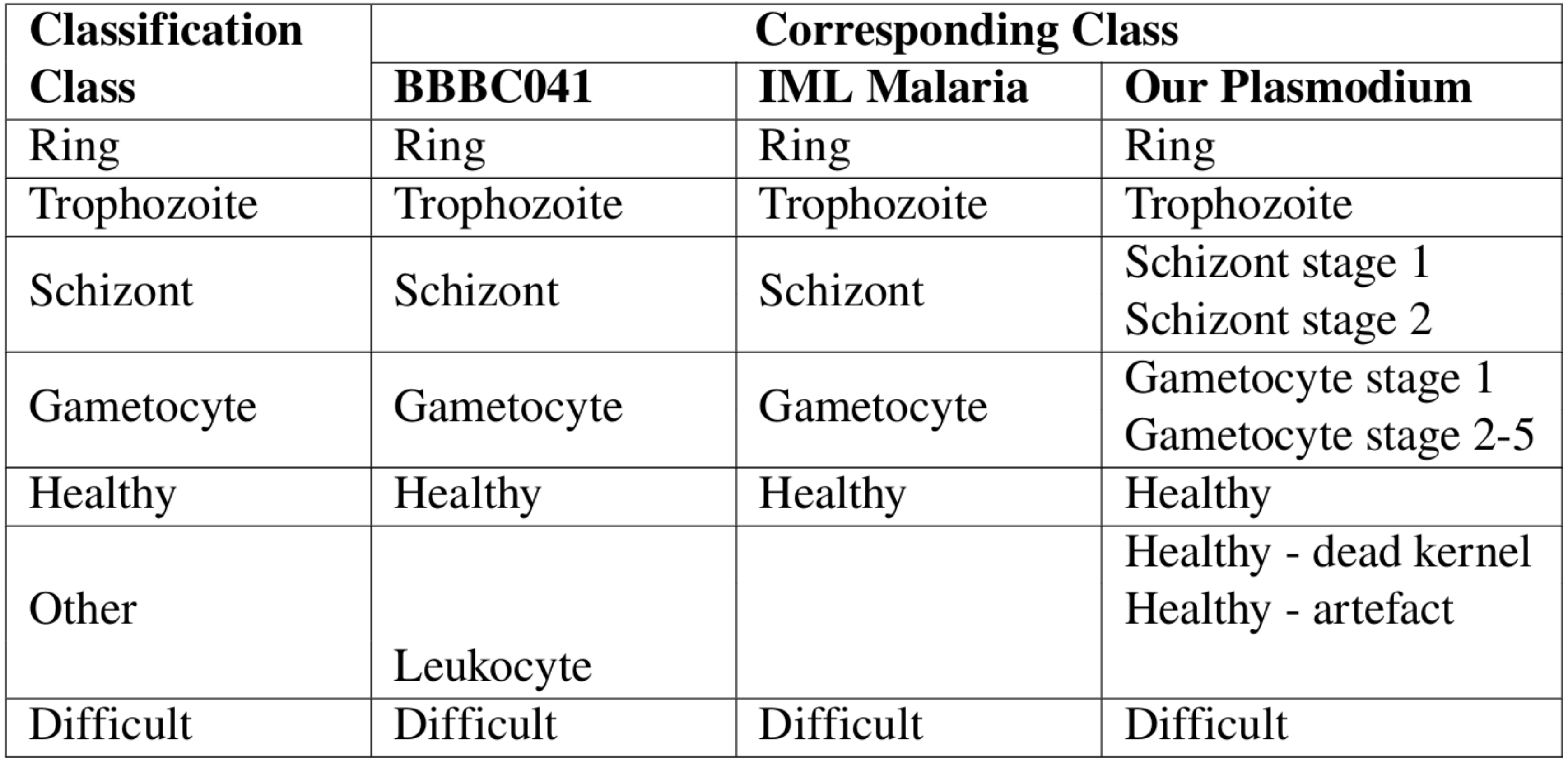
- BBBC041 [13], IML Malaria [18].
- Our Plasmodium (in progress of being published by CESPA).



Samples extracted from three datasets, grouped by class.

Datasets

- Classification classes (output of the model) and corresponding classes in three datasets.

