

Transfer learning as a regularized optimization : A mathematical and experimental model for medical image classification

Manasi Ghule *

Anita Chaware [†]

Department of Computer Science

SNDT Women's University

Juhu Campus

Mumbai 400049

Maharashtra

India

Abstract

Transfer learning has emerged as a pivotal technique for biomedical image classification, where labeled data is limited and computation cost of model training is very high. By leveraging pre-trained neural networks, transfer learning accelerates convergence and enhances model precision on domain-specific tasks. The efficiency and performance of transfer learning is governed by various factors that need to be fine-tuned. Despite its wide success, the mathematical understanding of transfer learning behavior remains limited. In this paper, we systematically studied the mathematical aspect of transfer learning as a regularized optimization problem, where adaptation to a target domain is balanced against deviation from pre-trained parameters, which plays an important role in classification of datasets including small and imbalanced datasets invariably found in medical diagnosis. We validated this model using controlled regularization across a simulated neural network and three models of deep convolutional networks pre-trained on two different domains for classification of ultrasound scans used in diagnosis of polycystic ovary syndrome. Our experimental results suggest that regularization improves generalization and preserves inductive bias, due to a principled adaptation-retention tradeoff, thereby bridging mathematical modeling and practical implementation of transfer learning in domain-adaptive medical imaging applications.

Subject Classification: 68T07, 68T05, 62H30.

Keywords: Transfer learning, Regularized optimization, Neural networks, Mathematical modelling, Medical image classification.

* E-mail: ghule.seema@gmail.com (Corresponding Author)

[†] E-mail: Anita.Chaware@computersc.sndt.ac.in

I. Introduction

Radiological tests performed using various imaging techniques play a very important role in diagnosis of many diseases. For many years, interpretation of results of such tests is done by radiologists or doctors. This process is time-consuming and depends on expert knowledge. Deep learning approaches have achieved significant success across multiple medical imaging modalities including computed tomography and ultrasound classification tasks [1]. The advent of image classification tools using Deep Learning models aided by transfer learning has improved accuracy as well as efficiency of diagnosis but still there are some limitations. Two major challenges arise when applying transfer learning to medical image classification tasks. First, scarcity of labelled data, due to high cost and time involved in annotation. Second, imbalanced nature of data, because of the natural bias created due to presence of more normal cases than the diseased ones in reality [2].

Transfer learning enables adaptation of neural network models trained on a large general database like ImageNet to a small domain specific dataset like MedMNIST. Several transfer learning methods focus on adapting feature representations across different domains and distributions. The very basic and general features like edges, textures, primary shapes learned by initial convolutional filters from the general dataset are directly used as pre-trained filters by transfer learning algorithm as a starting point [12]. Transfer learning leverages these pre-trained filters to accelerate convergence and improve generalization [3], [11]. However, the mathematical mechanism of “knowledge retention and adaptation” is not often formally described. Zhuang et al. [4] in their survey stated that “Transfer learning aims to improve the performance of target learners on target domains by transferring knowledge contained in different but related source domains”. Due to its wider application prospects, transfer learning has become a popular and upcoming area in machine learning. In practice, transfer learning is often treated as an engineering procedure rather than a mathematical formulation process. Deep transfer learning has emerged as an effective approach for improving medical image classification performance under limited data availability [5]. The efficiency of deep neural network models can be improved using models that are pre-trained on a large general dataset as training a deep learning model from scratch is very difficult, the major difficulty being less availability of labeled data and higher computational cost. Transfer learning overcomes these using pre-trained models. Transfer learning

using pretrained models has shown improved efficiency and classification performance in medical imaging applications [4], [5]. Recent healthcare-oriented AI frameworks have incorporated transfer learning and domain adaptation techniques to improve medical image analysis performance [6].

