Evolutionary Multi-Objective Calibration of MC-Dropout Predictive Uncertainty Quantification for Diabetic Retinopathy

Miguel Abreu-Cardenas
mabreu@estudiantec.ac.cr
Costa Rica Institute of Technology
Costa Rica

Saúl Calderón-Ramírez
sacalderon@itcr.ac.cr
Costa Rica Institute of Technology
Costa Rica

Sebastián Rojas González
sebastian.rojasgonzalez@ugent.be
Ghent University - IMEC
Belgium

Abstract

Monte carlo dropout is one of the most widely used methods for uncertainty quantification in deep learning, yet practitioners often rely on fixed heuristic settings for parameters such as the dropout rate and the ensemble size. Therefore, in this work, we formulate the joint tuning of dropout rate p, temperature T, and sample count N as a multi-objective optimisation problem that minimises expected calibration error and computational cost subject to an accuracy-retention constraint. We solve this problem with non-dominated sorting genetic algorithm II and reference vector guided evolutionary algorithm across three architectures—DenseNet121, MobileViT, and TinyViT—for binary diabetic retinopathy classification. Two main findings emerge, (1) calibration gains saturate at low sample counts ($N \approx 10$), enabling a 3× reduction in inference cost relative to common heuristics; and (2) optimal dropout rates are strongly architecture-dependent, with convolutional neural networks tolerating moderate perturbation ($p \approx 0.4$) while vision transformers require substantially lower rates ($p \approx 0.1$), challenging the widespread use of fixed Monte carlo dropout configurations. These findings demonstrate that evolutionary multi-objective optimisation can efficiently navigate the cost–calibration trade-off for uncertainty quantification in medical imaging.

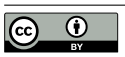
CCS Concepts

• **Computing methodologies** → **Scene anomaly detection**; **Neural networks**; **Computer vision problems**; **Genetic algorithms**.

Keywords

uncertainty quantification, calibration, multi-objective evolutionary optimisation, Monte Carlo Dropout, Temperature Scaling, deep learning

ACM Reference Format:
Miguel Abreu-Cardenas, Saúl Calderón-Ramírez, and Sebastián Rojas González. 2026. Evolutionary Multi-Objective Calibration of MC-Dropout Predictive Uncertainty Quantification for Diabetic Retinopathy. In *Genetic and Evolutionary Computation Conference (GECCO Companion '26), July 13–17, 2026, San Jose, Costa Rica.* ACM, New York, NY, USA, 4 pages. https://doi.org/10.1145/3795101.3805423



This work is licensed under a Creative Commons Attribution 4.0 International License.
GECCO Companion '26, San Jose, Costa Rica
© 2026 Copyright held by the owner/author(s).
ACM ISBN 979-8-4007-2488-6/2026/07
https://doi.org/10.1145/3795101.3805423

1 Introduction

Deep learning models used in ophthalmology can achieve expert-level accuracy in Diabetic Retinopathy (DR) screening [15], yet they are often poorly calibrated and prone to overconfident errors [6, 8]. For clinical deployment, reliable Uncertainty Quantification (UQ) is as important as predictive accuracy [12]. The literature commonly distinguishes between aleatoric uncertainty, caused by inherent data noise, and epistemic uncertainty, which reflects gaps in model knowledge [1]. The latter is especially relevant in medical imaging, where training datasets are often limited.

Monte Carlo Dropout (MCD) [7] approximates Bayesian inference by averaging N stochastic forward passes with dropout enabled at test time, yielding a predictive distribution that reflects epistemic uncertainty. Unlike Deep Ensembles, which require training multiple models and incur substantial memory overhead, MCD can be added to existing architectures without changing the training pipeline. However, it is still commonly used with fixed heuristic parameter settings that ignore both its interaction with post-hoc Temperature Scaling (TS) [8] and architecture-dependent sensitivity to dropout perturbations. In addition, the standard calibration metric, Expected Calibration Error (ECE) is non-differentiable because of its binning-based definition and is disconnected from proper scoring rules such as Negative Log-Likelihood (NLL) [2, 14], so minimising NLL does not necessarily minimise ECE.

To address this gap, we formulate the joint tuning of $\{p, T, N\}$ as a Constrained Multi-Objective Optimisation Problem (CMOP) and solve it with Non-dominated Sorting Genetic Algorithm II (NSGA-II) [5] and Reference Vector Guided Evolutionary Algorithm (RVEA) [4]. Although the search space has only three decision variables, naive enumeration is unattractive because the problem combines continuous and integer variables, stochastic Monte Carlo noise, and an accuracy-retention constraint. Evolutionary Multi-Objective Optimisation (EMO) algorithms therefore provide a natural fit through principled population management, constraint handling, and diversity preservation along the Pareto front.

