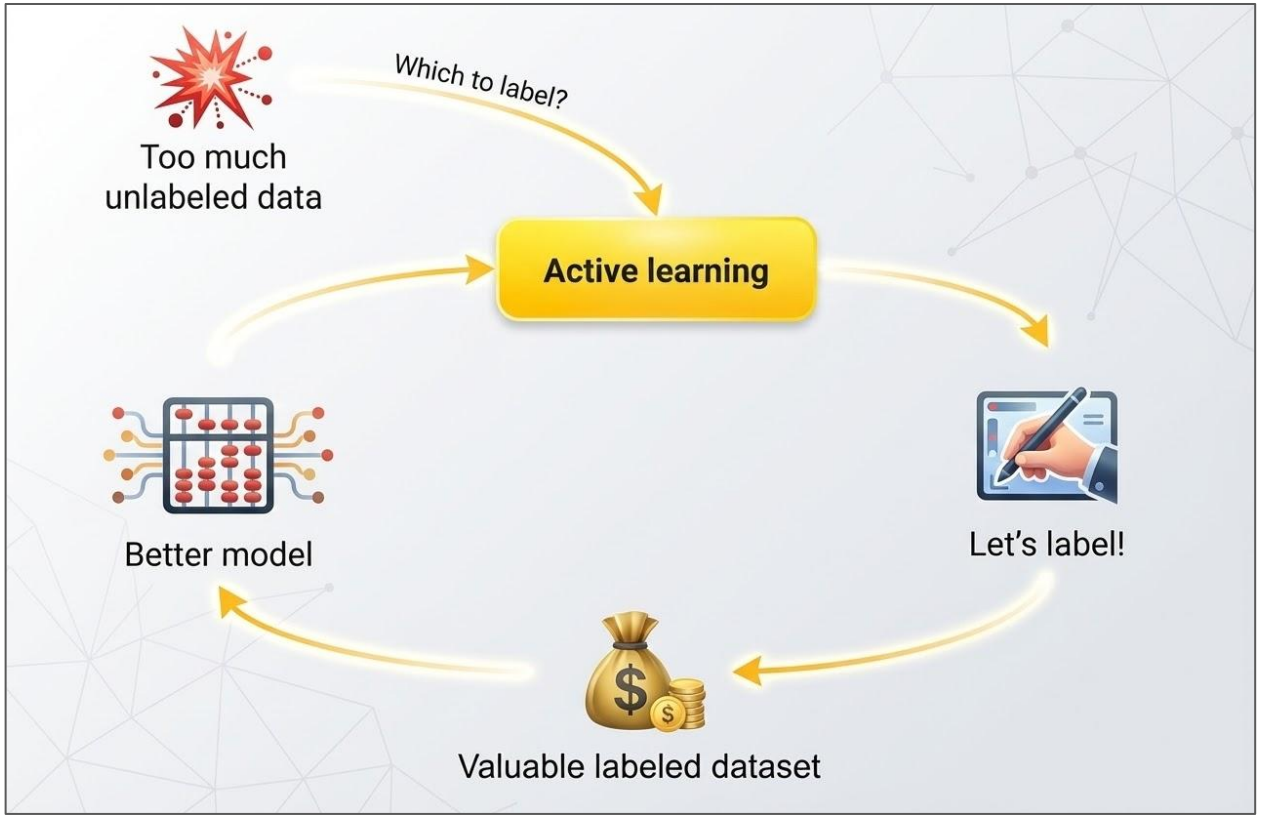
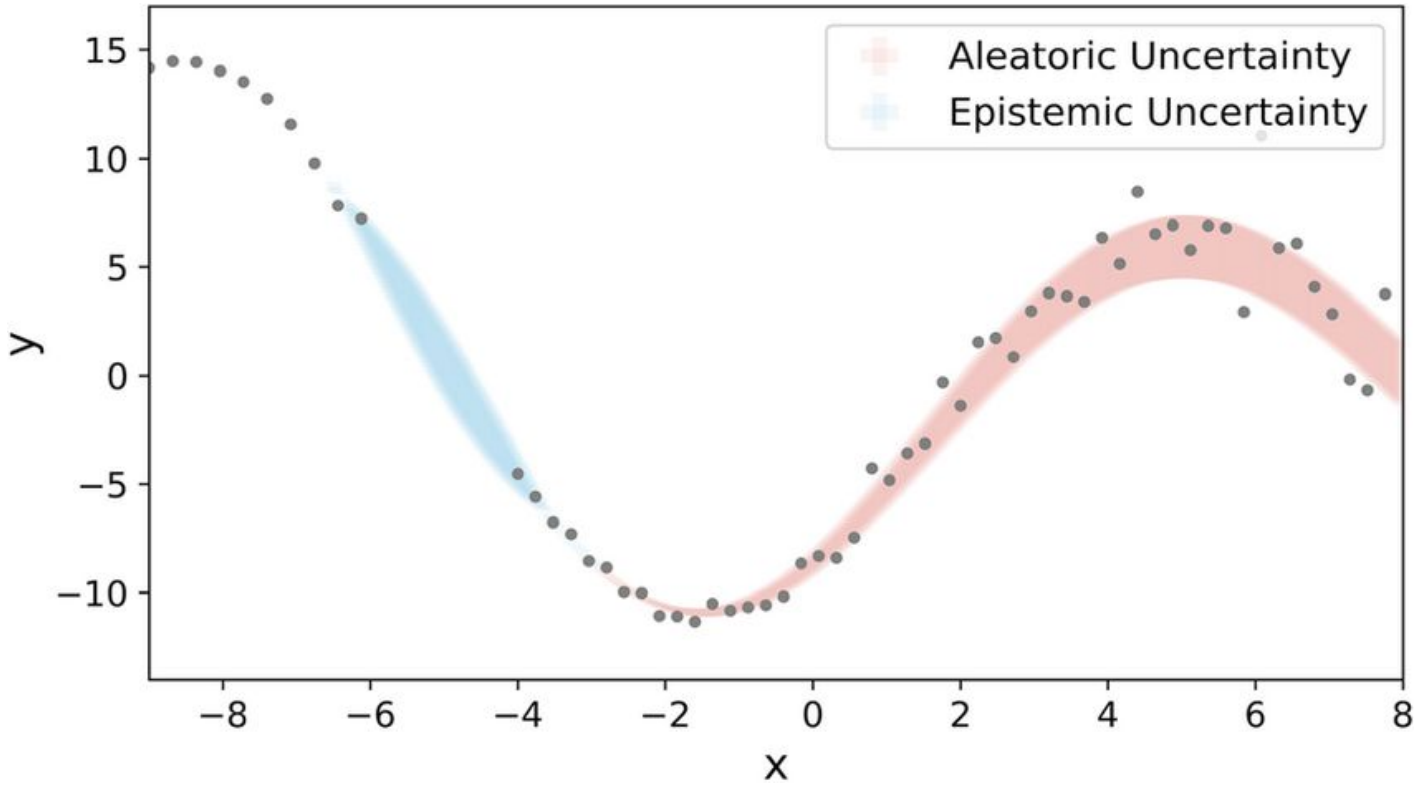
Enhancing Active-learning through Uncertainty-based Data Augmentation