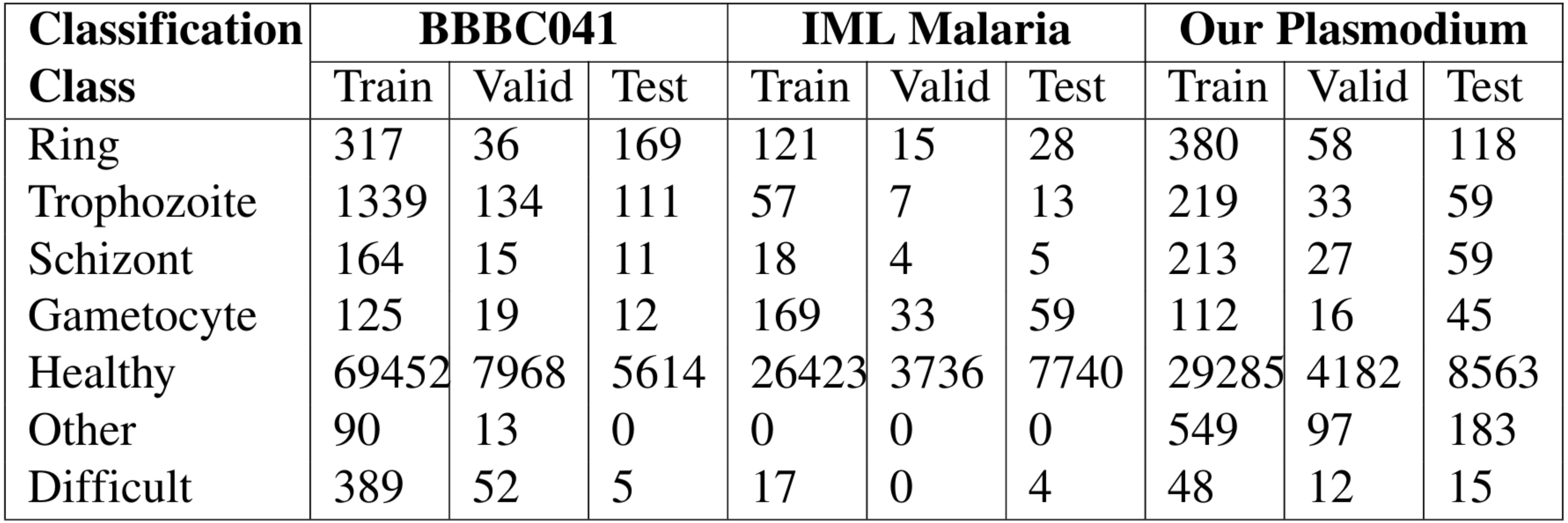


Classification Class	Corresponding Class		
	BBBC041	**IML Malaria**	**Our Plasmodium**
Ring	Ring	Ring	Ring
Trophozoite	Trophozoite	Trophozoite	Trophozoite
Schizont	Schizont	Schizont	Schizont stage 1 Schizont stage 2
Gametocyte	Gametocyte	Gametocyte	Gametocyte stage 1 Gametocyte stage 2-5
Healthy	Healthy	Healthy	Healthy
Other	Leukocyte		Healthy - dead kernel Healthy - artefact
Difficult	Difficult	Difficult	Difficult

Classification classes and their correspondents in three datasets.

Datasets

- Our Plasmodium, IML Malaria: random split with ratio 7/2/1 for train, test, and validation set.
- BBBC041: original train and test set (provided by authors), 1/10 of train set for validation purpose.

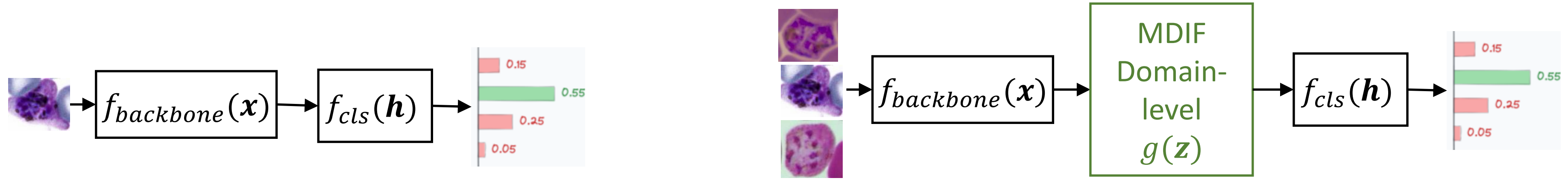


Classification Class	BBBC041			IML Malaria			Our Plasmodium		
	Train	Valid	Test	Train	Valid	Test	Train	Valid	Test
Ring	317	36	169	121	15	28	380	58	118
Trophozoite	1339	134	111	57	7	13	219	33	59
Schizont	164	15	11	18	4	5	213	27	59
Gametocyte	125	19	12	169	33	59	112	16	45
Healthy	69452	7968	5614	26423	3736	7740	29285	4182	8563
Other	90	13	0	0	0	0	549	97	183
Difficult	389	52	5	17	0	4	48	12	15

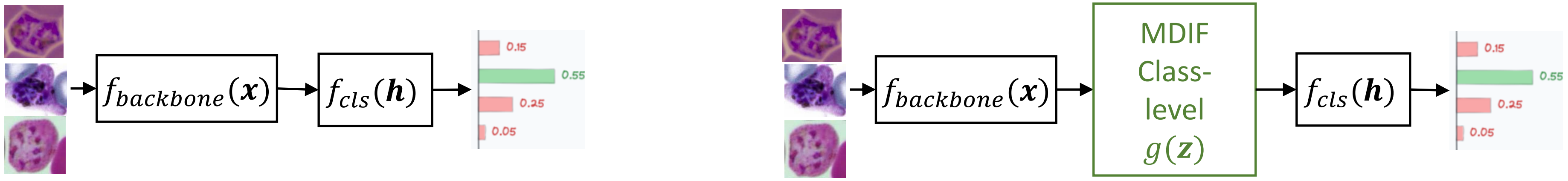
Number of instances by classification class for train, validation, and test set.

Experiment Design

- **Individual Training:** trained on individual dataset without MDIF.
- **Joint Training:** trained jointly on all datasets without MDIF.
- **MDIF Domain-level:** trained jointly on all datasets with MDIF Domain-level.
- **MDIF Class-level:** trained jointly on all datasets with MDIF Class-level.



Individual Training ***MDIF Domain-level***



Joint Training ***MDIF Class-level***

Illustration of experiment design.

Evaluation Strategy

- Evaluation metrics:
  - Accuracy
  - Life cycle micro-averaged recall $R_{LifeCycle}$
  - Life cycle micro-averaged precision $P_{LifeCycle}$
  - Life cycle micro-average F1 score $F1_{LifeCycle}$
- Model selection: best accuracy, best $R_{LifeCycle}$, and last epoch.

$$R_{LifeCycle} = \frac{\sum_{c\in\{R,T,S,G\}} TP_c}{\sum_{c\in\{R,T,S,G\}}(TP_c + FN_c)}$$

$$P_{LifeCycle} = \frac{\sum_{c\in\{R,T,S,G\}} TP_c}{\sum_{c\in\{R,T,S,G\}}(TP_c + FP_c)}$$

$$F1_{LifeCycle} = 2 \cdot \frac{R_{LifeCycle} \cdot P_{LifeCycle}}{R_{LifeCycle} + P_{LifeCycle}}$$

