CMD-GCN: Categorical Multi-Domain Graph Convolutional Network for *Plasmodium* Development Stage Recognition

Quoc Khanh Tran[1,2][0009−0006−2367−7761], Muriel Visani[3,4][0000−0001−7513−4749], Thierry Urruty[5][0000−0003−1339−1920], Océane Delandre[6,7,8,9], and Thi-Oanh Nguyen[1][0000−0002−6166−2011] ⋆

[1] School of Information and Communications Technology (SoICT), Hanoi University of Science and Technology (HUST), Hanoi, Vietnam
[2] Center for Environmental Intelligence (CEI), VinUniversity, Hanoi, Vietnam
[3] French Military Center for Epidemiology and Public Health (CESPA), Marseille, France
[4] Laboratoire L3i, La Rochelle Université, France
[5] Université de Poitiers, Univ. Limoges, CNRS, XLIM, Poitiers, France
[6] Unité Parasitologie et Entomologie, Département Risques Vectoriels, Institut de Recherche Biomédicale des Armées (IRBA), Marseille, France
[7] Aix Marseille Univ, SSA, APHM, RITMES, Marseille, France
[8] IHU Méditerranée Infection, Marseille, France
[9] Centre National de Référence du Paludisme, Marseille, France

Abstract. Malaria is a disease caused by the *Plasmodium* parasite. Automatic recognition of *Plasmodium* infection and its development stages in human blood can significantly reduce the cost of manual diagnosis and clinical trials. However, insufficient annotated samples per development stage and extreme data imbalance pose major challenges. When acquiring additional data is impractical, multi-domain learning provides an effective way to exploit heterogeneous data sources and improve model generalization. In this paper, we propose a two-step method with multi-domain learning for efficient recognition of *Plasmodium* development stages. Red blood cells are first detected and then classified as either uninfected (unparasitized) or belonging to a specific development stage. For the classification step, we introduce a new supervised multi-domain learning method, namely Categorical Multi-Domain Graph Convolutional Network (CMD-GCN), to reduce the effects of data scarcity for some rarely observed stages while mitigating domain discrepancies and enhancing feature representations. Moreover, CMD-GCN propagates information across data features at the fine-grained category-aware level, which helps in alleviating data imbalance. Experiments are conducted on three datasets, namely BBBC041-v1, IML Malaria, and RecoPlasmodiumV1. The experimental results demonstrate that CMD-GCN significantly improves the performance of end-to-end *Plasmodium* development stage recognition across all evaluated datasets.

⋆ Corresponding author: oanh.nguyenthi@hust.edu.vn

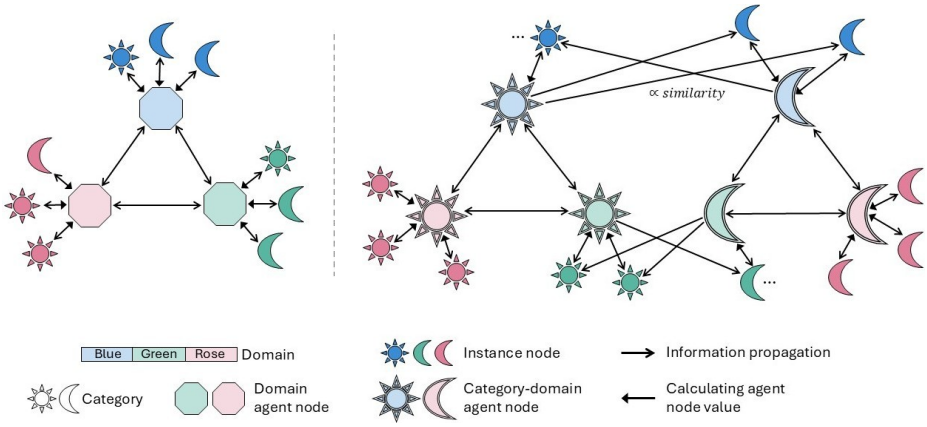
Keywords: Malaria disease · *Plasmodium* development stage recognition · Multi-domain learning · Domain adaptation · Graph Convolutional Network · Categorical Multi-Domain Graph Convolutional Network

1 Introduction

Malaria disease has strong impacts on public health, societies, and economics [1], [2], [3]. *Plasmodium* parasites are transmitted to humans through the bite of infected mosquitoes and undergo several development stages in the human liver and blood. Information about the development stages in human blood is crucial for the assessment of newly developed medications within the framework of clinical trials. To reduce the time and resources required for manual microscopic examination, we propose a deep learning method to automatically recognize *Plasmodium* development stages from thin blood smear images. Red blood cells (RBCs) are detected and then classified into the five classes that are necessary to estimate per-stage parasitaemia (percentage of infected RBCs per development stage), namely "Healthy", "Ring", "Trophozoite", "Schizont", and "Gametocyte". For examples of these five classes from different datasets (showcasing different species), please refer to Fig. 3. However, this classification problem is difficult, due to extreme data imbalance (the vast majority of RBCs are healthy), low inter-class variability among parasitized classes, large intra-class variability (some development stages consist of visually dissimilar sub-stages), and scarcity of labeled data (especially for some rarely observed stages, *e.g.*, gametocytes). To improve the model's ability to distinguish development stages despite data scarcity, we utilize *multi-domain learning* to make full use of the rich and diverse data from multiple domains[1].

