



Stratify or Die: Rethinking Data Splits for Image Segmentation

Naga Venkata Sai Jitin Jami^{1,2}, Thomas Altstidl¹, Jonas Leo Mueller¹, Jindong Li¹, Dario Zanca¹, Bjoern Eskofier¹, Heike Leutheuser²

¹Machine Learning and Data Analytics Lab, FAU Erlangen-Nürnberg, Germany

²Ambient Assisted Living and Medical Assistance Systems, Department of Computer Science, University of Bayreuth, Germany



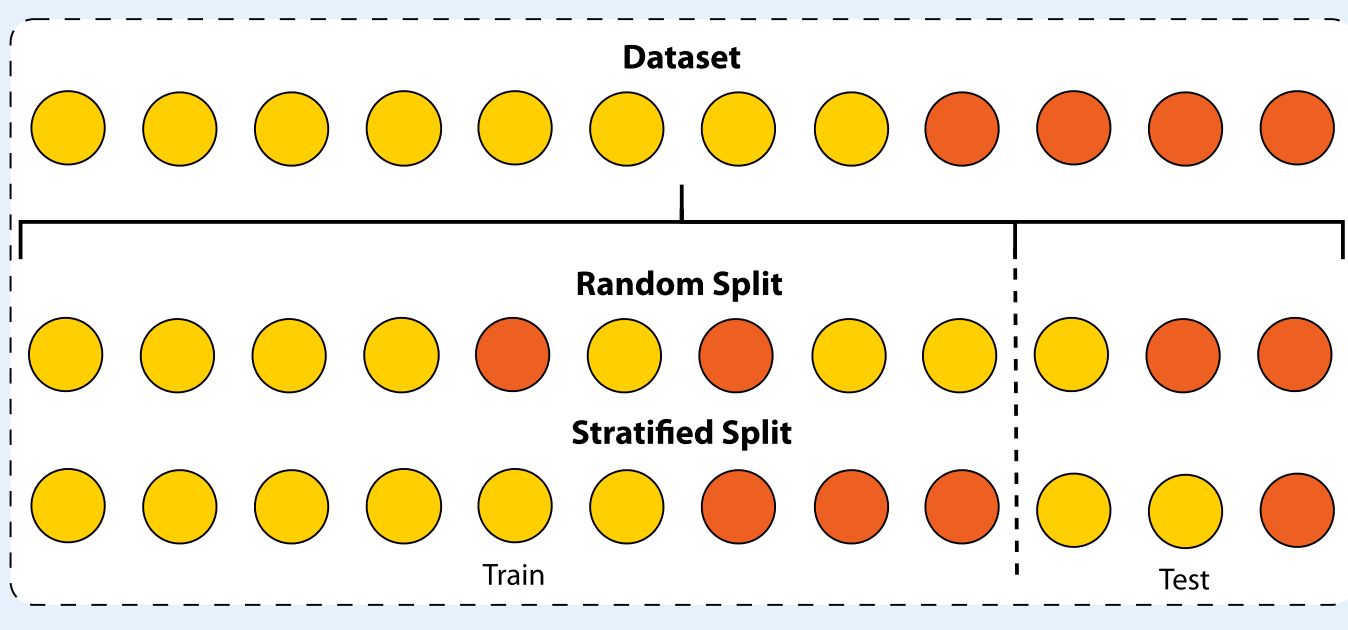
[Github](#)

1. Background

In Machine Learning, performing a **train-test split** of the dataset is crucial for **evaluation of a model on unseen data**.

Accurate evaluation can be carried out by evaluating the model on a **test set that is representative of the dataset**.

A representative test set can be created by **performing a stratified split** of the dataset **using ground truth labels**.