This paper interprets transfer learning as a regularized optimization problem - according to this concept, as shown in Figure 1, adaptation to a new medical imaging task competes with retention of pre-trained knowledge. Through mathematical modeling, a simplified single-neuron neural network model and convolutional architectures (ResNet50, VGG16, MobileNetV3) on ImageNet and MedMNIST, we validated λ -controlled regularization for PCOS ultrasound image classification. We model transfer learning as an optimization process that minimizes target-task loss while penalizing deviation from pretrained parameters. The proposed formulation captures the adaptation-retention equilibrium controlled by the regularization hyperparameter λ .

II. Related Work

Transfer learning has been extensively studied from both experimental and theoretical perspectives. Cao, Gu, and Guo [7] proposed a mathematical framework for transfer learning. They studied it as an optimization problem and practically measured its performance under different conditions resulting in the concept of “optimal transfer learning”. For example, in an image classification setup, only the last few layers of the model are retrained when solving the target task, while the feature extraction layers derived from the source task are directly utilized. This approach allows for leveraging the learned representations from the source task, optimizing the model specifically for the target task [7].

In deep learning, optimization algorithms minimize model loss by iteratively updating network parameters using gradient-based learning methods [12]. Proper application of optimization techniques helps in performance improvement as well as better and faster convergence [8]. “Domain transfer learning, which learns a target classifier using labeled data from a different distribution, has shown promising value in knowledge discovery yet still been a challenging problem”. A novel transfer learning framework based on the regularization theory is described in this paper by Long et al. [9]. Li, Grandvalet, and Davoine [10] experimentally showed that in transfer learning initial weights of parameters are changed to new values. The change in initial weight value

results in loss of useful knowledge. They argued that the regularization process that moves parameters towards initial weights doesn't fit in the framework of transfer learning resulting in failure to produce best results. They proposed a logical "parameter regularization approach, where the pre-trained model is both used as the starting point of the optimization process and as the reference in the penalty," which will preserve the initial knowledge. They tested simple regularization techniques for transfer learning with convolutional networks, concluding that there is "a trade-off between the solution to the target task and the pre-trained parameters" [10]. Transfer learning (TL) has become the de facto standard in medical image analysis, enabling models trained on large natural image datasets like ImageNet to adapt effectively to specialized domains such as ultrasound, X-ray or MRI. However, when the target dataset is small or imbalanced, transfer learning may lead to unstable gradient behavior, overfitting, and biased representations [4], [10], [17].

Although transfer learning methods achieve high classification performance on limited medical datasets, they still lack a comprehensive theoretical foundation. Despite broad empirical success, many practitioners continue to rely on heuristic fine-tuning strategies rather than formal mathematical analysis of why transfer learning works effectively. Recent surveys emphasize this gap in understanding the mathematical mechanisms underlying transfer learning [4], [7], [17]. Cross-domain transfer learning methods attempt to align feature representations between source and target domains [13]. Collaborative transfer learning frameworks have been proposed to improve knowledge transfer across heterogeneous domains [14].

III. Mathematical Framework

Formulation of Transfer Learning:

Definition of Domains

Let:

$D_s = \{(x_s, y_s)\}$: source domain and $D_t = \{(x_t, y_t)\}$: target domain Each domain has:

An input space $\{X\}$, A label space $\{Y\}$ and A probability distribution $(P(x, y))$ The source and target may differ in data distribution: $[P_s(x_s, y_s) \neq P_t(x_t, y_t)]$ [15].

Model Parameterization

Let, a neural network model be represented by $f_{\theta}(x)$, parameterized by weights $\theta \in \mathbb{R}^d$. After pre-training, parameters become θ_s ; during transfer, we seek adapted parameters θ_t .

Objective Function seeks to minimize the transfer learning loss so as to adapt the model pre-trained on source distribution to a target distribution.