A challenge in multi-domain learning is domain discrepancy. Different domains usually have different characteristics, *e.g.* in our case: staining agents, magnification scales, or even *Plasmodium* species. Thus, simply training a single classifier from multiple domains may not improve performance. It may even worsen it, because a classifier might sacrifice performance on the domains with the fewest number of samples [5] or suffer from domain conflict [6]. Domain adaptation, a closely related field to multi-domain learning, provides a potential solution to reduce the gap between source and target domains. In a recent study on unsupervised domain adaptation for person re-identification, Bai *et al.* [7] proposed an efficient method, Multi-Domain Information Fusion (MDIF), which fuses information from multiple domains using a graph convolutional network. Similar to many domain adaptation methods, the main objective of this work is to reduce the discrepancy between source and target domains. In contrast, our objective is not only to *mitigate domain discrepancy*, but also to *enhance feature representations* across all domains. Inspired by MDIF, we propose a **C**ategorical **M**ulti-**D**omain **G**raph **C**onvolutional **N**etwork (CMD-GCN), which transfers information not only at the domain level, but also exploits the rich

[1] A domain commonly refers to a dataset where samples follow the same underlying data distribution [4].



(a) MDIF [7]: Each agent node represents a domain and is computed from samples of all categories, leading to limited representativeness.

(b) Our CMD-GCN: Agent nodes are defined per category within each domain, enabling fine-grained and more representative information propagation.

Fig. 1: Information propagation mechanism in (a) MDIF [7] and (b) our proposed CMD-GCN, which operates at a more advanced level. Colors denote domains, sun/moon shapes represent different categories. For clearer illustration, instance–agent propagation across categories are shown within only two domains (blue and green). More information about CMD-GCN's information propagation mechanism is given in Section 3.1. Best viewed in color.

category information [8], [9], [10], which is not taken into account by MDIF. By incorporating category-aware relationships in its graph convolutional network (GCN), CMD-GCN refines high-level feature representations while simultaneously aligning domain distributions. Fig. 1 illustrates CMD-GCN's propagation mechanism, in comparison with MDIF's.

Our contributions are listed as follows: 1) A two-step *Plasmodium* development stage recognition pipeline is proposed, first detecting RBCs and then classifying them as healthy or one of four development stages. 2) To alleviate data scarcity in the classification stage, we introduce a new supervised multi-domain learning method, namely Categorical Multi-Domain Graph Convolutional Network (CMD-GCN), that exploits both domain and category labels to propagate information across data features, thereby simultaneously reducing domain discrepancy and enhancing high-level feature representations. 3) Extensive experiments are conducted on three datasets to demonstrate the effectiveness of the proposed end-to-end pipeline incorporating CMD-GCN.

2 Related Work

***Plasmodium* development stage recognition.** Recent studies use different recognition pipeline designs. Although some methods carry it out in a single step by formulating it as a multi-class object detection task on blood smear images [11], [12], [13], others successively perform detection and classification in two steps [11], [14], [15], [16] or three steps [14], [17]. Two-step approaches perform either infected RBC detection followed by classification, or, like ours, perform

RBC detection followed by classification, including the "Healthy" class. Three-step approaches usually consist of RBC detection, infected (or uninfected) RBC classification, and (for infected samples) development stage classification. The best performing method on the public BBBC041-v1 dataset [18], among those presented in Hung *et al.* [11], is a two-step approach that uses Faster R-CNN [19] for uninfected cell detection and AlexNet [20] for development stage classification (for negative samples). Zedda et *al.* [16] reported results on the public IML Malaria dataset [14], achieving their best performance by using YOLOv8m [21] to detect infected RBCs and a ViT-large model [22] to classify them into four developmental stages. Following a similar two-step strategy, the authors proposed, in subsequent work [13], a novel YOLOv8-based architecture to further improve the infected RBC detection stage. In addition to the whole pipeline, several studies focus exclusively on the classification stage, either on infected RBC development stage classification only [23], or for both development stage and healthy RBC classification [24], [25].

Lack of standard benchmark for *Plasmodium* recognition. In the studies on *Plasmodium* development stage recognition, there is no standardized evaluation protocol. Existing studies assess their methods using diverse protocols, that are rarely representative of real-life scenarios. Specifically, some studies report results for different pipeline components separately [14]; disregard samples labeled as difficult by experts [11]; or exclude incorrectly detected or misdetected samples from the initial detection stage when evaluating the subsequent classification step [11], [13], [16], which results in overly optimistic performance evaluations, compared to real-world applications. In addition, several works rely on private datasets [12], [17], [26] or augment the data with self-captured images [24]. Even when using public datasets such as BBBC041 and IML Malaria, fair comparison is further complicated by the lack of commonly adopted train/test splits. Indeed, while BBBC041 provides predefined splits, they are not consistently used, and IML Malaria does not offer standardized partitions.

Data challenges and multi-domain learning in *Plasmodium* recognition. The insufficient number of samples per development stage and severe data imbalance are major challenges in *Plasmodium* development stage recognition. While various techniques have been applied to mitigate the data imbalance, such as utilizing multi-stage recognition pipelines [14], [26], data augmentation [13], [16], [26], and under-sampling/removing some uninfected RBCs for training [17], [25], the data scarcity issue has received limited attention. To the best of our knowledge, the study of Li *et al.* [24] is the only one that tackles this issue by incorporating an additional dataset for training. However, the additional data used in their work is unlabeled with respect to development stages, resulting in under-utilization of available labeled information. Data scarcity can be more effectively addressed by exploiting labeled data from multiple sources, commonly referred to as multi-domain learning.

[2] Although the dataset is not explicitly specified in the paper, the images in BBBC041-v1, provided by Jane Hung (first author of [11]), appear to have been used.