Ehsan Garaaghaji, Farzad Hallaji, Elyar Esmaeilzadeh, Ida Fallah

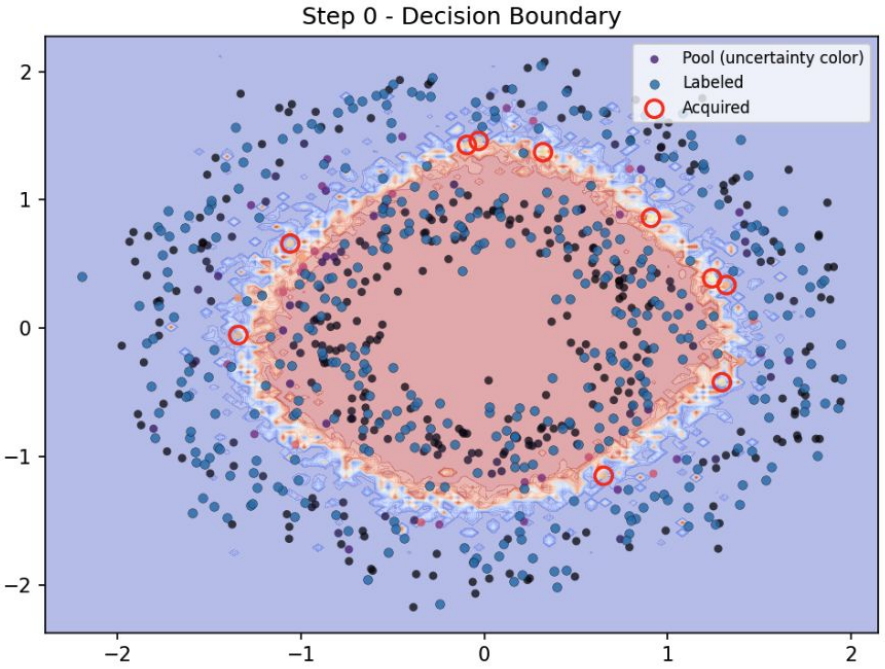
Active Learning

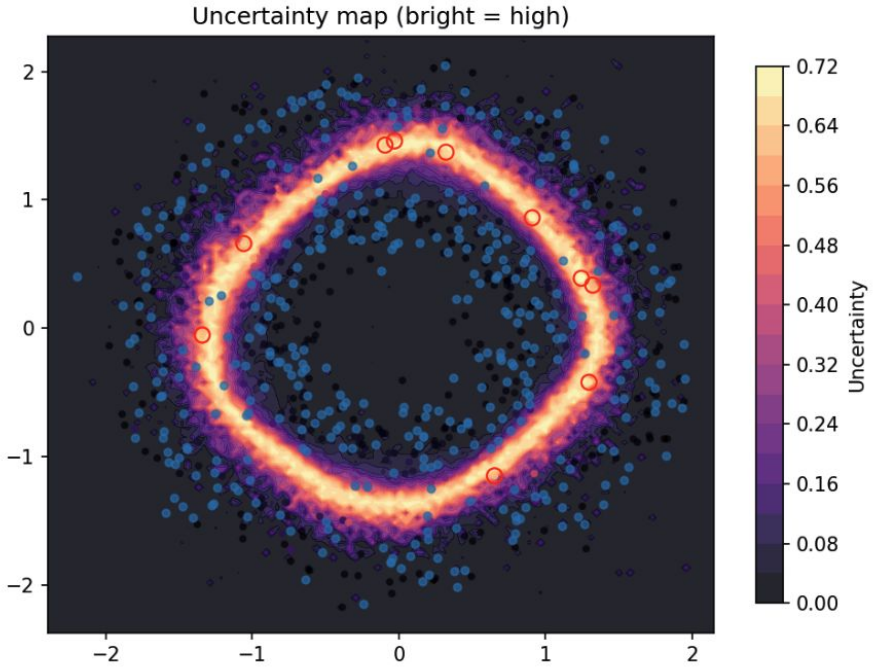


Uncertainty Estimation



Uncertainty Estimation in Active learning

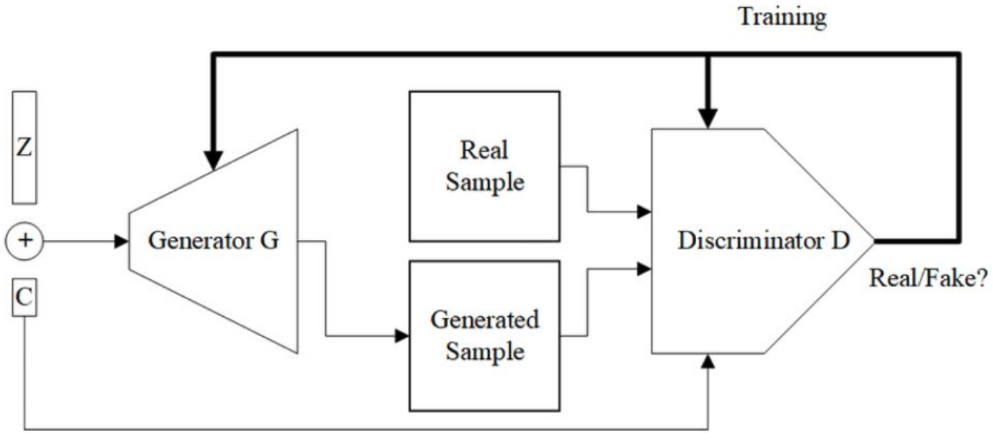




Methodology

