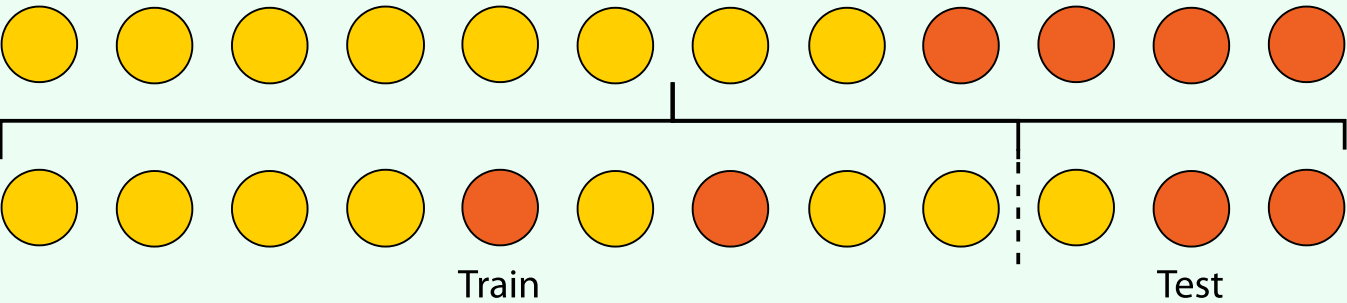


1. Background

- Train-Test Split is crucial:** Essential for testing models on **unseen data**.

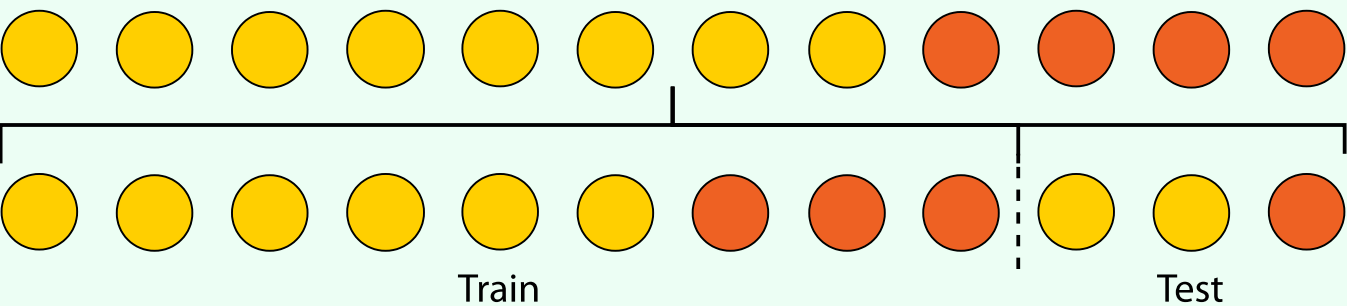
Random Split of Dataset



- Representative Test Set:** Ensures **accurate model evaluation** by mirroring the original data distribution.

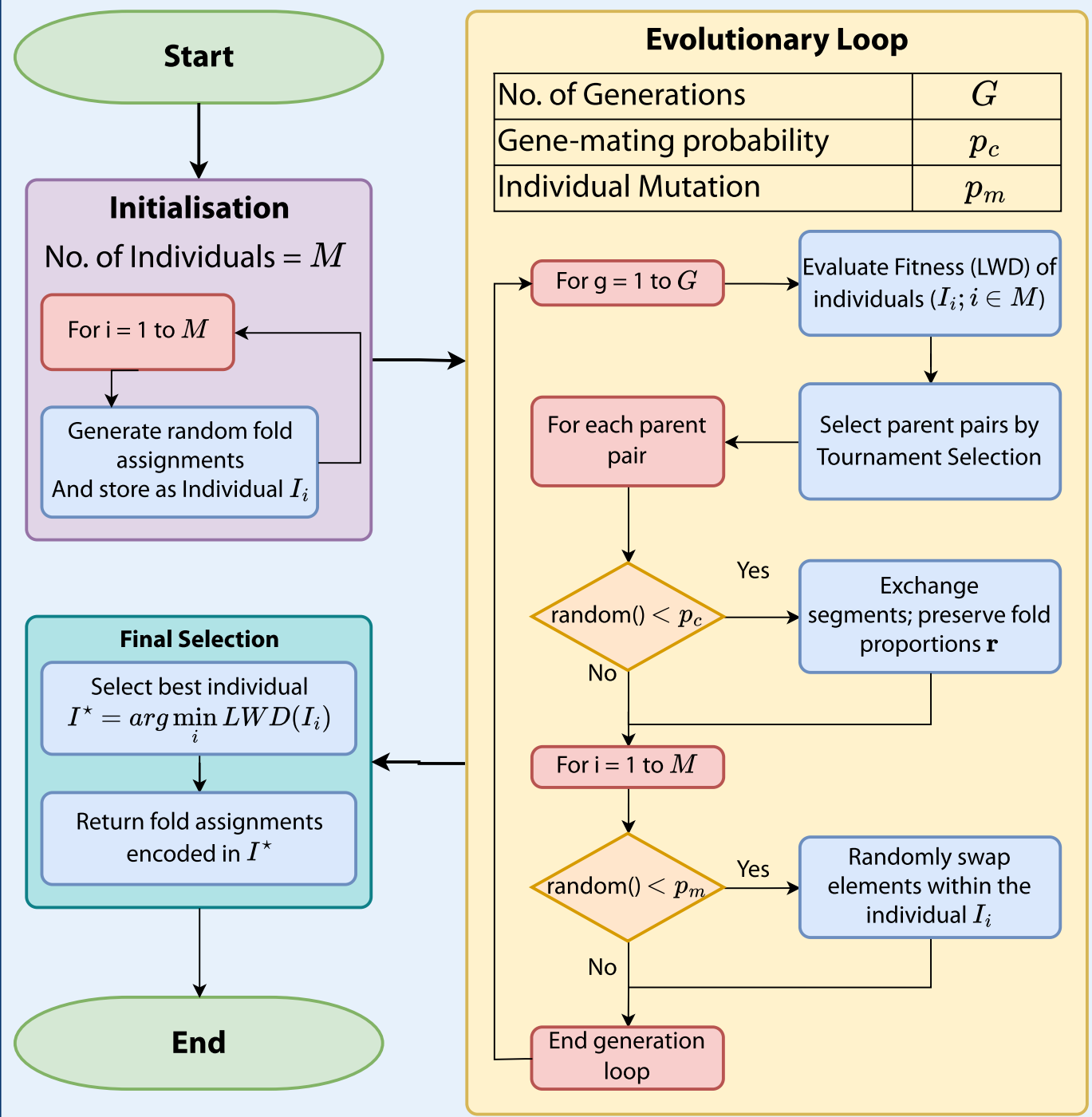
- Stratified Splitting:** Data splitting method based on the **ground truth** (class labels).