Multi-domain learning leverages diverse and complementary data from multiple datasets to improve model generalization. It is particularly effective for complex tasks that require costly and labor-intensive annotation. Recent computer vision applications of multi-domain learning include object detection [27], [28], segmentation [29], [30], action recognition [31], and crowd counting [5]. Common approaches in these works are vision-language alignment [27], [29], [30], domain-specific network partitioning [5], [28], and training loss regularization [31]. While vision-language methods rely on additional textual descriptions that are not always available, domain-specific network partitioning lacks scalability, as the number of per-domain components grow proportionally to the number of domains. Training loss regularization operates through optimization objective, consequently relies on the model to implicitly discover relationships among domains and samples. With the shared goal of mitigating domain discrepancy and enabling cross-domain knowledge transfer, *domain adaptation* employs techniques such as maximum mean discrepancy [32], [33], adversarial training [34], domain normalization [7], [35], and GCN-based alignment [7], [8]. Compared with other methods, GCN-based methods offer a structured way to model inter- and intra-domain relationships of the samples, which can be advantageous.

In this study, we propose a multi-domain learning method that – being supervised – captures relational semantics across both domains and categories. This method, named CMD-GCN, is designed in a way that facilitates domain distribution alignment while enhancing category-level feature representations. In addition, by grouping samples by category for information propagation in the GCN, CMD-GCN mitigates the bias toward the majority class(es), which would otherwise account for most information in the GCN's agent nodes.

3 Methodology

3.1 Overview

To recognize *Plasmodium* development stages from blood smear images, we propose the two-step pipeline illustrated in Fig. 2. In the first step, a detector $\mathcal{D}(\cdot)$ identifies all RBCs. In the second step, each detected RBC is classified into one of the five classes required to estimate per-stage parasitaemia: "Healthy" (H), "Ring" (R), "Trophozoite" (T), "Schizont" (S), and "Gametocyte" (G). In general, the background in blood smear images is relatively simple, making detection straightforward, while classification is more challenging due to the issues discussed in Section 1. Thus, we mainly focus on the classification stage. Multiple datasets are exploited to train a single classifier, with our proposed **CMD-GCN** placed on top of the backbone to perform distribution alignment and feature enhancement across domains.

Specifically, RBC images X from all domains are first embedded into a feature space using a feature extractor $\mathcal{F}(\cdot)$, resulting in a set of feature representations Z. These features are then enhanced by propagating information through CMD-GCN $\mathcal{G}(\cdot)$, before being passed to a classifier $\mathcal{C}(\cdot)$. The core idea is to facilitate information exchange between groups of features using a GCN. Data features

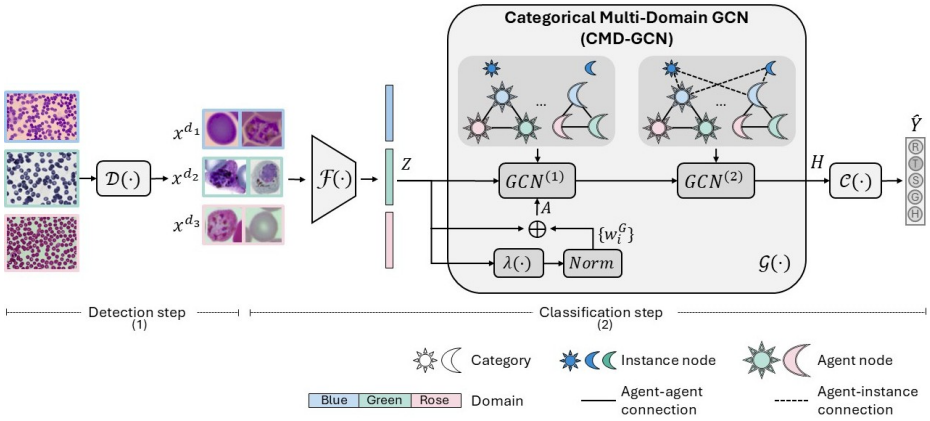


Fig. 2: Our proposed two-step pipeline for *Plasmodium* development stage recognition. (1) In the detection step, all RBCs are detected by a detector $\mathcal{D}(\cdot)$. (2) In the classification step, the model is trained using data from multiple domains, where CMD-GCN $\mathcal{G}(\cdot)$ is placed on top of the backbone (feature extractor) $\mathcal{F}(\cdot)$ for distribution alignment and feature enhancement before being classified by the classifier $\mathcal{C}(\cdot)$. For clarity, only instance nodes from one domain are shown. Best viewed in color.

are divided into *groups*, each represented by an *agent node*. These agent nodes communicate with each other to share global information through the first GCN layer and propagate this information to their associated instance nodes in the second one.

3.2 Agent Node

Suppose we have D domains, where each domain d consists of N_d images $\boldsymbol{x}_i^d$ and their corresponding labels $y_i^d \in [1, C]$. The sample set of domain d is denoted as $\{(\boldsymbol{x}_i^d, y_i^d) \mid i \in [1, N_d]\}$ with $d \in [1, D]$. An *agent node* $\boldsymbol{a}^G$ is determined based on the features of instances in G, $\{\boldsymbol{z}_i = \mathcal{F}(\boldsymbol{x}_i), \forall \boldsymbol{x}_i \in G\}$. Since each instance node contributes differently to the representation, a contribution coefficient $\boldsymbol{w}_i^G$ is assigned to each instance $\boldsymbol{x}_i$ in G for the calculation of its corresponding agent node $\boldsymbol{a}^G$: $\boldsymbol{w}_i^G = \lambda(\mathcal{F}(\boldsymbol{x}_i)) / \sum_{\boldsymbol{x}_j \in G} \lambda(\mathcal{F}(\boldsymbol{x}_j))$, in which $\lambda(.)$ is a fully connected layer. The calculation of an agent node is described in Eq. 1, where $\cdot$ denotes element-wise multiplication.