- Uncertainty-Conditioned CGAN
- VAE with Latent Interpolation

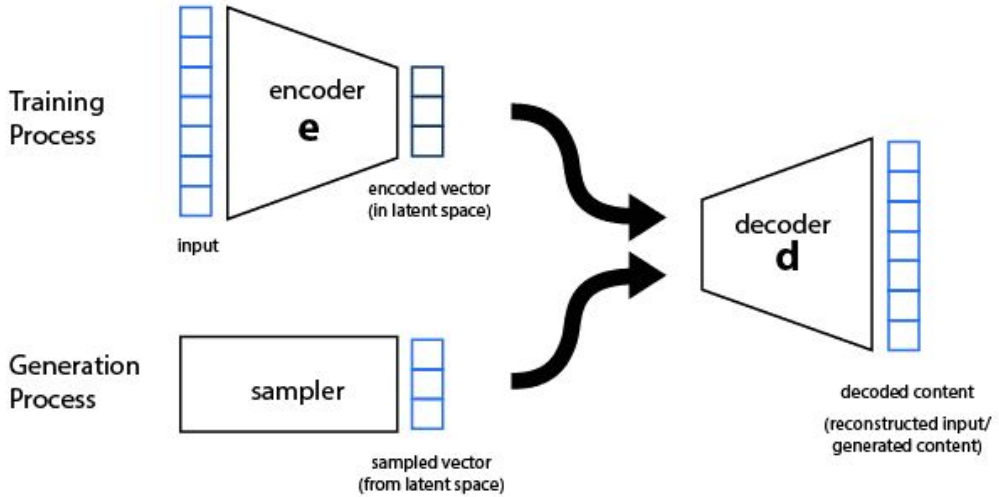
Uncertainty-Conditioned CGAN



Generator(z | c)

Discriminator(sample | c)

VAE Latent Interpolation

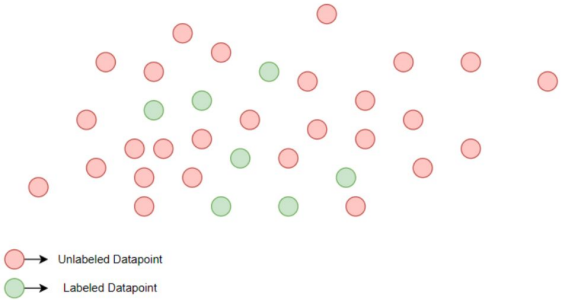


Apply **Farthest Point Sampling** in latent space to ensure diversity.