Stratified Split of Dataset



3. Wasserstein-Driven Evolutionary Stratification

Image segmentation dataset	$\mathbf{D} = \{\mathbf{x}_n, \mathbf{y}_n\}_{n=1}^N$
Total number of pixels in $\mathbf{D}$	P
Number of classes	C
K Disjoint Subsets	$\mathbf{S} = (S^1, \dots, S^K)$
Target proportion vector	$\mathbf{r} = (r^1, \dots, r^K)$



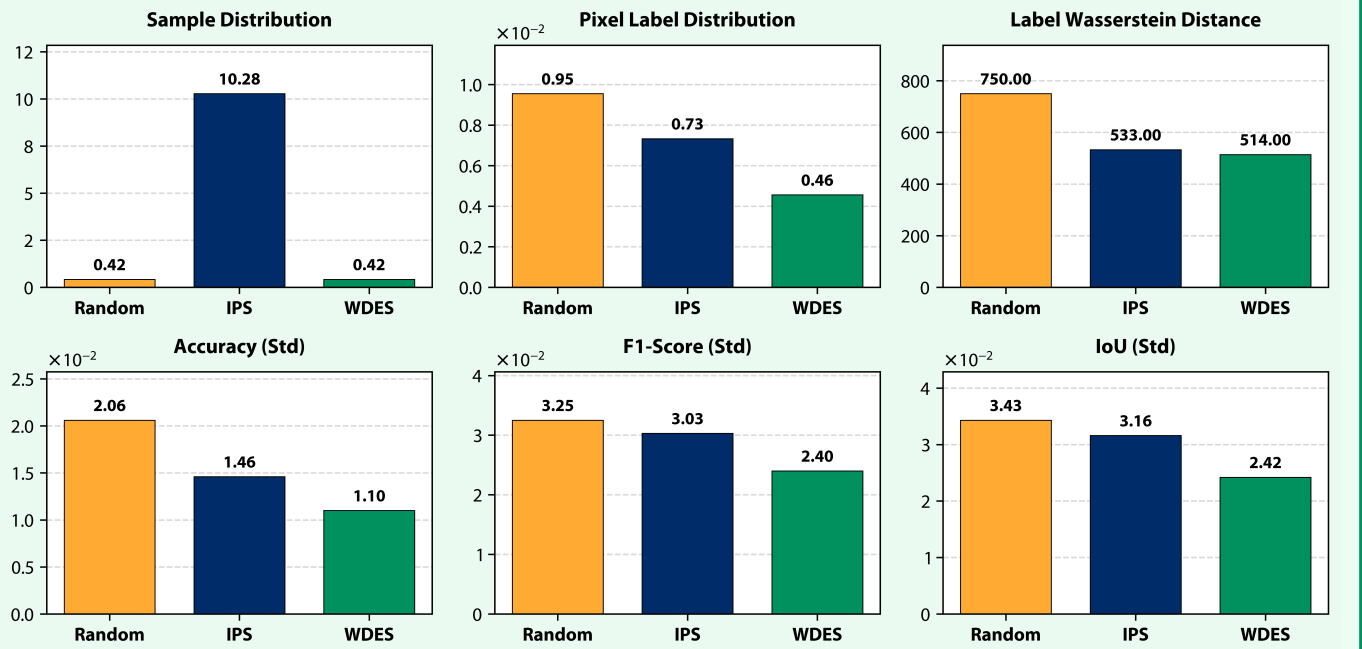
4. Evaluation Metrics

Sample Distribution	$SD = \frac{1}{K} \sum_{k=1}^K S^k - N^k $
Pixel Label Distribution	$PLD = \frac{1}{C} \sum_{c=1}^C \left(\frac{1}{K} \sum_{k=1}^K \left \frac{P_c^k}{P_c - P_c^k} - \frac{P_c}{P - P_c} \right \right)$
Label Wasserstein Distance	$LWD = \frac{1}{K} \sum_{k=1}^K \sum_{c=1}^C \left \sum_{i=1}^c P_i - \sum_{i=1}^c P_i^k \right $

5. Experiments and Results

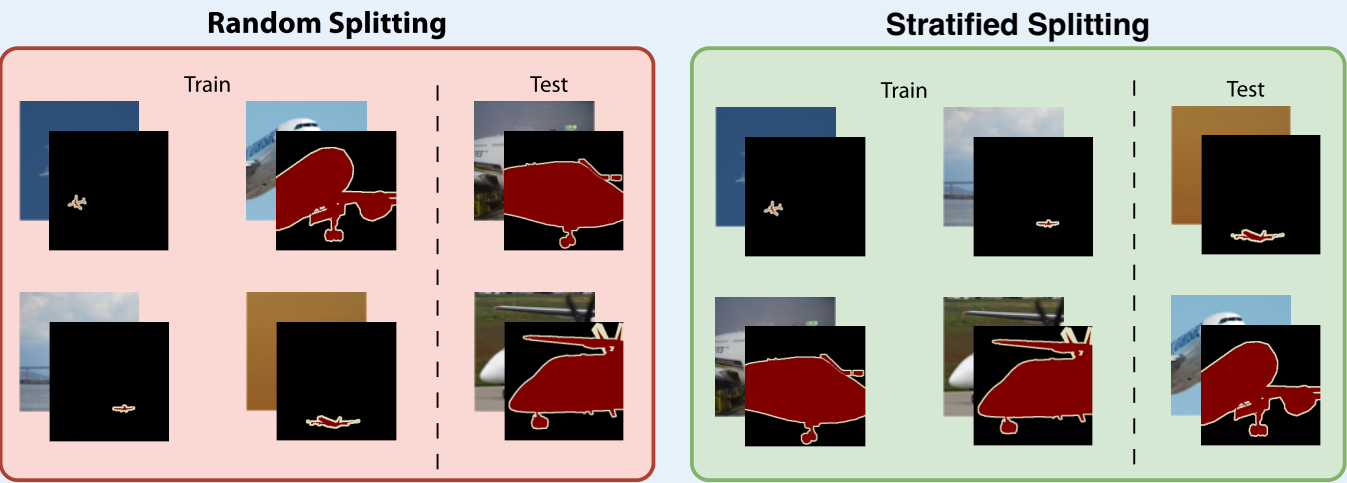
- Setup: 5-Fold Cross Validation (CV)** performed on **5 open-source segmentation datasets** (e.g., PascalVOC, Camvid, Endovis2018, LoveDA, Cityscapes).
- Methods Evaluated:** **Random**, **IPS** (iterative stratifier), and the proposed **WDES** splitting methods
- Evaluation:** Ranked methods using distribution metrics (**SD, PLD, LWD**) and model performance stability (std dev of **Accuracy, F1, IoU**).