Our main contribution is to show that. EMO-driven exploration of the UQ parameter space uncovers architecture-specific operating regimes for MCD that substantially reduce inference cost while preserving calibration quality, revealing behaviours that would remain hidden under fixed-parameter or single-objective approaches.

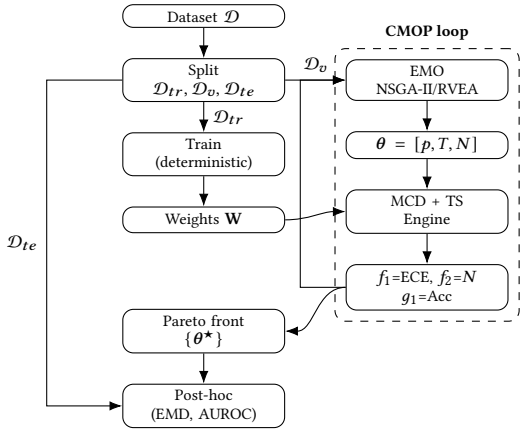


Figure 1: End-to-end pipeline of the proposed CMOP framework. The dataset is split for training, constrained multi-objective search on $\mathcal{D}_v$ (NSGA-II/RVEA with Latin Hypercube Sampling (LHS) initialisation), and held-out post-hoc diagnostics on $\mathcal{D}_{te}$. Each candidate $\theta = [p, T, N]$ is evaluated by the sample-wise calibrated MCD+TS engine to compute f_1=ECE, f_2=N, and the constraint g_1=Acc.

2 Proposed Method

We propose a multi-objective framework to configure a hybrid MCD + TS inference UQ engine. The method combines vectorised stochastic inference with evolutionary search to balance calibration error and computational cost.

2.1 CMOP Formulation

Let $\boldsymbol{\theta} = [p, T, N]$ denote the decision vector in search space Ω, where (i) the dropout rate is $p \in [0.1, 0.5]$, (ii) the temperature is $T \in [0.5, 3.0]$, and (iii) the number of Monte Carlo samples is $N \in \{10, \ldots, 50\}$. We minimise:

$$\min_{\boldsymbol{\theta}\in\Omega} \ \mathbf{F}(\boldsymbol{\theta}) = \big[f_1(\boldsymbol{\theta}),\ f_2(\boldsymbol{\theta})\big]^\top, \quad \text{s.t. } \ g_1(\boldsymbol{\theta}) \le 0, \tag{1}$$

where $f_1 = \mathrm{ECE}(\hat{P})$ is the calibration error of the averaged predictive distribution, $f_2 = N$ represents computational cost, and $g_1(\boldsymbol{\theta}) = \epsilon - \mathrm{Acc}(\boldsymbol{\theta})/\mathrm{Acc}_{\mathrm{base}}$ enforces an accuracy-retention constraint with ϵ=0.85, i.e., at most a 15% relative drop from the deterministic baseline.

2.2 Hybrid Inference Engine

Standard TS is usually applied to deterministic logits or to logits averaged after Bayesian integration. Here, we instead use a sample-wise calibrated MCD strategy to avoid damping the variance introduced by stochastic forward passes.

For a given $\boldsymbol{\theta}$, we run N stochastic forward passes with dropout rate p active and Batch Normalisation frozen. Each pass produces logits $\mathbf{z}_t = f_{\mathbf{W}_t}(\mathbf{x})$, where $\mathbf{W}_t$ denotes the weights after applying Bernoulli dropout masks. Temperature scaling is then applied to each logit vector *before* softmax:

$$\mathbf{q}_t = \mathrm{Softmax}\Big(\frac{\mathbf{z}_t}{T}\Big), \hat{P}(y|\mathbf{x}) \approx \frac{1}{N}\sum_{t=1}^{N} \mathbf{q}_t. \tag{2}$$

Calibrating before averaging preserves predictive variance, allowing T to correct overconfidence while p and N control the spread of the predictive distribution.