Implementation Details

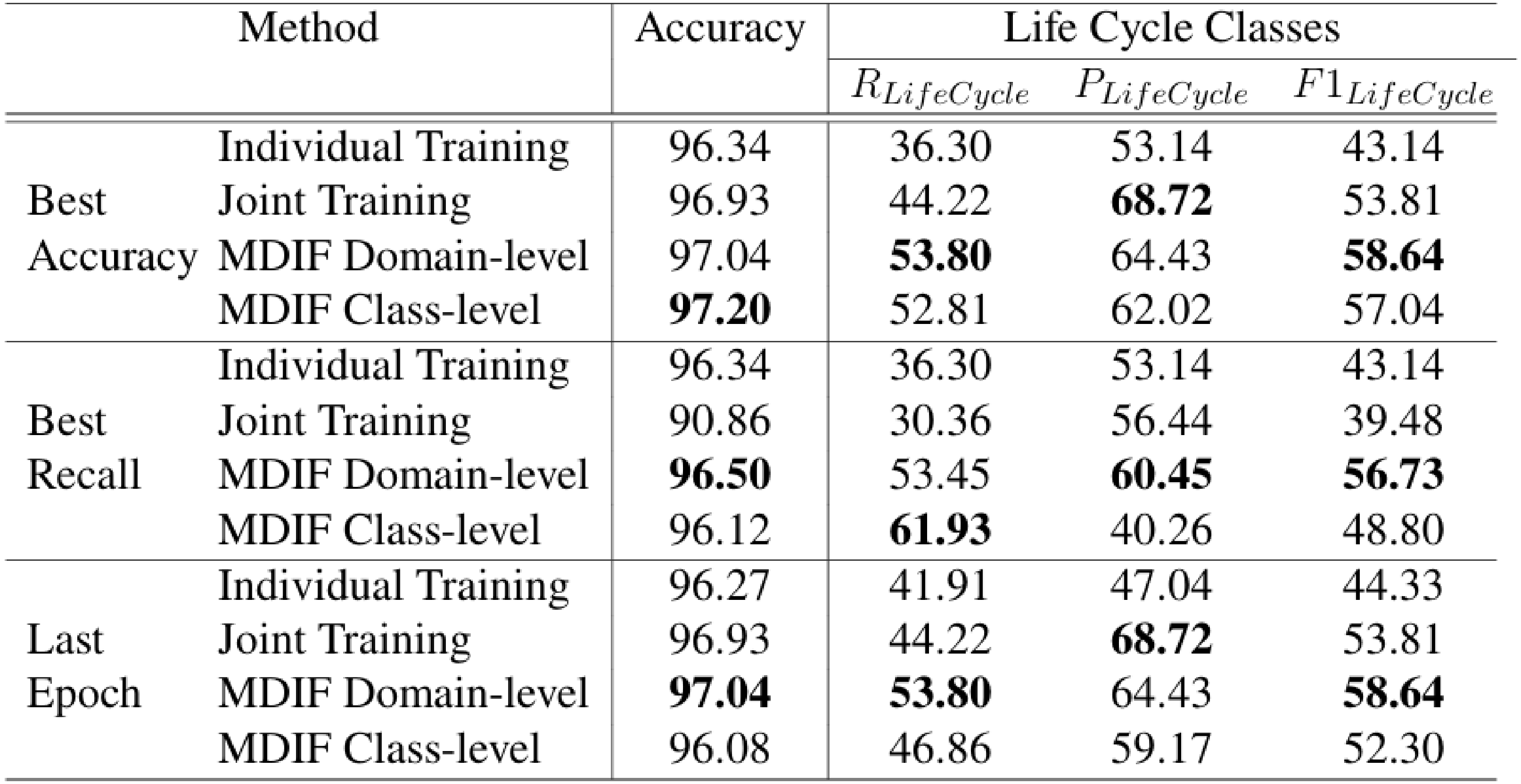
- Implementation with **MMPreTrain** framework.
- GCN are implemented with PyTorch Geometry.
- Experiments were carried out with NVIDIA P100 GPU on **Kaggle**.
- Backbone: **ResNet50**, pretrained on ImageNet.
- Training: **50 epochs** with **Adam** optimizer, learning rate 10^{-4}, reduced by 10^{-1} at 25th epoch.
- Evaluate on validation set every 5 epochs.

Source code is available at [github]()

Results and Analyses

Effectiveness of MDIF on BBBC041

- MDIF significantly improves performance on life cycle classes.