- Key Result:** **WDES excels on low-entropy datasets** (e.g., PascalVOC, Camvid, Endovis2018), while **Random splitting is superior for high-entropy datasets** (e.g., LoveDA, Cityscapes).

Dataset	Method	SD	PLD	LWD	Acc (Std)	F1 (Std)	IoU (Std)
PascalVOC	Random	0.42	9.55E-03	7.50E+02	2.06E-02	3.25E-02	3.43E-02
	IPS	10.28	7.33E-03	5.33E+02	1.46E-02	3.03E-02	3.16E-02
	WDES	0.42	4.56E-03	5.14E+02	1.10E-02	2.40E-02	2.42E-02



2. Data Splitting in Image Segmentation

- Need for Stratification:** Essential for a **representative test set** in image segmentation.
- The Challenge:** Images contain **multiple classes with varying proportions** (not single labels).
- Risk of Random Split:** High risk of **skewing distribution** or **omitting rare classes** entirely in test set.
- Goal:** The ideal split must **maintain the original data distribution** within the test set.



6. Conclusion

- Problem:** Stratification of image segmentation datasets is **non-trivial**.
- Solution:** We developed **WDES**, an **evolutionary algorithm**, capable of stratifying data for simple **train/test splits** or **K-Fold CV**.
- Validation:** Compared WDES, Random, and IPS in a 5-Fold CV setup. **Lower performance SD implies a better stratifier**.
- Result Summary:** **WDES is superior for low-entropy datasets**. Random splitting remains effective for **high-entropy and very large datasets** (Law of Large Numbers).