$$\boldsymbol{a}^G = \sum_{x_i \in G} \boldsymbol{w}_i^G \cdot \mathcal{F}(\boldsymbol{x}_i) \tag{1}$$

Unlike MDIF [7], where each agent node represents an entire domain and is therefore prone to domination by the majority class(es), in our CMD-GCN, agent nodes are defined at the category level within each domain. Specifically, a group G consists of all instances belonging to class c in domain d, and its corresponding agent node is denoted as $\boldsymbol{a}^{d,c}$. By maintaining distinct representations for each category within a domain, CMD-GCN preserves category-specific information and prevents dominant classes from overwhelming minority classes,

thereby facilitating more accurate information propagation. We denote A as the set of all agent nodes across domains, $A = \{\boldsymbol{a}^{d,c} \mid c \in [1, C], d \in [1, D]\}$, and A^d as the set of agent nodes in domain d, $A^d = \{\boldsymbol{a}^{d,c} \mid c \in [1, C]\}$. For training stability, agent node values are updated using an exponential moving average and are stored for subsequent inference.

3.3 Graph Construction

The graph topology defines how the information is propagated and integrated across nodes. The proposed graph contains two types of nodes: *agent nodes* A (described in Section 3.2), and *instance nodes* Z, which represent the feature embeddings of individual instances. The construction of the relational graph in CMD-GCN is described as follows. We denote the undirected graph as (V, E) where V is the set of all vertices, $V = A \cup Z$, and E defines the connections among nodes. We construct a two-layer graph architecture. The first layer focuses solely on exchanging global information across domains; therefore, only agent-agent connections are considered. In contrast, the second layer enables each agent node to propagate information to its associated instance nodes by incorporating both agent-agent and agent-to-instance connections. Two agent nodes are connected if they represent the same category across different domains, while each instance node is connected to all agent nodes within the same domain. The connections (adjacency matrices) of the first and second graph layers are defined in Eqs. 2 and 3, respectively, and are illustrated in Fig. 2.

$$\text{Layer 1: } E = \{(u, v) | u, v \in A^c; u \neq v; c \in [1, C]\} \tag{2}$$

$$\text{Layer 2: } E = \begin{cases} \{(u, v) | u, v \in A^c; u \neq v; c \in [1, C]\} \bigcup \\ \{(u, v) | u \in A^d; v \in Z^d; d \in [1, D]\} \end{cases} \tag{3}$$

where $Z^d = \{\boldsymbol{z}_i^d \mid i \in [1, N_d]\}$ denotes the set of instance nodes in domain d, and $A^c = \{\boldsymbol{a}^{d,c} \mid d \in [1, D]\}$ denotes the set of agent nodes corresponding to category c across all domains.

In practice, since categories often share visual similarities, each instance node is connected to all agent nodes within the same domain, allowing flexible information aggregation regardless of category. However, an instance node should not receive equal contributions from all agent nodes; instead, the influence should depend on its similarity to each agent node. Therefore, we define the connection weight based on feature similarity as follows: $s(\boldsymbol{z}_i^d, \boldsymbol{a}^{d,c}) = \frac{1/\|\boldsymbol{z}_i^d - \boldsymbol{a}^{d,c}\|_2}{\sum_{c'=1}^{C} 1/\|\boldsymbol{z}_i^d - \boldsymbol{a}^{d,c'}\|_2}$.

4 Experiment Settings

Datasets. We use three datasets, BBBC041-v1[3] [18], IML Malaria[4] [14], and RecoPlasmodiumV1 for the main experiments. The latter has been developed by us, is currently under publication, and will be made available publicly by spring 2026. BBBC041-v1 and IML Malaria contain images of the *Plasmodium vivax*

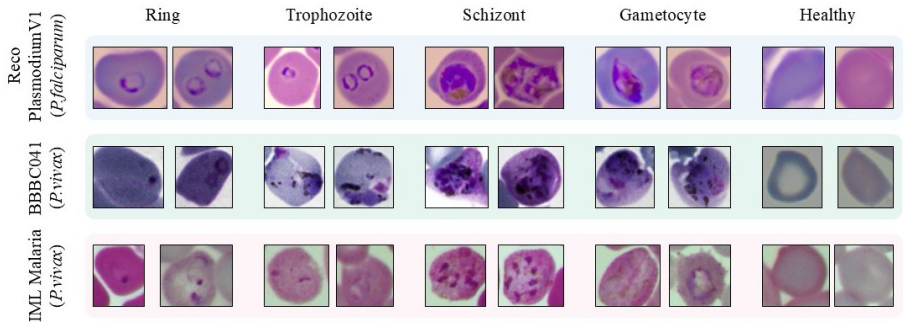


Fig. 3: Some RBCs images extracted from the three considered datasets. An RBC is either healthy or infected by *Plasmodium* at one of four development stages: Ring, Trophozoite, Schizont, or Gametocyte.

Table 1: Statistics of training, validation, and test instances for the three datasets.