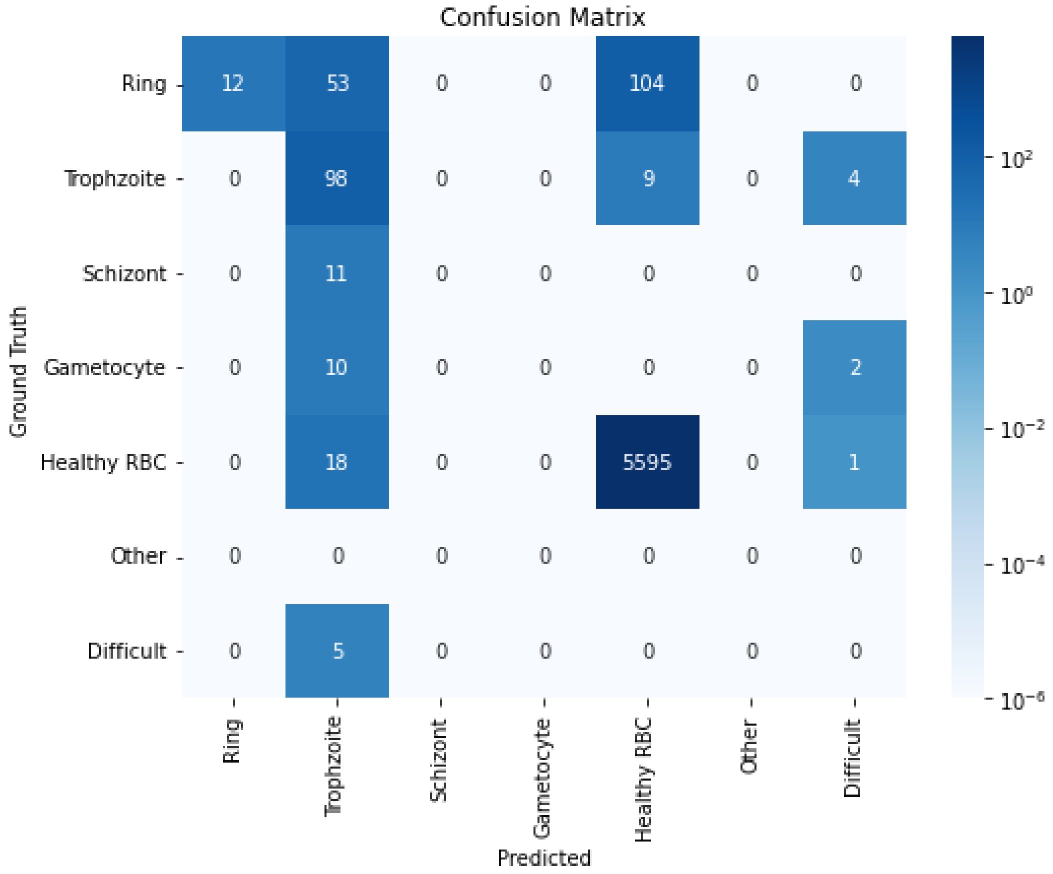


Method		Accuracy	Life Cycle Classes		
			$R_{LifeCycle}$	$P_{LifeCycle}$	$F1_{LifeCycle}$
Best Accuracy	Individual Training	96.34	36.30	53.14	43.14
	Joint Training	96.93	44.22	**68.72**	53.81
	MDIF Domain-level	97.04	**53.80**	64.43	**58.64**
	MDIF Class-level	**97.20**	52.81	62.02	57.04
Best Recall	Individual Training	96.34	36.30	53.14	43.14
	Joint Training	90.86	30.36	56.44	39.48
	MDIF Domain-level	**96.50**	53.45	**60.45**	**56.73**
	MDIF Class-level	96.12	**61.93**	40.26	48.80
Last Epoch	Individual Training	96.27	41.91	47.04	44.33
	Joint Training	96.93	44.22	**68.72**	53.81
	MDIF Domain-level	**97.04**	**53.80**	64.43	**58.64**
	MDIF Class-level	96.08	46.86	59.17	52.30

Table 5.1: Results on BBBC041.

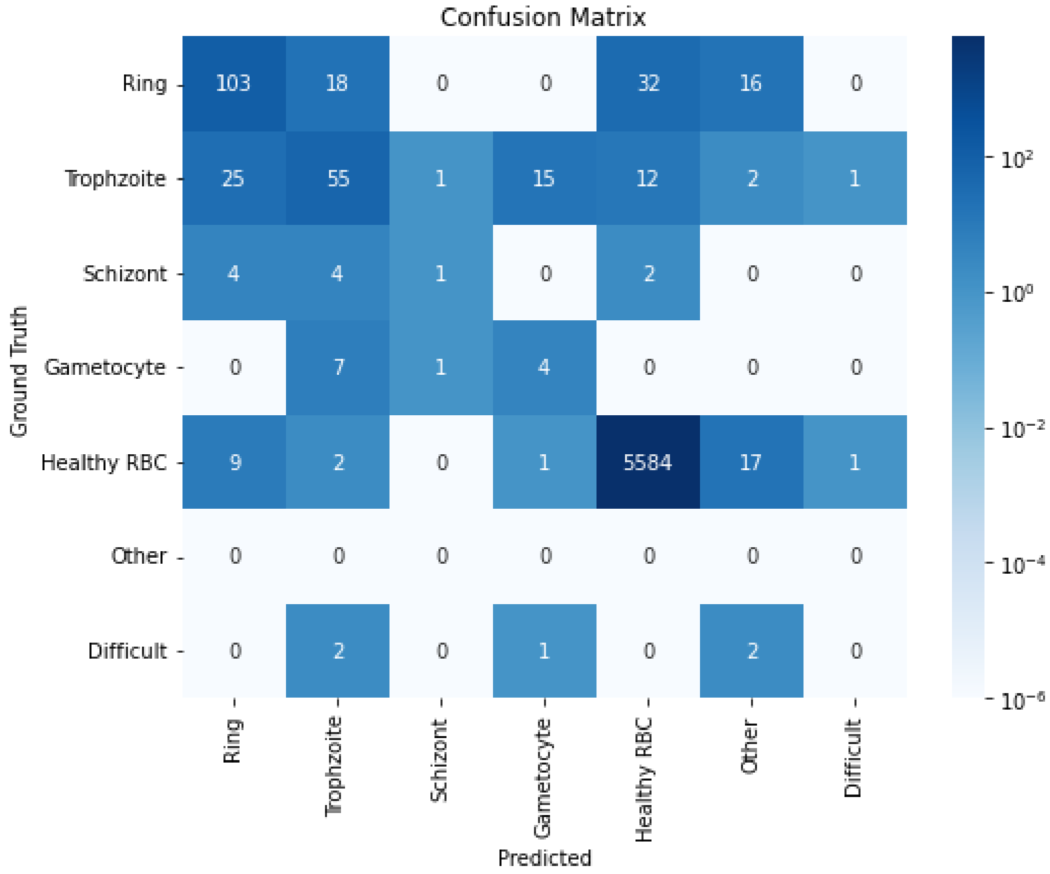
Effectiveness of MDIF on BBBC041

- Better separation among life cycle classes of MDIF Domain-level compared with Individual Training.



	Ring	Trophzoite	Schizont	Gametocyte	Healthy RBC	Other	Difficult
Ring	12	53	0	0	104	0	0
Trophzoite	0	98	0	0	9	0	4
Schizont	0	11	0	0	0	0	0
Gametocyte	0	10	0	0	0	0	2
Healthy RBC	0	18	0	0	5595	0	1
Other	0	0	0	0	0	0	0
Difficult	0	5	0	0	0	0	0