2.3 Calibration and Reliability Metrics

We optimise ECE directly as f_1. Although NLL is a proper scoring rule, minimising it does not guarantee smaller binning-based calibration gaps [14]. Following top-label calibration, for each sample $\mathbf{x}_i$ we define confidence as $\hat{c}_i = \max_k \hat{P}(y{=}k|\mathbf{x}_i)$, partition $[0, 1]$ into K=15 equal-width bins $\{B_k\}$, and compute:

$$\mathrm{ECE} = \sum_{k=1}^{K} \frac{|B_k|}{M}\big|\mathrm{acc}(B_k) - \mathrm{conf}(B_k)\big|, \tag{3}$$

where $\mathrm{acc}(B_k)$ and $\mathrm{conf}(B_k)$ are the empirical accuracy and mean confidence in bin B_k; empty bins contribute zero. As noted in prior work, ECE is sensitive to binning choices and may lack "non-robust completeness" [2, 3, 14].

As a post-hoc diagnostic of uncertainty *utility*, we measure the separability between the predictive entropy distributions of correct (P_C) and incorrect (P_I) predictions using the Earth Mover Distance (EMD):

$$W_1(P_C, P_I) = \inf_{\gamma\in\Pi(P_C,P_I)} \mathbb{E}_{(u,v)\sim\gamma}[\|u - v\|], \tag{4}$$

where $\Pi(\cdot,\cdot)$ is the set of couplings with prescribed marginals. Larger W_1 indicates better separation between correct and incorrect cases, complementing calibration by showing whether uncertainty concentrates on errors [16].

2.4 Evolutionary Search Strategy

The problem in Eq. (1) defines a non-convex, noisy landscape because of Monte Carlo approximation and the mixed continuous–integer search space. Continuous relaxation of integer variables can also create local plateaus [10]. We therefore use NSGA-II [5] as a Pareto-dominance baseline and RVEA [4] for reference-vector-guided diversity control.

The initial population P_0 is generated with LHS to cover Ω. Offspring are created using Simulated Binary Crossover (SBX) (p_c=0.9, η_c=15) and Polynomial Mutation (PM) (p_m=1/3, η_m=20). For the integer variable N, we apply continuous relaxation during variation and nearest-integer Baldwinian repair at evaluation time [17]. Constraints are handled with Deb's feasibility rules: feasible solutions dominate infeasible ones, and among infeasible solutions, lower constraint violation is preferred.

3 Experimental Setup

Dataset and architectures. We use the ODIR-5K dataset [11], reduced to binary classification (Normal vs. Diabetic Retinopathy; 4,642 fundus images, ≈60.7% Normal and 39.3% DR). This reduces confounding from the severe class imbalance present in other pathologies and helps isolate epistemic uncertainty. Images are resized to 224×224 and stratified into training (n=2,785), validation (n=928, used in the EMO loop), and test (n=929, held out for final evaluation). To cover different inductive biases, we evaluate **DenseNet121** [9] (CNN), **MobileViT** [13] (hybrid CNN–Transformer), and **TinyViT** [18] (Transformer). At inference, dropout is set to rate p and activated in `train()` mode, while

Batch Normalisation remains in `eval()` mode with frozen statistics to avoid corrupting normalisation estimates [7].

Optimiser configuration and validation protocol. NSGA-II and RVEA are implemented in `pymoo` [10] with population size 30. Because each evaluation requires costly MCD inference, the budget is limited to 1,200 function evaluations. Runs stop either at budget exhaustion or by hypervolume-based early stopping if Hypervolume (HV) improves by less than δ=10^{-4} for 5 consecutive generations. HV is computed with objective normalisation and a fixed reference point. Each algorithm is run with 15 independent seeds. Final populations are aggregated and filtered to obtain an approximate global Pareto front $\mathcal{P}^*_{\text{global}}$. Each selected configuration $\boldsymbol{\theta}^* \in \mathcal{P}^*_{\text{global}}$ is then re-evaluated 30 times on $\mathcal{D}_{te}$ to measure robustness to MCD stochasticity. We report mean±std of ECE, EMD, and Area Under the Receiver Operating Characteristic Curve (AUROC), and compare algorithms with two-sided Mann–Whitney U tests.