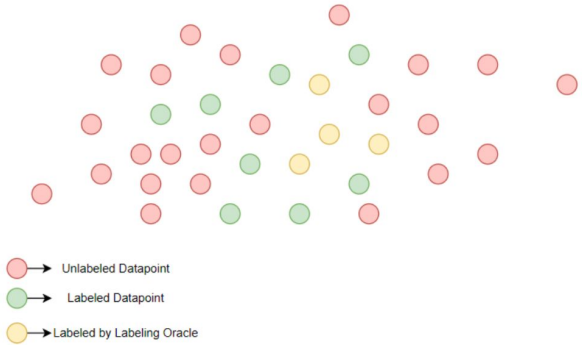
$$z_{new} = \alpha z_1 + (1 - \alpha) z_2$$

Pipeline

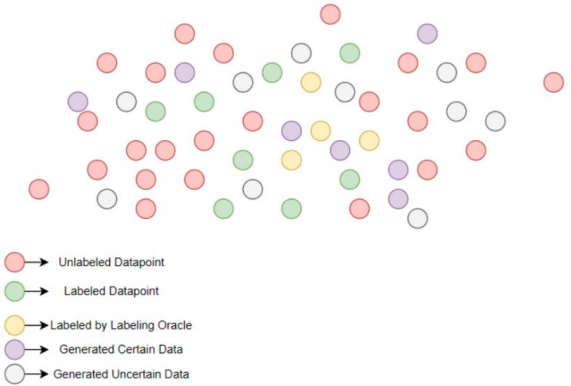
Step 1



Step 2



Step 3



Datasets

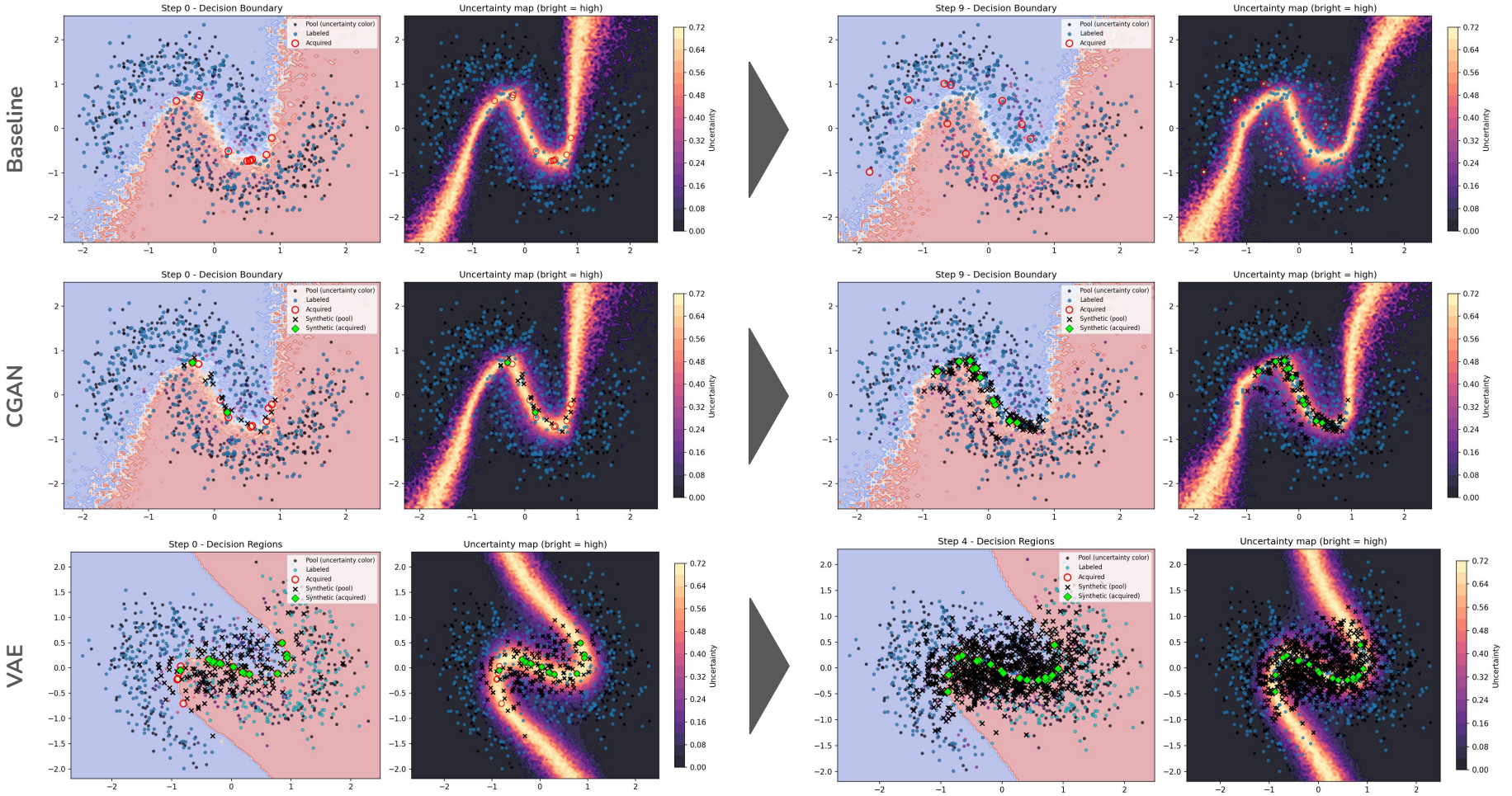
Classification: MNIST – Fashion-MNIST – CIFAR-10 – Breast Cancer – Wine – Iris – Two Moons – Circles