***Individual Training**. Best accuracy model.*

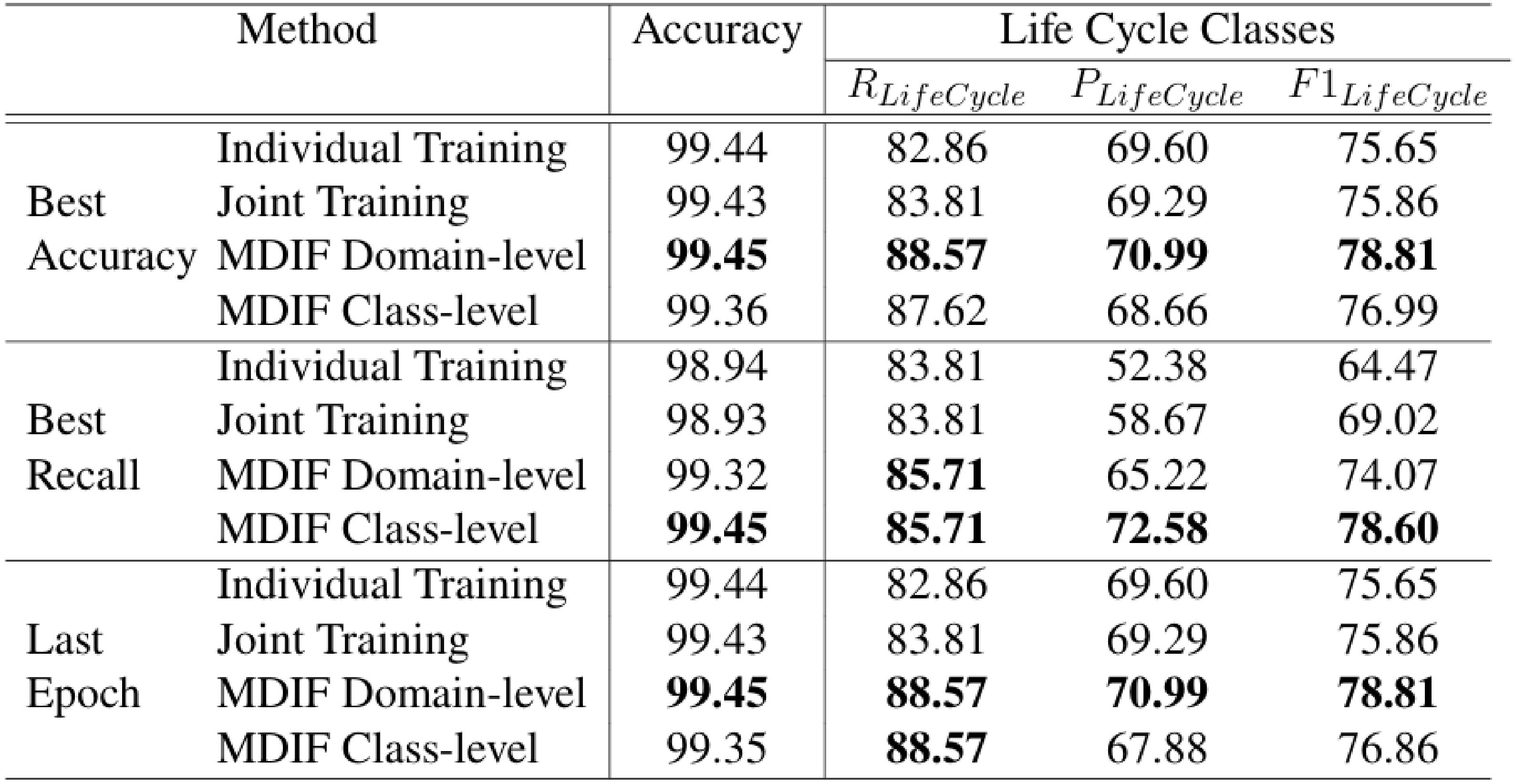


	Ring	Trophzoite	Schizont	Gametocyte	Healthy RBC	Other	Difficult
Ring	103	18	0	0	32	16	0
Trophzoite	25	55	1	15	12	2	1
Schizont	4	4	1	0	2	0	0
Gametocyte	0	7	1	4	0	0	0
Healthy RBC	9	2	0	1	5584	17	1
Other	0	0	0	0	0	0	0
Difficult	0	2	0	1	0	2	0

***MDIF Domain-level**. Best accuracy model.*

Effectiveness of MDIF on IML Malaria

- Either MDIF Domain-level or Class-level reports highest score on all evaluation metrics.