2. Data Splitting in Image Segmentation

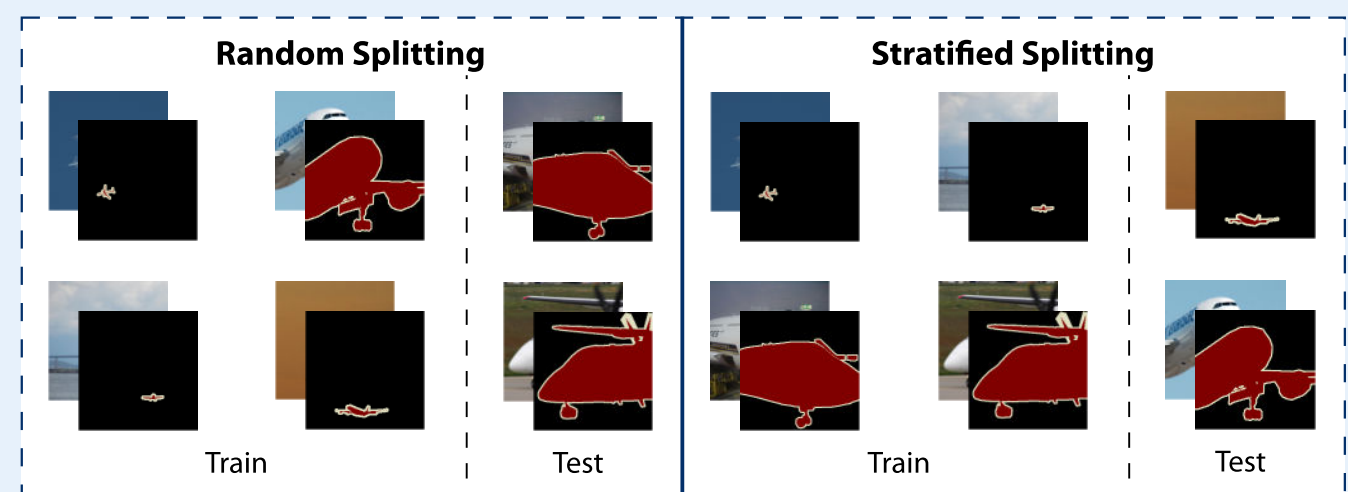
Image segmentation datasets are often **small and imbalanced**.

A simple **random train-test split** risks a **skewed test set** or **complete absence of rare classes** leading to **inaccurate model evaluation**.

Stratified splitting of the dataset is required for **accurate model evaluation**.