The general form of equation for transfer learning loss:

$$L(\theta) = \min_{\theta} \mathbb{E}_{(x_t, y_t)} \mathbb{P}_t[L(f_{\theta}(x_t), y_t)] + \lambda R(\theta, \theta_s)$$

Where, θ = Current model weights θ_s = Pretrained source weights

R = Regularization term that prevents new model from “forgetting” previous knowledge

λ = Regularization parameter controlling the balance between preserving pretrained knowledge and adapting to the target task [9], [10], [16].

Thus, we model transfer learning as solving: $LTL(\theta) = \mathbb{P}_t[L(f_{\theta}(x_t), y_t)] + \lambda \|\theta - \theta_s\|^2$

$$LTL(\theta) = \underbrace{L_t(\theta)}_{\text{Target Loss}} + \underbrace{\lambda \|\theta - \theta_s\|^2}_{\text{Regularization term}} \quad (\text{Equation 1})$$

where $L_t(\theta)$ denotes the target task loss and the regularization term constrains deviation from pretrained source weights [9], [10], [20].

The first term encourages adaptation to the target domain, while the second term penalizes deviation from the pretrained source weights, thereby preserving knowledge, where λ balances adaptation and retention.

The goal of transfer learning is to find new parameters that perform well on the target data while staying close to the pretrained parameters or balancing **new learning** (first term) and **memory of source knowledge** (second term).

Hence, transfer learning involves constrained optimization to minimize a combined loss ($LTL(\theta)$) using gradient descent as shown in the following equation.

Derivation of gradients

Taking gradients with respect to θ : $\frac{\partial}{\partial \theta} LTL = \frac{\partial}{\partial \theta} L_t + 2\lambda (\theta - \theta_s)$

The gradient consists of two competing components:

1. Reduces target error
2. Stays close to pre-trained weights

This captures the dual optimization behavior of adaptation and knowledge retention.

Equilibrium Condition can be depicted as follows: At convergence:

- Small gradients $\rightarrow$ parameters remain close to pretrained values.
- Large gradients $\rightarrow$ greater deviation to capture new patterns.

Thus, fine-tuning in transfer learning behaves like controlled optimization-driven perturbation of pre-trained parameters.

This regularization can be interpreted as follows:

The term $\lambda \|\theta - \theta_s\|^2$ prevents catastrophic forgetting.

Intuitively, λ regulates the deviation from pre-trained parameters.

- $\lambda \rightarrow 0$ Pure fine-tuning (maximum adaptation)
- $\lambda \rightarrow \infty$ Feature reuse (maximum retention)
- Intermediate λ yields best generalization
- High λ : stronger memory, slower adaptation.
- Low λ : more flexibility, faster adaptation but higher risk of forgetting.

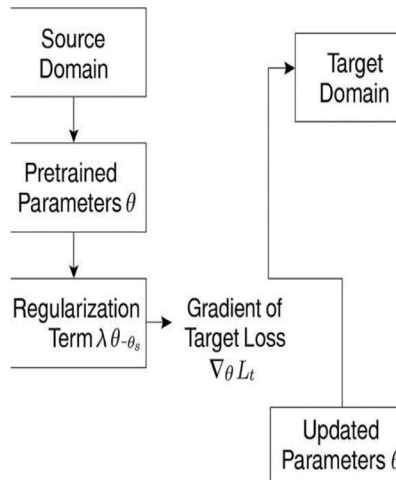


Figure 1

Conceptual view of transfer learning as a regularized optimization process balancing target-domain adaptation and retention of pretrained representations

This reflects a **bias–variance tradeoff** in transfer learning. Final parameters are a balance between old model and new gradient influence. The smaller the divergence between source and target, the more effective the transfer process [18]. Various regularization schemes such as Lasso, L1-group Lasso, and elastic net regularization are widely used in machine learning [19]. Similarly, L2, L2-SP, Fisher-weighted L2SP regularization methods are used for fine tuning of parameters during transfer learning in Convolutional Neural Networks to preserve pretrained representations while enabling domain adaptation [20].

IV. Methodology

1. Datasets

Pre-training sources:

ImageNet for natural-image pretraining and MedMNIST for medical-image pretraining [21], [22].