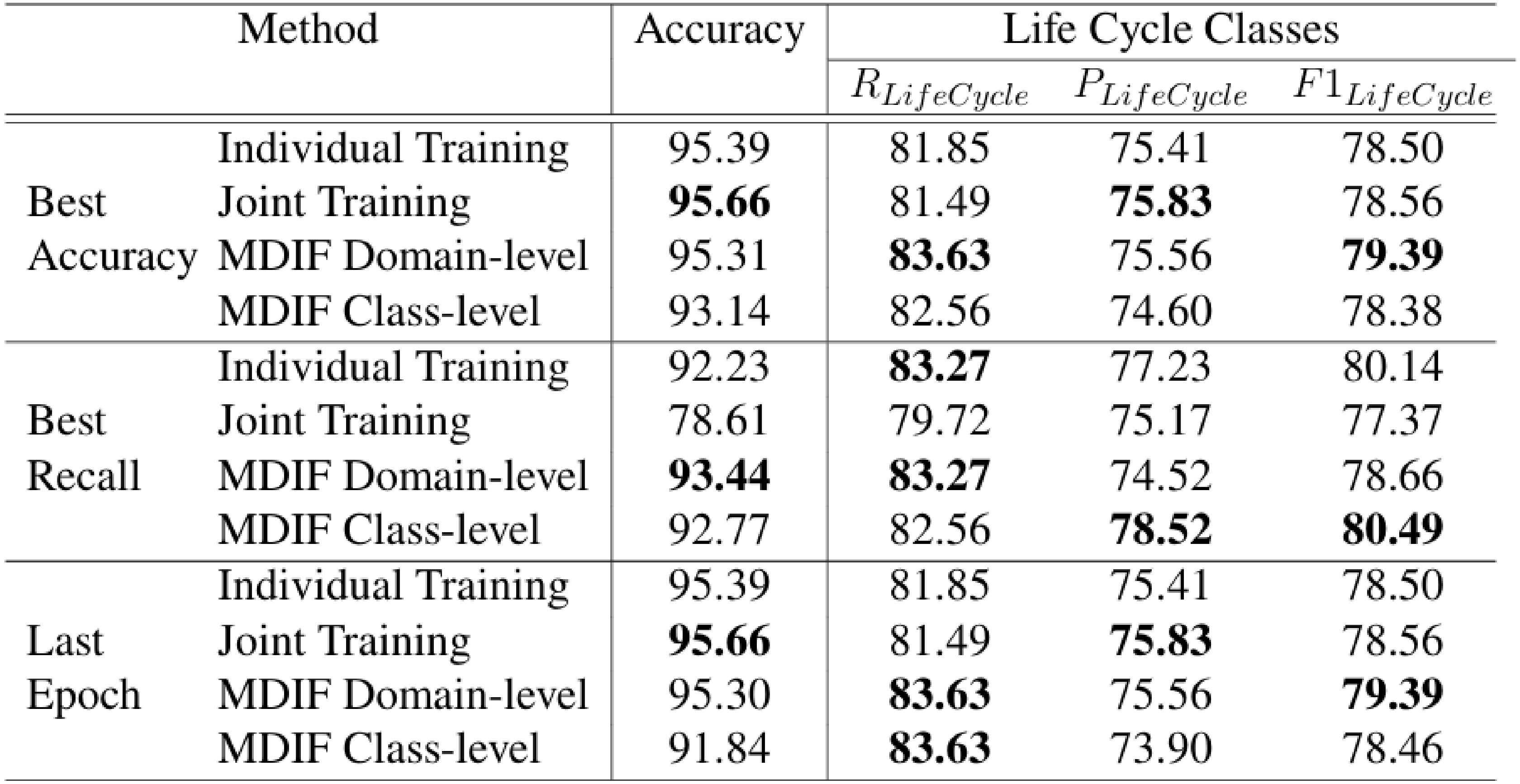


Method		Accuracy	Life Cycle Classes		
			$R_{LifeCycle}$	$P_{LifeCycle}$	$F1_{LifeCycle}$
Best Accuracy	Individual Training	99.44	82.86	69.60	75.65
	Joint Training	99.43	83.81	69.29	75.86
	MDIF Domain-level	**99.45**	**88.57**	**70.99**	**78.81**
	MDIF Class-level	99.36	87.62	68.66	76.99
Best Recall	Individual Training	98.94	83.81	52.38	64.47
	Joint Training	98.93	83.81	58.67	69.02
	MDIF Domain-level	99.32	**85.71**	65.22	74.07
	MDIF Class-level	**99.45**	**85.71**	**72.58**	**78.60**
Last Epoch	Individual Training	99.44	82.86	69.60	75.65
	Joint Training	99.43	83.81	69.29	75.86
	MDIF Domain-level	**99.45**	**88.57**	**70.99**	**78.81**
	MDIF Class-level	99.35	**88.57**	67.88	76.86

Table 5.2: Results on IML Malaria.

Effectiveness of MDIF on Our Plasmodium

- Slightly better performance with MDIF.



Method		Accuracy	Life Cycle Classes		
			$R_{LifeCycle}$	$P_{LifeCycle}$	$F1_{LifeCycle}$
Best Accuracy	Individual Training	95.39	81.85	75.41	78.50
	Joint Training	**95.66**	81.49	**75.83**	78.56
	MDIF Domain-level	95.31	**83.63**	75.56	**79.39**
	MDIF Class-level	93.14	82.56	74.60	78.38
Best Recall	Individual Training	92.23	**83.27**	77.23	80.14
	Joint Training	78.61	79.72	75.17	77.37
	MDIF Domain-level	**93.44**	**83.27**	74.52	78.66
	MDIF Class-level	92.77	82.56	**78.52**	**80.49**
Last Epoch	Individual Training	95.39	81.85	75.41	78.50
	Joint Training	**95.66**	81.49	**75.83**	78.56
	MDIF Domain-level	95.30	**83.63**	75.56	**79.39**
	MDIF Class-level	91.84	**83.63**	73.90	78.46

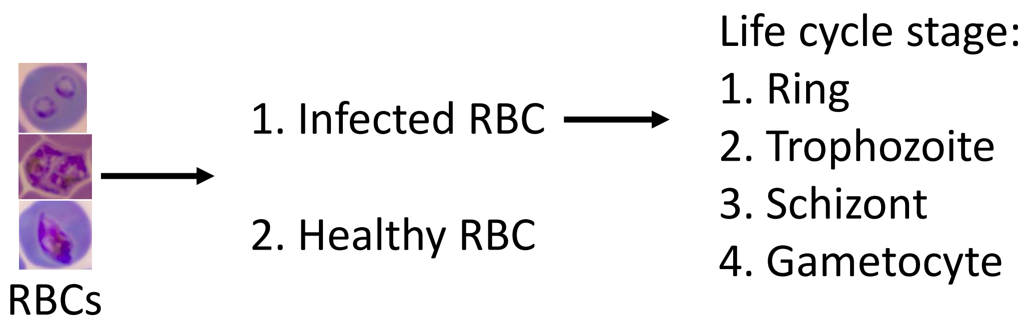
Table 5.3: Results on Our Plasmodium.

Attempt to Compare Our Results with Existing Studies

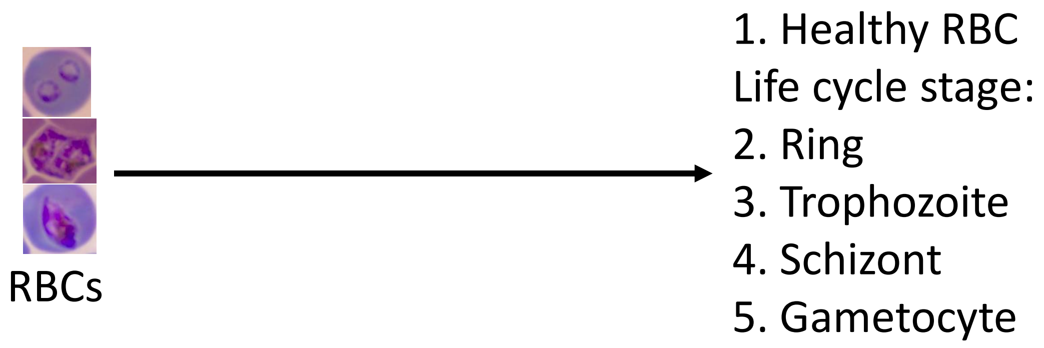
Only two studies on life cycle classification from RBCs:

- **Different pipeline** (a): Li et al. [10].
- Similar pipeline (b), but **different data split:** Auraujo et al. [11] did not use original train/test split provided by BBBC041, too many samples of minor classes for testing, only a few for training.

→ **Hard to draw an absolute comparison** on life cycle classification.