Class	**RecoPlasmodiumV1**			**BBBC041-v1**			**IML Malaria**		
	Train	*Valid*	*Test*	*Train*	*Valid*	*Test*	*Train*	*Valid*	*Test*
Ring	482	78	115	317	36	169	121	15	28
Trophozoite	117	26	41	1339	134	111	57	7	13
Schizont	215	22	63	164	15	11	18	4	5
Gametocyte	101	18	41	125	19	12	169	33	59
Healthy	33066	4972	9742	69452	7968	5614	26423	3736	7740
#Blood smear images	280	39	75	1087	121	120	241	35	69

species, while RecoPlasmodiumV1 contains *Plasmodium falciparum*. Sample images from these datasets are shown in Fig. 3.

As we evaluate the complete detection–classification pipeline and the IML Malaria dataset does not provide predefined splits, we randomly partition the data at the full (blood smear) image level into 70%/20%/10% for training, testing, and validation, respectively. RecoPlasmodiumV1 is split in the same manner. At the RBC level, the partitions are approximately stratified across the five classes. For BBBC041-v1, we follow the original train/test split and randomly sample 10% of the training set for validation. Dataset statistics are reported in Table 1. Please note that, in each dataset, more than 95% of RBCs are healthy.

Implementation Details. We train our detection and classification models separately. For detection, YOLOv11n [36] pretrained on COCO [37] is fine-tuned using the SGD optimizer [38], with a learning rate of 10^{-2} over 100 epochs. For classification, the feature extractor ResNet50 [39] pretrained on ImageNet [40] and CMD-GCN implemented with PyTorch Geometric [41] are trained using the Adam optimizer [42] with an initial learning rate of 10^{-4}, reduced by a

[3] Available at https://bbbc.broadinstitute.org/BBBC041
[4] Available at https://www.kaggle.com/datasets/qaziammararshad/iml-malaria1

10^{-1} factor after half of the total 50 epochs. Experiments are performed on an NVIDIA-P100 GPU.

Evaluation Strategy. We evaluate four methods to assess the effectiveness of our proposed CMD-GCN. The first is Individual Training (IT), where the model is trained separately on each domain. The second is Joint Training (JT), where data from all domains are combined to train a single model. JT serves as a basic multi-domain learning approach; both IT and JT are trained without CMD-GCN or any specialized multi-domain learning modules. As an additional baseline to assess CMD-GCN, we re-implement Multi-Domain Information Fusion (MDIF) [7], originally proposed for unsupervised multi-source domain adaptation in person re-identification, and adapt it to our context. MDIF is placed on top of the same backbone as CMD-GCN (ResNet50) for feature refinement.

Evaluation Metrics. To evaluate the performance of our *Plasmodium* recognition pipeline, we employ standard metrics for multiclass detection, including mean Average Precision over all classes (mAP) and over the four parasitized classes (mAP@4) at an IoU threshold of 0.5. In addition, we report accuracy (Acc), macro F1-score ($F1$) across all classes, and weighted F1-score over the four parasitized classes ($wF1$@4), computed at an IoU threshold of 0.5 and a confidence threshold of 0.3. Given the F1-scores of the four parasitized classes $F1^c, c \in \{R, T, S, G\}$, the healthy class $F1^H$, and the corresponding sample counts N^c, $F1$ and $wF1$@4 are computed as follows: $F1 = \frac{\sum_{c\in\{R,T,S,G,H\}} F1^c}{5}$, $wF1@4 = \sum_{c\in\{R,T,S,G\}} \frac{N^c}{N^R+N^T+N^S+N^G} F1^c$. Please note that all false positives and false negatives from the detection stage are taken into account when evaluating the full pipeline, making this experimental protocol more representative of real-world conditions than most protocols from the literature (*c.f.* Section 2).

5 Experiment Results

5.1 Effectiveness of CMD-GCN for the Recognition Pipeline

We evaluate the four aforementioned methods applied to the classification step and report their performance in the complete recognition pipeline. Table 2 shows experimental results on the three datasets. We found that *incorporating multiple datasets for training does not always yield positive effects* without appropriate methods. As shown in Table 2, the performance achieved with joint training (JT) or with domain-level fusion (MDIF) is not consistently superior to individual training (IT). This can be explained by the loss of category information during cross-domain information propagation and/or poor management of data imbalance. Indeed, in MDIF, the per-domain agent nodes are certainly dominated by healthy RBC samples. In contrast, our proposed method, CMD-GCN, reaches the highest scores for most metrics across all datasets, especially for parasitized classes. Significant gains compared to MDIF are observed on both *P.*

Table 2: Performance comparison of four training methods for the *Plasmodium* recognition pipeline on three thin blood smear image datasets, reported using the total accuracy *Acc*, *mAP*, macro F1-score over all classes ($F1$), and the weighted F1-score over the four parasitized classes ($wF1@4$) (all in %). **Bold** and underlined values indicate the **best** and second-best results, respectively.

Method	RecoPlasmodiumV1					BBBC041-v1					IML Malaria				
	Acc	*mAP*	*F1*	*mAP@4*	*wF1@4*	*Acc*	*mAP*	*F1*	*mAP@4*	*wF1@4*	*Acc*	*mAP*	*F1*	*mAP@4*	*wF1@4*
IT	**93.62**	82.73	85.42	79.44	83.61	82.99	29.51	37.91	13.69	30.90	75.95	64.88	69.87	57.06	74.55
JT	93.58	79.97	84.19	75.25	82.60	83.40	32.85	35.08	17.69	31.59	75.94	60.07	66.87	51.06	73.22
MDIF [7]*	93.57	82.38	85.29	78.27	83.26	83.37	33.52	40.57	18.57	38.26	75.93	60.05	66.12	51.02	74.44
CMD-GCN	93.59	**84.16**	**85.89**	**80.49**	**84.03**	**83.78**	**36.69**	**41.33**	**22.50**	**46.51**	**75.99**	**69.02**	**73.53**	**62.23**	**78.00**

* We implemented MDIF in our pipeline following the description in [7].