Target Task:

PCOS ultrasound classification (grayscale, 224×224 , normalized images).

2. Experimental Setup:

A. λ -Sensitivity Analysis Using a Simplified Simulation Model

This experiment analyzes how varying regularization strengths influence domain adaptation across increasing levels of model complexity.

B. Numerical Simulation

We implemented a simple linear model, a single-neuron linear model to illustrate the Equation (1). Let $\theta_s = [0.8, 1.2, -0.5]$, $\theta \in \mathbb{R}^3$. The target sample is, $x_t = [1.0, 2.0, 3.0]$ and $y_t = 4.5$.

We computed numerical gradients by finite differences ($\epsilon=1e-4$) and ran gradient descent updates for 20 epochs with learning rate $\eta=0.05$. We swept $\lambda \in \{0.01, 0.05, 0.1, 0.5, 1.0\}$.

C. Validation on Real Medical Imaging Data

We carried out the experiment using pre-trained CNN models to visualize training loss against varying values of λ .

- Architectures: ResNet50, VGG16, MobileNetV3.

- Optimization: SGD (momentum=0.9, LR=1e-3, batch=16, 25 epochs).
- Regularization λ sweep = {0, 0.01, 0.05, 0.1, 0.5, 1.0}.
- Evaluation: Accuracy, F1, AUC, Parameter Drift, Loss curves.

D. Metrics and Implementation

- Parameter drift measured as $\|\theta - \theta_s\|^2$.
- Target loss curves across varying λ values.
- Environment: Google Colab + Google Drive for persistence.

V. Results

Results of single-neuron simulation model:

Table 1
 λ vs Final Loss and Weight Change

Λ	Final Loss	Weight Drift $\ \theta - \theta_s\ $
0.01	3.1683	0.3831
0.05	3.0861	0.853
0.1	2.9879	1.3998
0.5	2.3516	4.5832
1.0	1.8229	6.8465

Results of validation on real dataset:

Table 1 displays results of the numerical simulation showing a clear trade-off:

Smaller λ leads to higher final loss but smaller weight deviation $\|\theta - \theta_s\|$.

Larger λ achieves lower final loss but gives slower convergence and greater deviation from the starting weights.

This precisely matches the analytical equilibrium expression discussed in the Mathematical Framework indicating that increasing λ improves convergence in terms of loss minimization, but it destabilizes weight preservation. Conversely, keeping λ small maintains weight stability but sacrifices accuracy.