(a) Two-stage classification

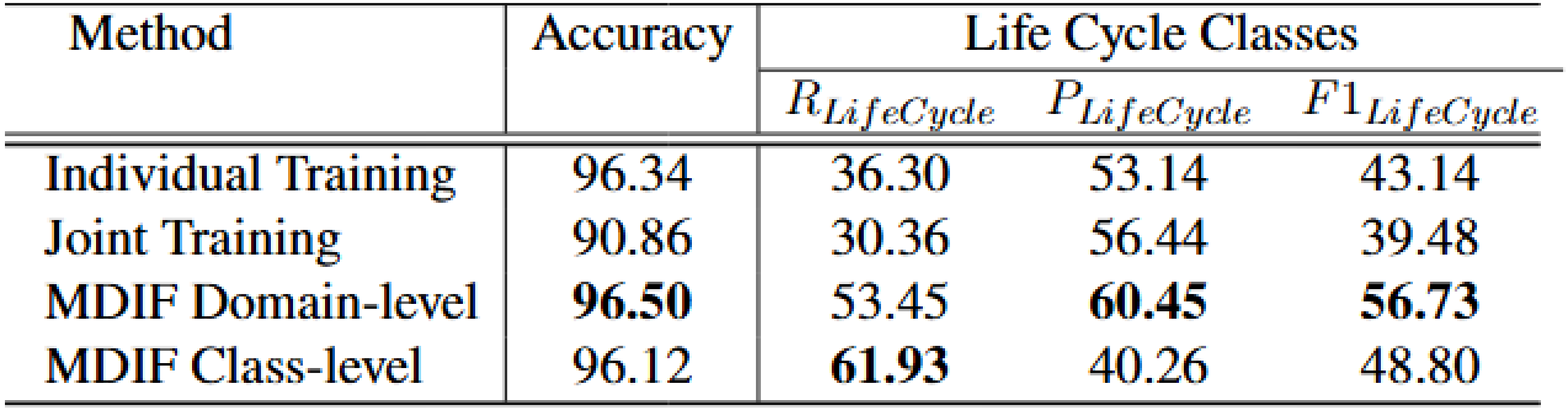


(b) One-stage classification

Two different pipeline for life cycle classification from RBC.

More Data, Better Result?

- Individual Training and Joint Training: **“Physically” adding more data does not always ensure better model.**
- MDIF achieves higher performance compared with Individual Training and Joint Training.

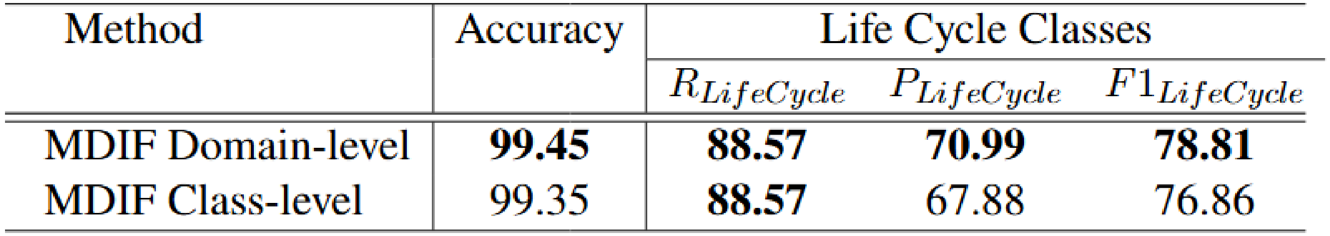


Method	Accuracy	Life Cycle Classes		
		$R_{LifeCycle}$	$P_{LifeCycle}$	$F1_{LifeCycle}$
Individual Training	96.34	36.30	53.14	43.14
Joint Training	90.86	30.36	56.44	39.48
MDIF Domain-level	**96.50**	53.45	**60.45**	**56.73**
MDIF Class-level	96.12	**61.93**	40.26	48.80

Result on BBBC041. Best recall selection.

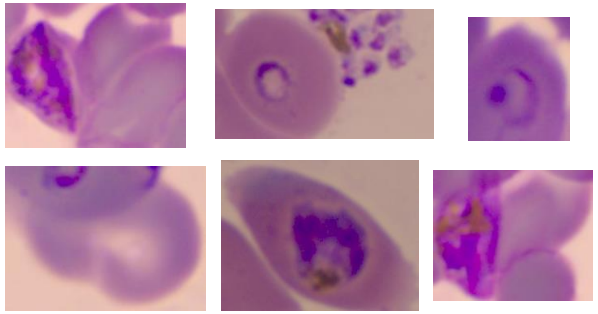
Negative Affect of Poor Annotation on MDIF Class-level?

- MDIF Class-level was expected to be refinement of Domain-level.
- Might be due to **annotation mistakes**.

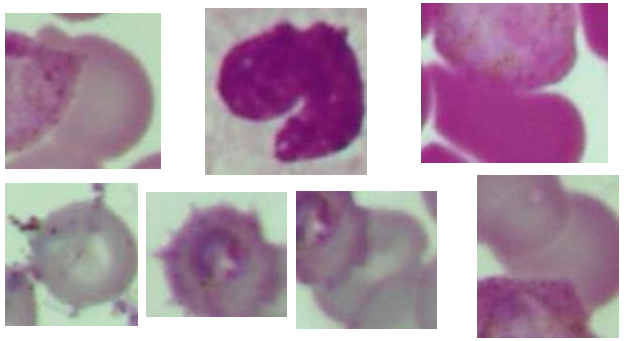


Method	Accuracy	Life Cycle Classes		
		$R_{LifeCycle}$	$P_{LifeCycle}$	$F1_{LifeCycle}$
MDIF Domain-level	**99.45**	**88.57**	**70.99**	**78.81**
MDIF Class-level	99.35	**88.57**	67.88	76.86

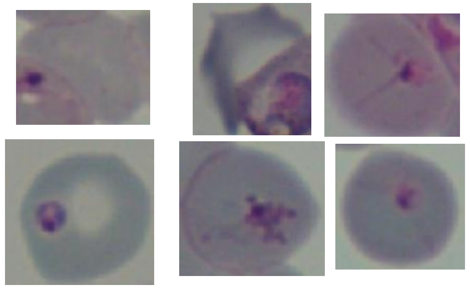
Results on IML Malaria. Last epoch selection.