Regression: Boston/California Housing

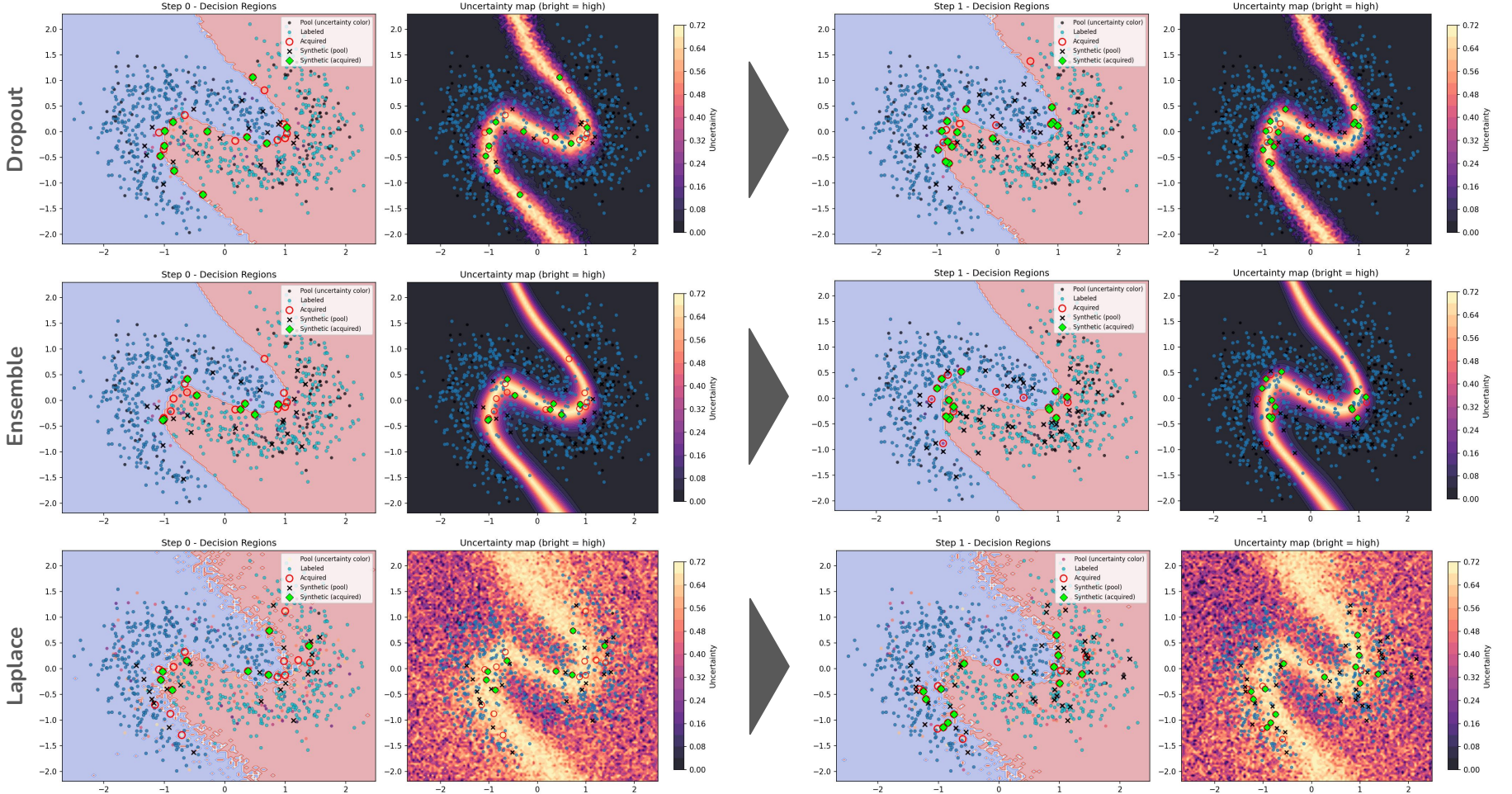
Ablations

Uncertainty estimation methods:	Dropout – Ensemble – Laplace
Data fraction:	0.1 – 0.3 – 0.5 – 0.8
Generation methods:	CGAN – VAE
Selection methods:	Top-k percent – Random – Uniform

CGAN vs. VAE

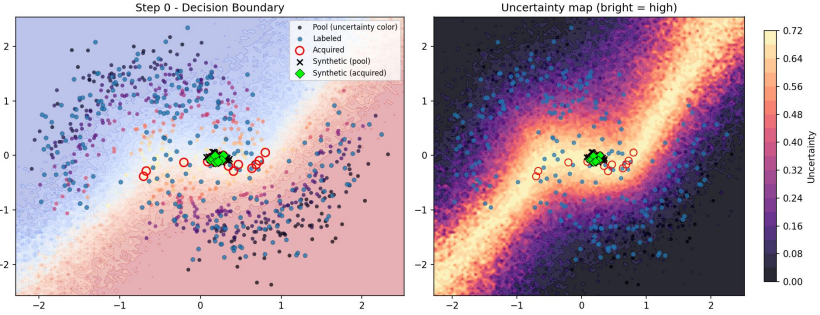

Baseline
CGAN
VAE
Step 0 - Decision Boundary