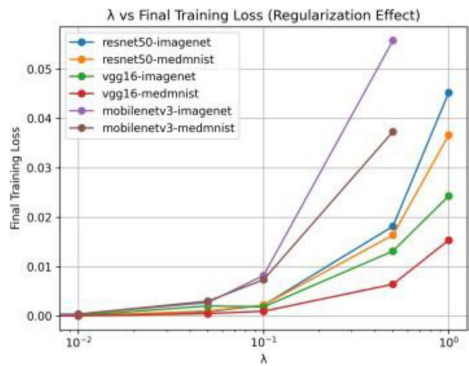


Figure 2
Regularization effect

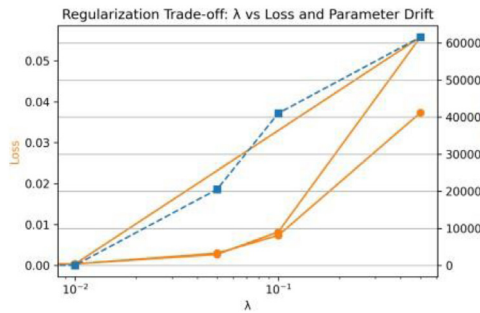


Figure 3
Parameter drift

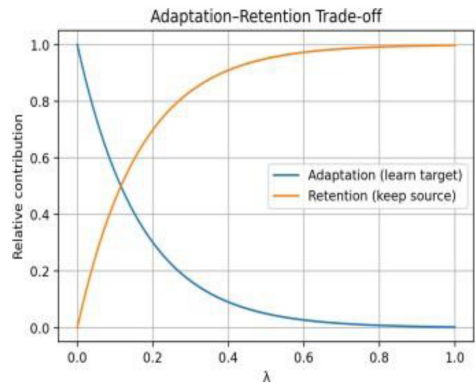


Figure 4
Adaptation-retention

Figure 2 illustrates the variation of final training loss with regularization strength (λ) for three convolutional architectures—ResNet50, VGG16, and MobileNetV3—under two pre-training sources (ImageNet and MedMNIST). Across all models, the training loss increases with λ , confirming that stronger regularization constrains network adaptation during fine-tuning. This behavior demonstrates the proposed mathematical formulation.

From the models studied, MobileNetV3 showed the steepest rise in loss, indicating higher sensitivity to λ due to its lightweight architecture. VGG16 displayed a smoother trend, reflecting stable optimization across its convolutional blocks, while ResNet50 showed moderate loss increase, due to internal residual regularization.

For a given λ , MedMNIST-pre-trained models achieved lower losses than ImageNet-pre-trained, highlighting that domain similarity allows stronger retention (higher λ) without hampering adaptation.

Figure 3 presents the variation of parameter drift with increasing λ for the same models and pre-training sources. The results show an inverse relationship between weight stability and model adaptability. As λ increases, the total drift initially decreases, indicating stronger retention of pre-trained representations. Comparatively, MedMNIST-pre-trained models exhibit smaller drift across λ , suggesting that domain similarity is bringing weight alignment. ImageNet-pre-trained models, show greater drift variation due to the larger domain gap.

Figure 4 shows that at low λ values, adaptation dominates as the model aggressively fine-tunes to the target data, potentially forgetting pre-trained knowledge. At high λ values, retention dominates as strong regularization preserves source knowledge but limits adaptation. The intermediate λ region represents the optimal balance between the two, minimizing total transfer loss.

Figure 5 indicates that MedMNIST pre-training, representing a medically relevant domain, exhibits its loss minimum at a higher λ than ImageNet pre-training. This shift of the optimal λ indicates that greater domain alignment allows stronger retention of pre-trained weights without hindering target adaptation. When the source domain is dissimilar, as with ImageNet, weaker regularization (smaller λ) is required to facilitate sufficient adaptation.

All standard metrics such as Accuracy, F1 score, AUC, ROC curves, loss curves, and confusion matrices for all CNNs showed similar trends indicating that these metrics did not depend on value of λ compared to parameter-drift.

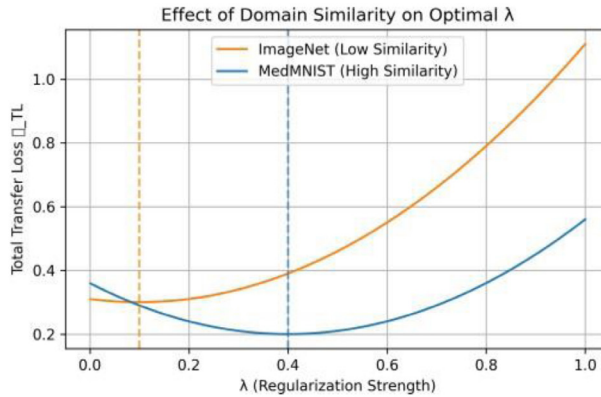


Figure 5
Domain similarity

VI. Discussion

These results show that λ serves as a control for knowledge retention in transfer learning and validate the regularized-optimization interpretation across diverse CNN architectures.

Beyond a certain threshold, however, drift values may rise slightly, reflecting optimization instability caused by over-regularization. This nonlinear behavior indicates that transfer learning can be interpreted as a regularized optimization problem in which λ governs the equilibrium between *adaptation* (large weight updates) and *retention* (limited deviation from pre-trained weights).