Healthy RBC samples with ***wrong annotation?*** *Our Plasmodium*



Healthy RBC samples with ***wrong annotation?*** *IML Malaria*



Healthy RBC samples with ***wrong annotation?*** *BBBC041*

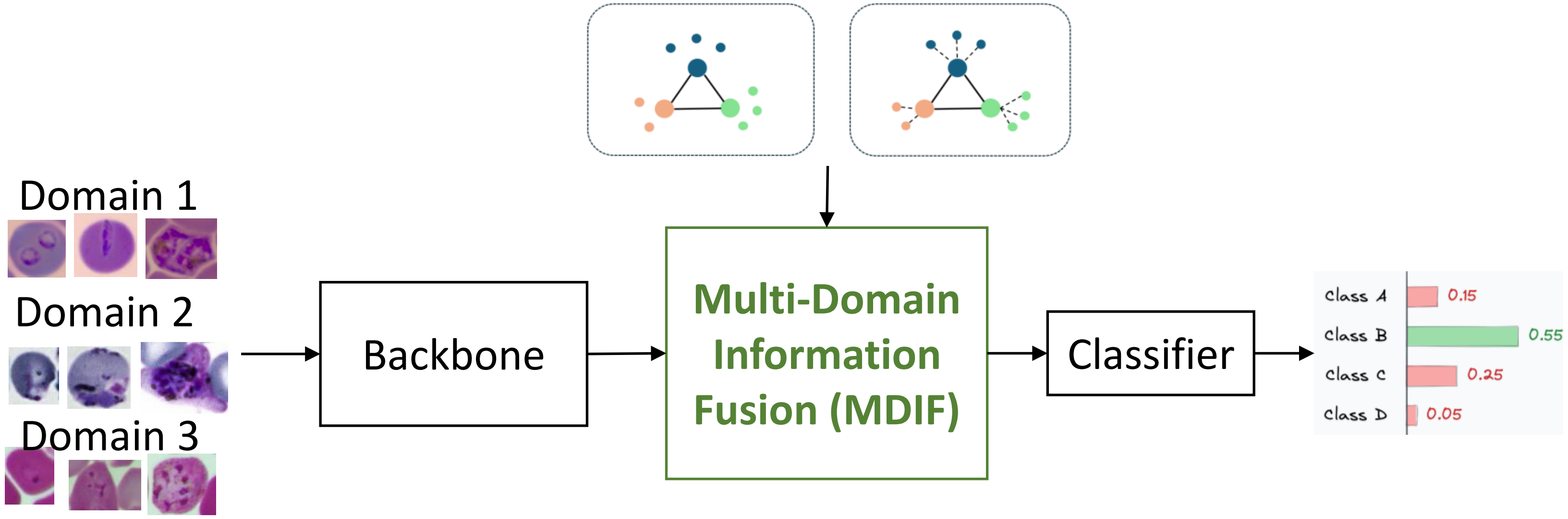
HUST

hust.edu.vn fb.com/dhbkhn

Conclusion and Future Works

Conclusion

- An automatic classification of plasmodium life cycle development.
- Address **data imbalance** with **data enrichment** from multiple domains enrichment.
- **Bridge domain gap** by Multi-Domain Information Fusion at feature level.



Domain 1

Domain 2

Domain 3

Backbone

Multi-Domain Information Fusion (MDIF)

Classifier

Life cycle development classification framework with MDIF.

Future Works

- Input-level alignment methods:
  - Domain-specific normalization
  - Style transfer
- Two-stage classification to reduce impact of data imbalance.

References

[1] World Health Organization, *Malaria*, https://www.who.int/news room/fact-sheets/detail/malaria, Accessed: December 28, 2024, 2024.

[2] J. Sachs and P. Malaney, “The economic and social burden of malaria,” *Nature*, vol. 415, no. 6872, pp. 680–685, 2002.

[3] K. E. Halliday, S. S. Witek-McManus, C. Opondo, et al., “Impact of schoolbased malaria case management on school attendance, health and education outcomes: A cluster randomised trial in southern malawi,” *BMJ global health*, vol. 5, no. 1, e001666, 2020.

[4] Z. Jan, A. Khan, M. Sajjad, K. Muhammad, S. Rho, and I. Mehmood, “A review on automated diagnosis of malaria parasite in microscopic blood smears images,” *Multimedia Tools and Applications*, vol. 77, pp. 9801–9826, 2018.

[5] X.-z. Su and J. Wu, “Zoonotic transmissions and host switches of malaria parasites,” *Zoonoses (Burlington, Mass.)*, vol. 1, no. 1, 2021.

[6] S. Rajaraman, S. K. Antani, M. Poostchi, et al., “Pre-trained convolutional neural networks as feature extractors toward improved malaria parasite detection in thin blood smear images,” *PeerJ*, vol. 6, e4568, 2018.

[7] F. Abdurahman, K. A. Fante, and M. Aliy, “Malaria parasite detection in thick blood smear microscopic images using modified yolov3 and yolov4 models,” *BMC bioinformatics*, vol. 22, pp. 1–17, 2021.

References

[8] R. Nakasi, E. Mwebaze, A. Zawedde, J. Tusubira, B. Akera, and G. Maiga, “A new approach for microscopic diagnosis of malaria parasites in thick blood smears using pre-trained deep learning models,” *SN Applied Sciences*, vol. 2, pp. 1–7, 2020.

[9] J. Hung and A. Carpenter, “Applying faster r-cnn for object detection on malaria images,” in *Proceedings of the IEEE conference on computer vision and pattern recognition workshops, 2017*, pp. 56–61.

[10] S. Li, Z. Du, X. Meng, and Y. Zhang, “Multi-stage malaria parasite recognition by deep learning,” *GigaScience*, vol. 10, no. 6, giab040, 2021.

[11] F.Araujo,N.Colares, U. Carvalho, C. F. Costa Filho, and M.G.Costa, “Plasmodium life cycle-stage classification on thick blood smear microscopy images using deep learning: A contribution to malaria diagnosis,” in *2023 19th International Symposium on Medical Information Processing and Analysis (SIPAIM), IEEE, 2023*, pp. 1–4.

[12] Q. A. Arshad, M. Ali, S.-u. Hassan, et al., “A dataset and benchmark for malaria life-cycle classification in thin blood smear images,” *Neural Computing and Applications*, vol. 34, no. 6, pp. 4473–4485, 2022.

HUST

hust.edu.vn fb.com/dhbkhn

Acknowledgement

Special thanks to CESPA project members, **Assoc.Prof. Muriel Visani** and **Asst.Prof. Thierry Urruty**, for constructive feedback and suggestions.

HUST

hust.edu.vn fb.com/dhbkhn

THANK YOU !