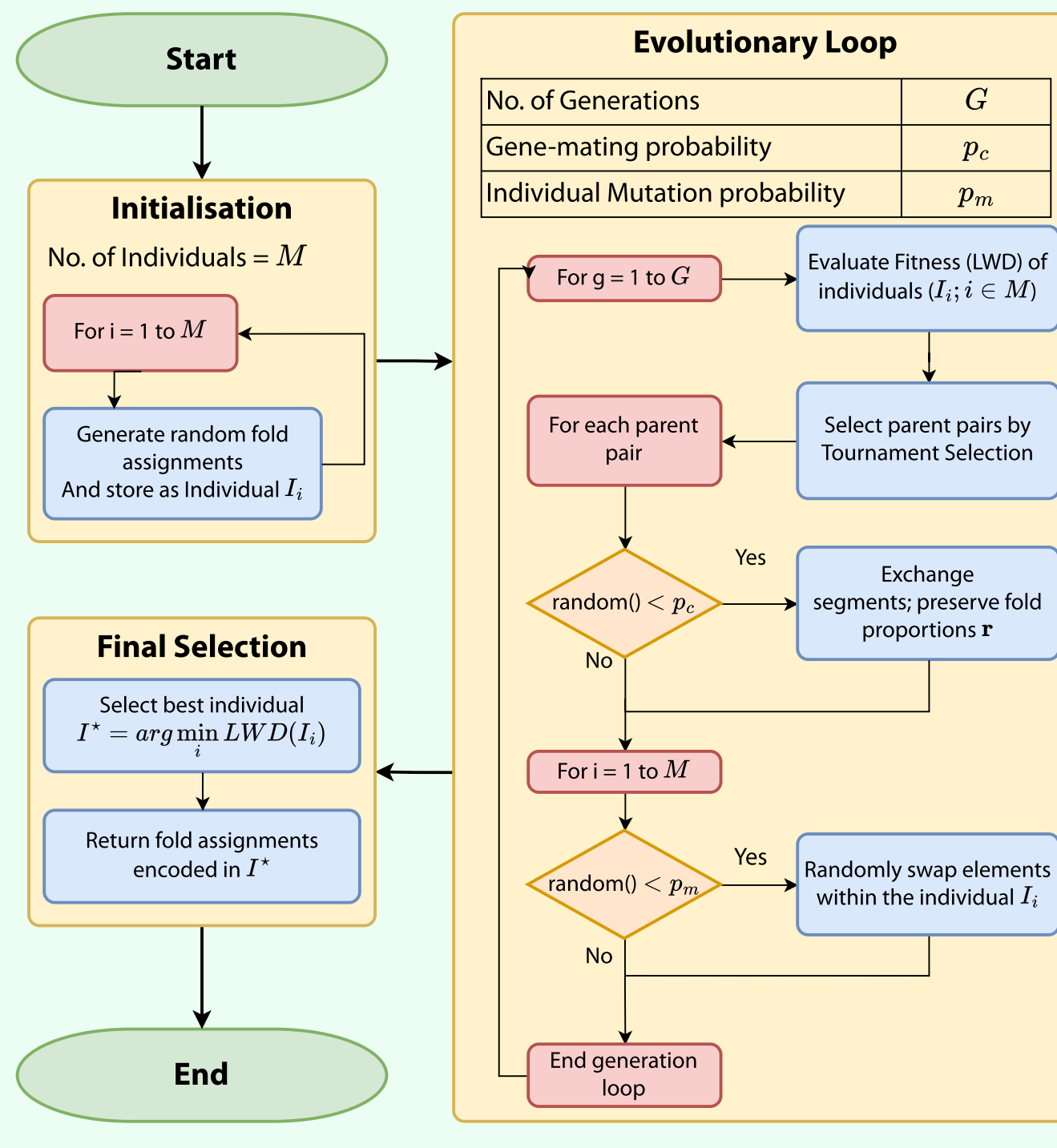
However, **stratification** of such datasets is **non-trivial** as an **image has multiple classes** with varying proportions.

The **ideal test set** must closely match the base dataset's **pixel-level data distribution**.



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	$PLD = \frac{1}{C} \sum_{c=1}^C \left(\frac{1}{K} \sum_{k=1}^K \frac{P_c^k}{P_c - P_c^k} - \frac{P_c}{P - P_c} \right)$
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 $