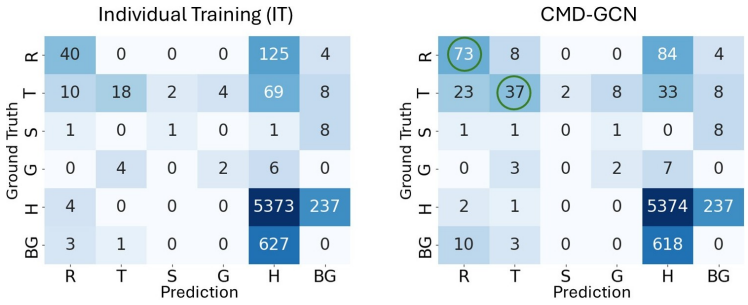


Fig. 4: Confusion matrices for IT (left) and CMD-GCN (right) on the BBBC041-v1 test set. R, T, S, G, H, BG respectively stand for **R**ing, **T**rophozoite, **S**chizont, **G**ametocyte, **H**ealthy and **B**ack**g**round classes. With CMD-GCN, substantial improvements are observed for the **R**ing and **T**rophozoite classes, as highlighted by the two circles in the matrix diagonal.

vivax datasets: BBBC041-v1 (roughly +4% in $mAP@4$ and +8.3% in $wF1@4$), and IML Malaria +11.2% in $mAP@4$ and +3.6% in $wF1@4$). On the *P. falciparum* dataset RecoPlasmodiumV1, CMD-GCN improves $mAP@4$ by roughly 2.2% and $wF1@4$ by about 0.8% compared with MDIF.

Fig. 4 shows the confusion matrices of the models trained with IT and CMD-GCN on the BBBC041-v1 test set. Compared with IT, CMD-GCN correctly recognizes more parasitized samples, particularly rings and trophozoites, notably by reducing their misclassifications with healthy RBCs.

5.2 Comparison with Existing Studies

Difficulties in making appropriate comparison. In the literature, there are few studies that report results for *Plasmodium* development stages recognition. As detailed in Section 2, the lack of a standard benchmark and improper evaluation protocols make the published results not accurately representative of the performance of most existing methods in real-life applications. For this reason, we focus here on the three methods published in [11], [13], [16], that employ comparatively more appropriate evaluation protocols and report results on BBBC041-v1 and/or IML Malaria.

Table 3: Performance comparison (all in %) with existing studies on the four parasitized classes, using the same evaluation strategy as in [13], [16] (*i.e.* disregarding some classes).

Method	**IML Malaria**			
	$Accuracy^{\dagger}$	$F1^{\dagger}$	$Precision^{\dagger}$	$Recall^{\dagger}$
Zedda *et al.* [16]	80.0	76.0	**85.0**	73.0
Zedda *et al.* [13]	75.8	72.7	81.3	68.8
CMD-GCN	**89.5**	**82.1**	83.0	**81.4**

Assessment against compatible studies. The train and test sets used for BBBC041-v1 by [11] notably differ from the publicly released ones (that we use) and the classes considered for evaluation differ from our method; thus, a comparison with our method is inappropriate. For the IML Malaria dataset, no official test set is provided, but Zedda *et al.* [13], [16] split the dataset using the same 70/20/10 ratio as in our study for training, validation, and testing. Thus, to offer an approximate indication of CMD-GCN's performance relative to prior work, we compare our results with those of Zedda *et al.* [13], [16] on IML Malaria by re-computing our metrics under the same protocol, *i.e.* disregarding background, healthy RBCs misclassified as parasites, and infected RBCs misclassified as healthy RBCs. Specifically, the metrics include accuracy ($Accuracy^{\dagger}$), macro F1 score ($F1^{\dagger}$), macro precision ($Precision^{\dagger}$), and macro recall (sensitivity) ($Recall^{\dagger}$) over the four parasitized classes. As reported in Table 3, our CMD-GCN outperforms the others in terms of $Accuracy^{\dagger}$, $F1^{\dagger}$, and $Recall^{\dagger}$.

Clinical practicability of our *Plasmodium* recognition pipeline. By relying on a detection step that provides only infected RBCs, the methods in [11], [13], [16] do not provide the healthy RBC counts, and thus are not enough to estimate per-stage parasitaemia, a crucial statistics for healthcare professionals. On the other hand, we first detect all RBCs from blood smear images, and then classify them as healthy RBC or one of four development stages, thereby providing the information required to estimate per-stage parasitaemia. Thus, our pipeline is suitable for deployment in real-world clinical scenarios.

5.3 Discussion

Data noise in the ground truth of public datasets. During our experiments, we observed relatively low performance on the two public datasets, even under individual training (IT). We conducted a detailed inspection of samples from BBBC041-v1 and IML Malaria and found that annotation noise is a likely contributing factor. As illustrated in Fig. 5, many samples labeled as "Healthy" actually include parts of infected RBCs belonging to one of the four parasitized classes (Fig. 5(a) and 5(b)) which could have been corrected with more precise bounding boxes. In addition, some infected RBCs are incorrectly annotated as "Healthy" (Fig. 5(c)). Such annotation noise can confuse the models.

(a) BBBC041-v1: RBCs annotated as "Healthy" but containing part of infected RBC.

(b) IML Malaria: RBCs annotated as "Healthy" but containing part of infected RBC.

(c) IML Malaria: incorrect annotations as "Healthy".