Step 9 - Decision Boundary
Step 0 - Decision Regions
Step 4 - Decision Regions
Uncertainty map (bright = high)
Pool (uncertainty color)
Labeled
Acquired
Synthetic (pool)
Synthetic (acquired)
Uncertainty

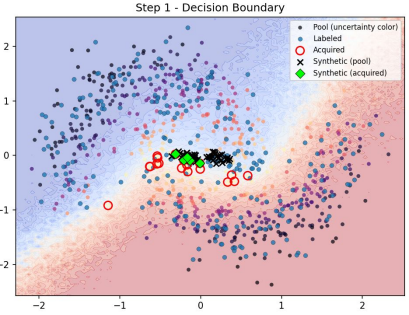
Uncertainty estimations methods (w/VAE & f = 0.8)

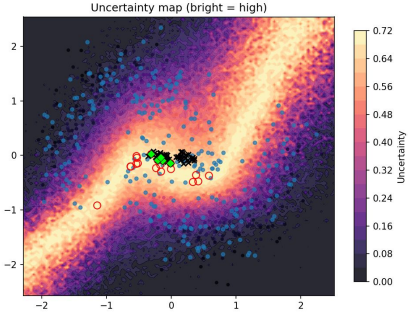


Uncertainty estimations methods (w/CGAN & f = 0.3)