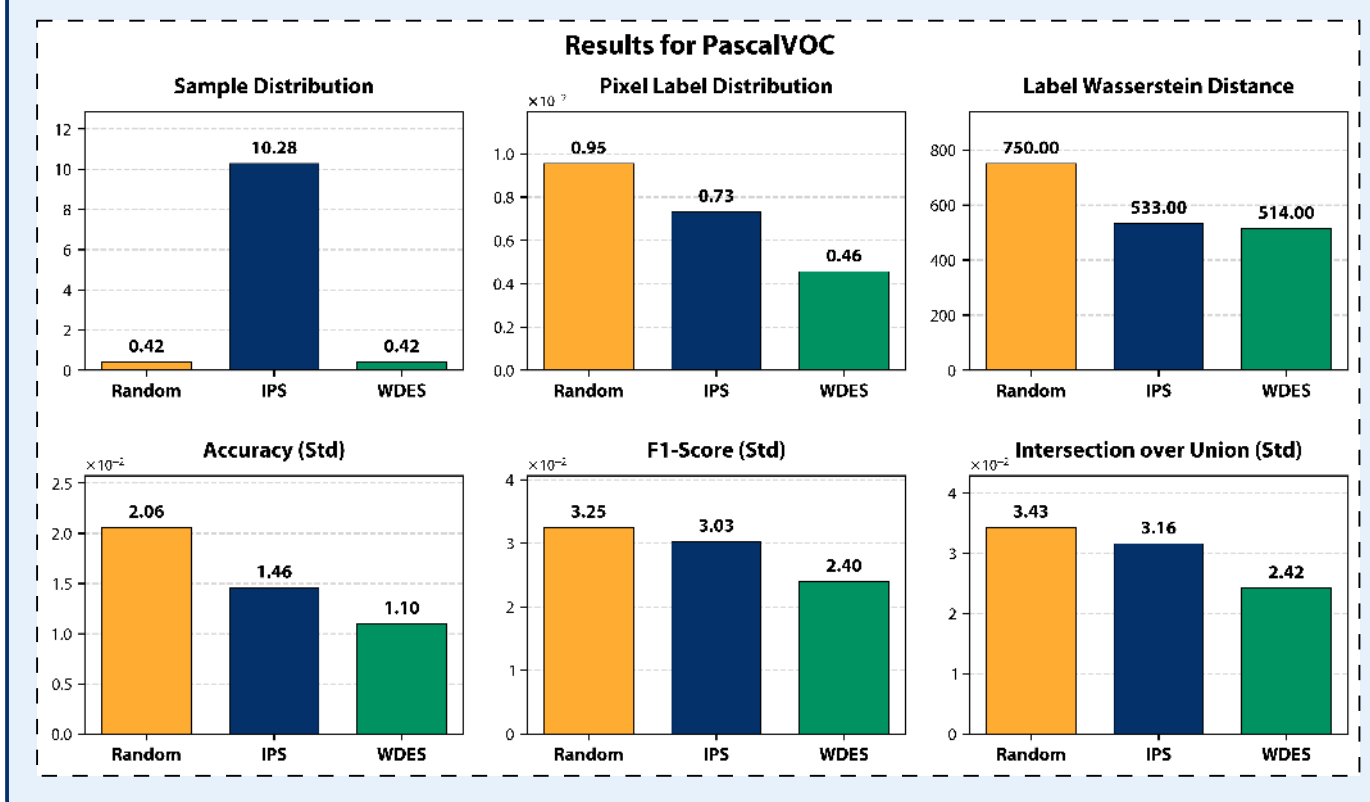
5. Experiments and Results

Performed a **5-Fold Cross Validation test** on **5 open-source segmentation datasets** (e.g., PascalVOC, Camvid, Endovis2018, LoveDA, Cityscapes) and train a **UNet** with **Resnet34** backbone.

Compared the performance of **Random**, **IPS** (a modified iterative stratifier)^[1], and the proposed **WDES** splitting methods.

Evaluated the above splitting methods by ranking distribution metrics (**SD**, **PLD**, **LWD**) and model performance stability (standard deviation of **Accuracy**, **F1**, **IoU** over 5 folds). Lower is better.

WDES excels on **smaller low-entropy high-diversity datasets** (e.g., PascalVOC, Camvid, Endovis2018), while **Random splitting is superior** for **larger high-entropy balanced-diversity datasets** (e.g., LoveDA, Cityscapes).



6. Conclusion

Stratification of Image Segmentation datasets is **non-trivial**.

We propose **WDES**, an algorithm that **optimises for closest subset distribution** using **Wasserstein distance** between subset and dataset

Targeted stratification is useful in **small, low-entropy, high-diversity datasets**.

For large, high-entropy, low-diversity datasets, random splitting is sufficient.

[1] Sechidis, Konstantinos, Grigorios Tsoumakas, and Ioannis Vlahavas. "On the stratification of multi-label data." Joint European conference on machine learning and knowledge discovery in databases. Berlin, Heidelberg: Springer Berlin Heidelberg, 2011.