Fig. 5: Examples of noisy annotations from public datasets that may negatively affect model performance. Correctly annotated "Healthy" samples are shown in Fig. 3 for reference.

Robustness of CMD-GCN against data noise. Given the above observations, we assess the robustness of CMD-GCN against annotation noise by training the classification model under two dataset combinations: {BBBC041-v1, IML Malaria} and {BBBC041-v1, IML Malaria, RecoPlasmodiumV1}. Note that RecoPlasmodiumV1 has been collegially annotated by a pool of five expert microscopists, offering higher label quality than the other two datasets. The weighted F1-scores over the four parasitized classes ($wF1@4$) of the four methods described in Section 4 are shown in Table 4.

Despite the annotation noise in BBBC041-v1 and IML Malaria, CMD-GCN achieves the best performance when being trained with these two combinations. Moreover, when using {BBBC041-v1, IML Malaria, RecoPlasmodiumV1}, the inclusion of RecoPlasmodiumV1 further yields significant gains for CMD-GCN on BBBC041-v1, with $wF1@4$ increasing from 33.85% (using IT) to 51.74%. On the other hand, CMD-GCN's performance in RecoPlasmodiumV1 is quite similar to IT and JT, showing its robustness towards noise from BBBC041-v1 and IML Malaria. These results demonstrate CMD-GCN's ability to improve performance in noisier domains (*e.g.*, *P. vivax* in BBBC041-v1 and IML Malaria) by leveraging cleaner annotations from another domain (*P. falciparum* in RecoPlasmodiumV1), without any loss of performance in the cleaner domain.

Table 4: Performance ($wF1@4$) of different training strategies for the **classification phase** on two dataset combinations: {*BBBC041-v1, IML Malaria*} and {*BBBC041-v1, IML Malaria, RecoPlasmodiumV1*}. Values in **bold** indicate the **best result within each combination**.

Dataset Group/Method		RecoPlasmodiumV1	BBBC041-v1	IML Malaria
IT (baseline)		88.83	33.85	72.93
{BBBC041-v1 + IML Malaria}	JT	-	20.75	79.20
	MDIF [7]*	-	40.14	75.47
	CMD-GCN	-	**43.04**	**81.05**
{BBBC041-v1 + IML Malaria + RecoPlasmodiumV1}	JT	88.36	37.69	74.58
	MDIF [7]*	**88.87**	42.60	77.11
	CMD-GCN	**88.87**	**51.74**	**80.39**

* We implemented MDIF in our pipeline following the description in [7].

6 Conclusion

In this study, a two-step pipeline is proposed for recognizing *Plasmodium* development stages and healthy RBCs, which is ready for clinical applications. We leverage the diversity of multiple datasets when training the classification model to address data scarcity (at least for some classes). For that purpose, we propose CMD-GCN, a new supervised multi-domain learning approach, to align distributions and enhance feature representations. By simultaneously incorporating domain and category information during the construction of the relational graph, CMD-GCN enables information propagation at a fine-grained level, outperforming different types of methods (multi-domain or not) on three existing datasets. Additionally, CMD-GCN is also robust to annotation noise.

Although the limited number of studies and the lack of a standard benchmark in *Plasmodium* development stage recognition hinders a fair and accurate comparison of our results over existing works, CMD-GCN is a generic method that can be seamlessly integrated into diverse backbones to solve other fine-grained classification tasks, beyond this specific application. An interesting direction to explore further would be to adapt CMD-GCN to semi-supervised learning, where a large amount of data is available, but mostly unlabeled.

References

[1] World Health Organization, *Malaria*, https://www.who.int/news-room/fact-sheets/detail/malaria, Accessed: July 29, 2025.

[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 et al., "Impact of school-based malaria case management on school attendance, health and education outcomes: A cluster randomised trial in southern malawi," *BMJ global health*, vol. 5, no. 1, 2020.

[4] T. Xiao, H. Li, W. Ouyang, and X. Wang, "Learning deep feature representations with domain guided dropout for person re-identification," in *Proceedings of the IEEE conference on CVPR*, 2016, pp. 1249–1258.

[5] B. Chen et al., "Variational attention: Propagating domain-specific knowledge for multi-domain learning in crowd counting," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2021, pp. 16 065–16 075.

[6] M. Cho, T. Kim, M. Shim, D. Wee, and S. Lee, "Towards multi-domain learning for generalizable video anomaly detection," *Advances in Neural Information Processing Systems*, vol. 37, pp. 50 256–50 284, 2024.

[7] Z. Bai, Z. Wang, J. Wang, D. Hu, and E. Ding, "Unsupervised multi-source domain adaptation for person re-identification," in *Proceedings of the IEEE/CVF conference on CVPR*, 2021, pp. 12 914–12 923.

[8] Y. Liu, J. Wang, C. Huang, Y. Wang, and Y. Xu, "Cigar: Cross-modality graph reasoning for domain adaptive object detection," in *Proceedings of the IEEE/CVF Conference on CVPR*, 2023, pp. 23 776–23 786.

[9] X. Ma, T. Zhang, and C. Xu, "Gcan: Graph convolutional adversarial network for unsupervised domain adaptation," in *Proceedings of the IEEE/CVF Conference on CVPR*, 2019, pp. 8266–8276.

[10] S. Ma, D. Qian, K. Ye, and S. Zhang, "Cake: Category aware knowledge extraction for open-vocabulary object detection," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 39, 2025, pp. 5982–5990.

[11] J. Hung and A. Carpenter, "Applying faster r-cnn for object detection on malaria images," in *Proceedings of the IEEE conference on CVPRW*, 2017, pp. 56–61.