4 Results and Discussion

Figure 2 shows the approximate global Pareto fronts obtained by NSGA-II and RVEA for the three architectures, while Table 1 summarises selected non-dominated solutions from the 30-repetition stress test on $\mathcal{D}_{te}$. Together, these results support two main findings. First, calibration gains saturate quickly. Across all architectures, non-dominated solutions cluster near the lower bound of the search space (N≈10–16; Figure 2), indicating that ECE stabilises with relatively few stochastic samples when p and T are jointly tuned. This translates into a 3×–5× reduction in inference cost relative to common heuristics (N≥30, $p = 0.5$) while satisfying the accuracy-retention constraint. Second, Table 1 reveals a clear architectural pattern. DenseNet121 tolerates moderate dropout (p≈0.40–0.45), whereas TinyViT favours much lower rates (p≈0.11–0.16), with MobileViT lying in between. This supports the view that self-attention models are more sensitive to feature removal, so aggressive dropout can hurt predictive quality before uncertainty benefits appear.

Mann–Whitney U tests over the 30 stochastic repetitions (Table 1) show significant differences in EMD and AUROC between the selected NSGA-II and RVEA solutions for all three architectures ($p < 0.01$). ECE differences are significant for MobileViT and TinyViT ($p < 0.01$), but not for DenseNet121 ($p = 0.11$). Normalised HV exceeded 0.95 for all model–optimiser pairs within 10–25 generations, indicating robust convergence under the available budget. TinyViT showed the largest HV dispersion, consistent with a more rugged search landscape for attention-based models.

5 Conclusions, Limitations and Future Work

We presented an evolutionary multi-objective framework for automatically tuning MCD + TS parameters $\{p, T, N\}$ under a safety-oriented accuracy constraint, formulated as a constrained bi-objective problem minimising ECE and computational cost. Evaluating across three architectures (DenseNet121, MobileViT, TinyViT) with 15 independent seeds and 30-repetition stress tests, two empirical findings emerged: (1)calibration saturates at N≈10, enabling 3×–5× inference cost reduction relative to common heuristics; (2)optimal dropout rates are architecture-specific—Convolutional Neural Network (CNN) tolerate considerably more perturbation (p≈0.4) than Vision Transformers (p≈0.1), challenging the widespread practice of fixed-rate MCD. Additionally, preliminary evidence from the MobileViT experiments suggests that low ECE does not necessarily yield clinically useful uncertainty: one Pareto-optimal configuration achieved reasonable calibration yet showed substantially reduced separation between the uncertainty score distributions of correct and incorrect predictions (EMD drops from ∼0.223 to ∼0.076). While this observation is limited to a single architecture–algorithm pair and cannot be generalised from the current evidence, it motivates caution when relying solely on binning-based calibration for safety-critical UQ evaluation. Finally, we note some limitations and directions for future work. This study uses a single binary classification dataset and does not include a comparison against random search baselines, although the mixed-variable, noisy, constrained nature of the landscape favours EMO. Future work should explore: (i)multi-class and segmentation settings where spatial calibration becomes critical; (ii)alternative calibration estimators; (iii)incorporating the EMD directly as an optimisation objective in the CMOP formulation, transforming it from a post-hoc diagnostic into an active driver of the evolutionary search to ensure that Pareto-optimal solutions preserve meaningful separation between correct and incorrect uncertainty score distributions; and (iv) extending the framework to jointly tune architectural and training-time UQ mechanisms such as deep ensembles, conformal predictors, or test-time augmentation.

Acknowledgments

Computational support was provided by the Kabré supercomputer at the National High Technology Center of Costa Rica (CeNAT). Sebastián Rojas González acknowledges support by FWO Belgium (grant 1216021N).

References

[1] Moloud Abdar, Farhad Pourpanah, Sadiq Hussain, Dana Rezazadegan, Li Liu, Mohammad Ghavamzadeh, Paul Fieguth, Xiaochun Cao, Abbas Khosravi, U. Rajendra Acharya, Vladimir Makarenkov, and Saeid Nahavandi. 2021. A review of uncertainty quantification in deep learning: Techniques, applications and challenges. *Information Fusion* 76 (2021), 243–297. doi:10.1016/j.inffus.2021.05.008

[2] Jarosław Błasiok, Parikshit Gopalan, Lunjia Hu, and Preetum Nakkiran. 2023. A Unifying Theory of Distance from Calibration. In *Proceedings of the 55th Annual ACM Symposium on Theory of Computing* (Orlando, FL, USA) *(STOC 2023)*. Association for Computing Machinery, New York, NY, USA, 1727–1740. doi:10.1145/3564246.3585182

[3] Jarosław Błasiok, Parikshit Gopalan, Lunjia Hu, and Preetum Nakkiran. 2023. A unifying theory of distance from calibration. In *Proceedings of the 55th Annual ACM Symposium on Theory of Computing*. 1727–1740.