These results suggest that λ regulates the stability of parameter transfer, proving the theoretical framework experimentally across architectures and pre-training conditions.

The proposed framework provides a theoretical interpretation for commonly used transfer learning heuristics such as layer freezing and reduced learning rates and stresses the importance of λ as a tuning knob.

The λ shift indicates that adaptation–retention equilibrium depends on domain similarity and suggests that selection of λ can be guided by domain similarity.

Limitations:

- The PCOS ultrasound dataset used for target training was relatively small which limits the evaluation of λ under clinical conditions.

- A single regularization term exhibiting the adaptation–retention balance is used in the study. Deeper network or other architectural study is missing.
- Computational limitations restricted the experiments to only fixed hyperparameters.

Future work:

The work can be taken ahead with more datasets of medical images of different modalities to perform cross-dataset evaluation. It can be extended to Transformer-based as well as hybrid deep learning architectures.

Further work can integrate more complex structures suggested in the literature like domain divergence ($D(P_s, P_t)$), Optimal Transport penalties [23], or Fisher-information geometry of parameter space [24] for deeper theoretical understanding.

VII. Conclusion

This study presents transfer learning as a regularized optimization providing a mathematical framework for theoretical formalization of transfer learning. The results are analyzed numerically across simulation as well as CNN models for validating the regularized optimization concept. The paper offers mathematical basis for balance between feature reuse and fine tuning. It proposes that λ can be chosen during fine-tuning of domain-adaptive medical-image classification tasks where datasets tend to be small and imbalanced.

References

- [1] A. Shah, W. Patel, and H. Koyuncu, “Empowering healthcare innovation: IoT-enabled smart systems and deep learning for enhanced diabetic retinopathy in the telehealth landscape,” *J. Interdiscip. Math.*, vol. 27, no. 2, pp. 355–367 (2024), doi: 10.47974/JIM-1836.
- [2] H. Zhang and Y. Qie, “Applying deep learning to medical imaging: A review,” *Appl. Sci.*, vol. 13, no. 18, p. 10521 (2023), doi: 10.3390/app131810521.
- [3] D. P. Sahoo, M. Rout, P. K. Mallick, and S. R. Samanta, “Comparative analysis of medical images using transfer learning based deep learning models,” *2022 International Conference on Advancements in Smart,*

- Secure and Intelligent Computing (ASSIC)*, Bhubaneswar, India, pp. 1-8 (2022), doi: 10.1109/ASSIC55218.2022.10088373.
- [4] F. Zhuang, Q. Qi, K. Duan, D. Xi, Y. Zhu, H. Zhu, H. Xiong, and Q. He, "A comprehensive survey on transfer learning," arXiv:1911.02685 (2020), doi: 10.48550/arXiv.1911.02685.
 - [5] V. Gunturu, N. Maiti, B. Toure, P. Kunekar, S. B. Banu, and D. Sahaya Lenin, "Transfer learning in biomedical image classification," in *Proc. Int. Conf. Adv. Comput., Commun. Appl. Informatics (ACCAI)*, pp. 1-5 (2024), doi: 10.1109/ACCAI61061.2024.10601862.
 - [6] A. Seethalakshmy, T. Tamilvizhi, K. Naga Sowjanya, and B. Bala, "Deep learning based computed tomography image classification of COVID-19 patients," *J. Interdiscip. Math.*, vol. 26, no. 3, pp. 371-381 (2023), doi: 10.47974/JIM-1668.
 - [7] H. Cao, H. Gu, and X. Guo, "Feasibility of transfer learning: A mathematical framework," arXiv:2305.12985 (2023), doi: 10.48550/arXiv.2305.12985.
 - [8] N. Mamatha and B. Shetty, "Machine learning mathematical optimization techniques for enhancing model efficiency and convergence," *World J. Adv. Res. Rev.*, vol. 10, no. 3, pp. 471-481 (2021), doi: 10.30574/wjarr.2021.10.3.0235.
 - [9] M. Long, J. Wang, G. Ding, S. J. Pan, and P. S. Yu, "Adaptation regularization: A general framework for transfer learning," *IEEE Trans. Knowl. Data Eng.*, vol. 26, no. 5, pp. 1076-1089 (2014), doi: 10.1109/TKDE.2013.111.
 - [10] X. Li, Y. Grandvalet, and F. Davoine, "A baseline regularization scheme for transfer learning with convolutional neural networks," *Pattern Recognit.*, vol. 98, p. 107049 (2020), doi: 10.1016/j.patcog.2019.107049.
 - [11] S. J. Pan and Q. Yang, "A survey on transfer learning," *IEEE Trans. Knowl. Data Eng.*, vol. 22, no. 10, pp. 1345-1359 (2010), doi: 10.1109/TKDE.2009.191.
 - [12] I. Goodfellow, Y. Bengio, and A. Courville, *Deep Learning*. Cambridge, MA, USA: MIT Press (2016). [Online]. Available: <http://www.deeplearningbook.org>
 - [13] S. J. Pan, X. Ni, J.-T. Sun, Q. Yang, and Z. Chen, "Cross-domain sentiment classification via spectral feature alignment," in *Proc. 19th Int. Conf. World Wide Web (WWW'10)*, New York, USA: Association for Computing Machinery, pp. 751-760 (2010), doi: 10.1145/1772690.1772767.