[12] G. Wang, G. Luo, H. Lian, L. Chen, W. Wu, and H. Liu, "Application of deep learning in clinical settings for detecting and classifying malaria parasites in thin blood smears," in *Open forum infectious diseases*, Oxford University Press US, vol. 10, 2023.

[13] L. Zedda, A. Loddo, and C. Di Ruberto, "A deep architecture based on attention mechanisms for effective end-to-end detection of early and mature malaria parasites in a realistic scenario," *Computers in Biology and Medicine*, vol. 186, p. 109 704, 2025.

[14] Q. A. Arshad 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.

[15] L. Zedda, A. Loddo, and C. Di Ruberto, "A deep learning based framework for malaria diagnosis on high variation data set," in *International Conference on Image Analysis and Processing*, Springer, 2022, pp. 358–370.

[16] L. Zedda, A. Loddo, C. Di Ruberto, et al., "Sammi: Segment anything model for malaria identification," *VISAPP*, vol. 3, pp. 367–374, 2024.

[17] M. S. Davidson et al., "Automated detection and staging of malaria parasites from cytological smears using convolutional neural networks," *Biological imaging*, vol. 1, 2021.

[18] V. Ljosa, K. L. Sokolnicki, and A. E. Carpenter, "Annotated high-throughput microscopy image sets for validation," *Nature methods*, vol. 9, no. 7, p. 637, 2012.

[19] R. Girshick, "Fast r-cnn," in *Proceedings of the IEEE international conference on computer vision*, 2015, pp. 1440–1448.

[20] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," in *Advances in Neural Information Processing Systems*, vol. 25, Curran Associates, Inc., 2012.

[21] G. Jocher, A. Chaurasia, and J. Qiu, *Ultralytics yolov8*, version 8.0.0, 2023. [Online]. Available: https://github.com/ultralytics/ultralytics.

[22] A. Dosovitskiy et al., "An image is worth 16x16 words: Transformers for image recognition at scale," *arXiv preprint arXiv:2010.11929*, 2020.

[23] H. A. H. Chaudhry, M. S. Farid, A. Fiandrotti, and M. Grangetto, "A lightweight deep learning architecture for malaria parasite-type classification and life cycle stage detection," *Neural Computing and Applications*, vol. 36, no. 31, pp. 19 795–19 805, 2024.

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

[25] 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.

[26] D. Sukumarran et al., "Automated identification of malaria-infected cells and classification of human malaria parasites using a two-stage deep learning technique," *IEEE Access*, 2024.

[27] Y. Chen et al., "Scaledet: A scalable multi-dataset object detector," in *Proceedings of the IEEE/CVF Conference on CVPR*, 2023, pp. 7288–7297.

[28] Z. Wang, Y. Li, H. Zhao, and S. Wang, "One for all: Multi-domain joint training for point cloud based 3d object detection," *Advances in Neural Information Processing Systems*, vol. 37, pp. 56 859–56 877, 2024.

[29] R. Zheng et al., "Tmt-vis: Taxonomy-aware multi-dataset joint training for video instance segmentation," *Advances in Neural Information Processing Systems*, vol. 36, pp. 38 844–38 862, 2023.

[30] Y. Liu et al., "Multi-space alignments towards universal lidar segmentation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024, pp. 14 648–14 661.

[31] J. Liang, E. Zhang, J. Zhang, and C. Shen, "Multi-dataset training of transformers for robust action recognition," *Advances in Neural Information Processing Systems*, vol. 35, pp. 14 475–14 488, 2022.

[32] H. Yan, Y. Ding, P. Li, Q. Wang, Y. Xu, and W. Zuo, "Mind the class weight bias: Weighted maximum mean discrepancy for unsupervised domain adaptation," in *Proceedings of the IEEE conference on CVPR*, 2017, pp. 2272–2281.

[33] M. Long, H. Zhu, J. Wang, and M. I. Jordan, "Deep transfer learning with joint adaptation networks," in *International conference on machine learning*, PMLR, 2017, pp. 2208–2217.

[34] Y. Ganin and V. Lempitsky, "Unsupervised domain adaptation by backpropagation," in *Proceedings of the 32nd ICML - Volume 37*, ser. ICML'15, Lille, France: JMLR.org, 2015, pp. 1180–1189.

[35] W.-G. Chang, T. You, S. Seo, S. Kwak, and B. Han, "Domain-specific batch normalization for unsupervised domain adaptation," in *Proceedings of the IEEE/CVF conference on CVPR*, 2019, pp. 7354–7362.

[36] R. Khanam and M. Hussain, "Yolov11: An overview of the key architectural enhancements," *arXiv preprint arXiv:2410.17725*, 2024.

[37] T.-Y. Lin et al., "Microsoft coco: Common objects in context," in *European conference on computer vision*, Springer, 2014, pp. 740–755.

[38] H. Robbins and S. Monro, "A stochastic approximation method," *The annals of mathematical statistics*, pp. 400–407, 1951.

[39] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE CVPR*, 2016, pp. 770–778.

[40] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "Imagenet: A large-scale hierarchical image database," in *2009 IEEE conference on CVPR*, IEEE, 2009, pp. 248–255.

[41] M. Fey and J. E. Lenssen, "Fast graph representation learning with PyTorch Geometric," in *ICLR Workshop on Representation Learning on Graphs and Manifolds*, 2019.

[42] D. P. Kingma, "Adam: A method for stochastic optimization," *arXiv preprint arXiv:1412.6980*, 2014.