[4] Ran Cheng, Yaochu Jin, Markus Olhofer, and Bernhard Sendhoff. 2016. A Reference Vector Guided Evolutionary Algorithm for Many-Objective Optimization. *IEEE Transactions on Evolutionary Computation* 20, 5 (2016), 773–791. doi:10.1109/TEVC.2016.2519378

[5] K. Deb, A. Pratap, S. Agarwal, and T. Meyarivan. 2002. A fast and elitist multiobjective genetic algorithm: NSGA-II. *IEEE Transactions on Evolutionary Computation* 6, 2 (2002), 182–197. doi:10.1109/4235.996017

[6] Shahriar Faghani, Mana Moassefi, Pouria Rouzrokh, Bardia Khosravi, Francis I. Baffour, Michael D. Ringler, and Bradley J. Erickson. 2023. Quantifying Uncertainty in Deep Learning of Radiologic Images. *Radiology* 308, 2 (2023), e222217. arXiv:https://doi.org/10.1148/radiol.222217 doi:10.1148/radiol.222217 PMID: 37526541.

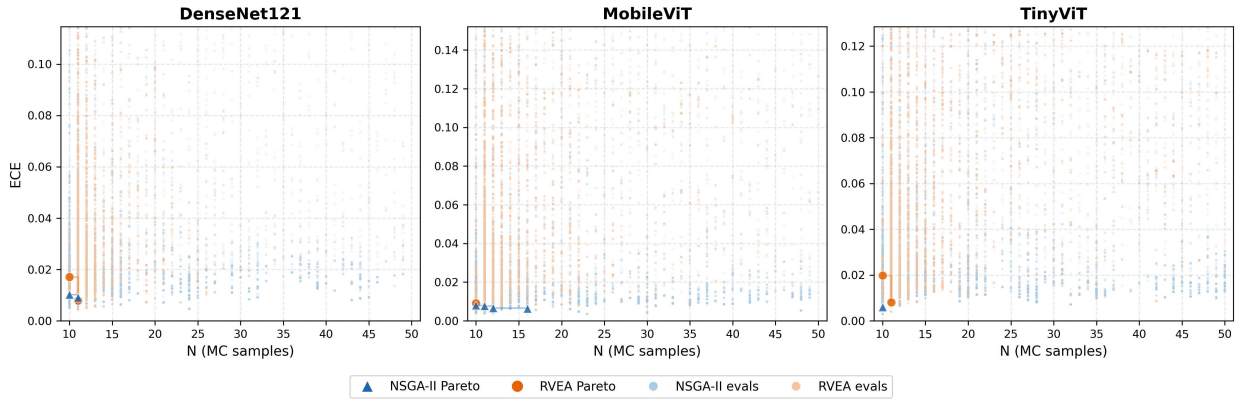


Figure 2: Global Pareto fronts for calibration–cost trade-offs. Each panel shows one architecture, all evaluated configurations, and Pareto-optimal solutions from NSGA-II (blue) and RVEA (orange) aggregated over 15 seeds. Grey points are dominated or infeasible. The fronts show diminishing ECE gains beyond $N \approx 10-16$, supporting low-cost operating regimes without major accuracy loss.

Table 1: Performance and stability of selected Pareto-optimal solutions: mean±std over 30 stochastic repetitions on $\mathcal{D}_{te}$.

Model	Algorithm	Config. (N, p, T)	ECE	EMD	AUROC
DenseNet121	NSGA-II	N=11, p=0.45, T=0.78	0.025 ± 0.006	0.194 ± 0.007	0.764 ± 0.011
	RVEA	N=11, p=0.40, T=0.82	0.023 ± 0.007	0.206 ± 0.007	0.779 ± 0.011
MobileViT	NSGA-II	N=12, p=0.12, T=0.86	0.016 ± 0.005	0.223 ± 0.004	0.796 ± 0.006
	RVEA	N=10, p=0.41, T=0.76	0.036 ± 0.008	0.076 ± 0.003	0.676 ± 0.010
TinyViT	NSGA-II	N=10, p=0.11, T=0.97	0.030 ± 0.004	0.143 ± 0.002	0.743 ± 0.005
	RVEA	N=11, p=0.16, T=0.81	0.019 ± 0.006	0.107 ± 0.003	0.698 ± 0.007