- [14] W. Zhang, P. Zhang, B. Zhang, X. Wang, and D. Wang, "A collaborative transfer learning framework for cross-domain recommendation," arXiv:2306.16425 (2023), doi: 10.48550/arXiv.2306.16425.
- [15] T. Cody and P. A. Beling, "A systems theory of transfer learning," arXiv:2107.01196 (2021), doi: 10.48550/arXiv.2107.01196.
- [16] P. Patil, S. Charkha, and S. V. Kakade, "Transfer learning optimization in medical imaging enhancing diagnostic accuracy with limited data sets," *J. Inf. Optim. Sci.*, vol. 46, no. 4-A, pp. 1105–1116 (2025), doi: 10.47974/JIOS-1895.
- [17] F. Yu, X. Xiu, and Y. Li, "A survey on deep transfer learning and beyond," *Mathematics*, vol. 10, no. 19, p. 3619 (2022), doi: 10.3390/math10193619.
- [18] Y. Tan and X. Xiang, "Illustrating Ben-David et al.'s theory of learning from different domains," (2021), doi: 10.13140/RG.2.2.26939.13607.
- [19] K. P. Murphy, *Machine Learning: A Probabilistic Perspective*. Cambridge, MA, USA: MIT Press (2012).
- [20] X. Li, "Regularization schemes for transfer learning with convolutional networks," *Ph.D. dissertation*, Université de Technologie de Compiègne (2019). [Online]. Available: <https://theses.hal.science/tel-02439193>.
- [21] 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 Computer Vision and Pattern Recognition*, pp. 248–255 (2009), doi: 10.1109/CVPR.2009.5206848.
- [22] J. Yang, R. Shi, D. Wei, Z. Liu, L. Zhao, B. Ke, H. Pfister, B. Ni, and M. Xu, "MedMNIST v2: A large-scale lightweight benchmark for 2D and 3D biomedical image classification," *Sci. Data*, vol. 10, no. 1, p. 41 (2023), doi: 10.1038/s41597-022-01721-8.
- [23] G. Peyré and M. Cuturi, "Computational optimal transport," arXiv:1803.00567 (2020), doi: 10.48550/arXiv.1803.00567.
- [24] W. Li, J. Lu, and L. Wang, "Fisher information regularization schemes for Wasserstein gradient flows," *J. Comput. Phys.*, vol. 416, p. 109449 (2020), doi: 10.1016/j.jcp.2020.109449.

Received December, 2025