[7] Yarin Gal and Zoubin Ghahramani. 2016. Dropout as a Bayesian Approximation: Representing Model Uncertainty in Deep Learning. In *Proceedings of The 33rd International Conference on Machine Learning (Proceedings of Machine Learning Research, Vol. 48)*, Maria Florina Balcan and Kilian Q. Weinberger (Eds.). PMLR, New York, New York, USA, 1050–1059. https://proceedings.mlr.press/v48/gal16.html

[8] Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q. Weinberger. 2017. On Calibration of Modern Neural Networks. In *Proceedings of the 34th International Conference on Machine Learning (Proceedings of Machine Learning Research, Vol. 70)*, Doina Precup and Yee Whye Teh (Eds.). PMLR, 1321–1330. https://proceedings.mlr.press/v70/guo17a.html

[9] Gao Huang, Zhuang Liu, Laurens van der Maaten, and Kilian Q. Weinberger. 2017. Densely Connected Convolutional Networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

[10] Koki Ikeda and Isao Ono. 2023. Natural Evolution Strategy for Mixed-Integer Black-Box Optimization. In *Proceedings of the Genetic and Evolutionary Computation Conference* (Lisbon, Portugal) *(GECCO '23)*. Association for Computing Machinery, New York, NY, USA, 831–838. doi:10.1145/3583131.3590518

[11] Ning Li, Tao Li, Chunyu Hu, Kai Wang, and Hong Kang. 2020. A benchmark of ocular disease intelligent recognition: One shot for multi-disease detection. In *International symposium on benchmarking, measuring and optimization*. Springer, 177–193.

[12] Leopoldo Julian Lechuga Lopez, Shaza Elsharief, Dhiyaa Al Jorf, Firas Darwish, Congbo Ma, and Farah E. Shamout. 2025. Uncertainty Quantification for Machine Learning in Healthcare: A Survey. In *Proceedings of the sixth Conference on Health, Inference, and Learning (Proceedings of Machine Learning Research, Vol. 287)*, Xuhai Orson Xu, Edward Choi, Pankhuri Singhal, Walter Gerych, Shengpu Tang, Monica Agrawal, Adarsh Subbaswamy, Elena Sizikova, Jessilyn Dunn, Roxana Daneshjou, Tasmie Sarker, Matthew McDermott, and Irene Chen (Eds.). PMLR, 862–907. https://proceedings.mlr.press/v287/lopez25a.html

[13] Sachin Mehta and Mohammad Rastegari. 2022. MobileViT: Light-weight, General-purpose, and Mobile-friendly Vision Transformer. In *International Conference on Learning Representations*. https://openreview.net/forum?id=vh-0sUt8HlG

[14] Teodora Popordanoska, Raphael Sayer, and Matthew Blaschko. 2022. A Consistent and Differentiable Lp Canonical Calibration Error Estimator. In *Advances in Neural Information Processing Systems*, S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, and A. Oh (Eds.), Vol. 35. Curran Associates, Inc., 7933–7946. https://proceedings.neurips.cc/paper_files/paper/2022/file/33d6e648ee4fb24acec3a4bbcd4f001e-Paper-Conference.pdf

[15] Madhushan Ramalingam, Yaish Riaz, Priyanthi Rajamanoharan, and Piyumi Dasanayaka. 2025. Enhancing Safety in Diabetic Retinopathy Detection: Uncertainty-Aware Deep Learning Models with Rejection Capabilities. arXiv:2510.00029 [eess.IV] https://arxiv.org/abs/2510.00029

[16] Joachim Sicking, Maram Akila, Maximilian Pintz, Tim Wirtz, Stefan Wrobel, and Asja Fischer. 2024. Wasserstein dropout. *Machine Learning* 113, 5 (2024), 3161–3204.

[17] Ryoji Tanabe. 2024. Benchmarking Parameter Control Methods in Differential Evolution for Mixed-Integer Black-Box Optimization. In *Proceedings of the Genetic and Evolutionary Computation Conference* (Melbourne, VIC, Australia) *(GECCO '24)*. Association for Computing Machinery, New York, NY, USA, 712–721. doi:10.1145/3638529.3654019

[18] Kan Wu, Jinnian Zhang, Houwen Peng, Mengchen Liu, Bin Xiao, Jianlong Fu, and Lu Yuan. 2022. TinyViT: Fast Pretraining Distillation fornbsp;Small Vision Transformers. In *Computer Vision – ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part XXI* (Tel Aviv, Israel). Springer-Verlag, Berlin, Heidelberg, 68–85. doi:10.1007/978-3-031-19803-8_5

Received 21 January 2026; revised 12 March 2026; accepted 5 June 2026