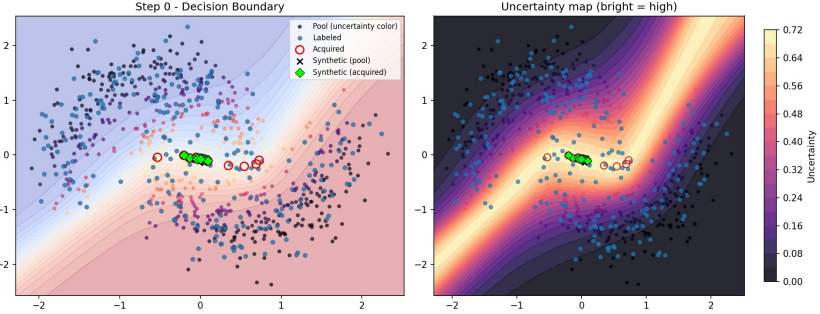
Dropout

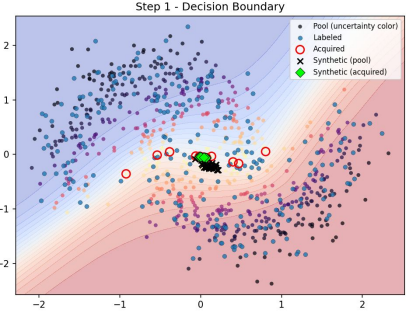


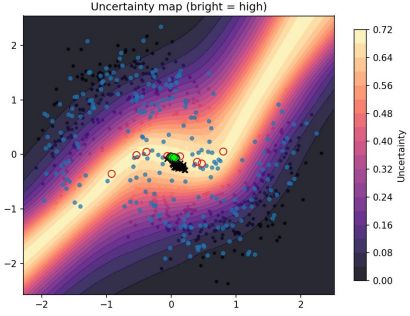




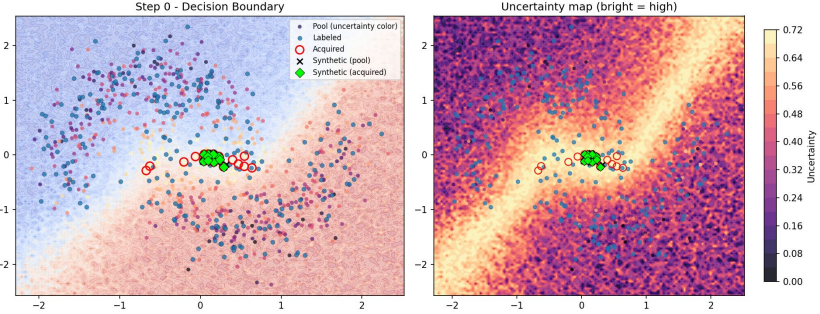
Ensemble

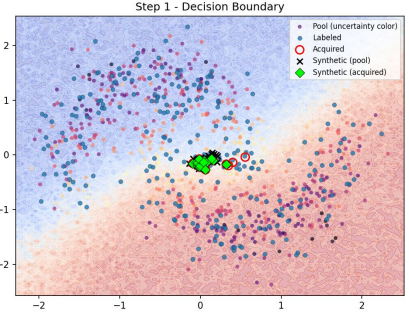


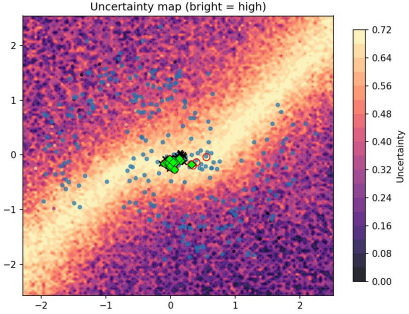




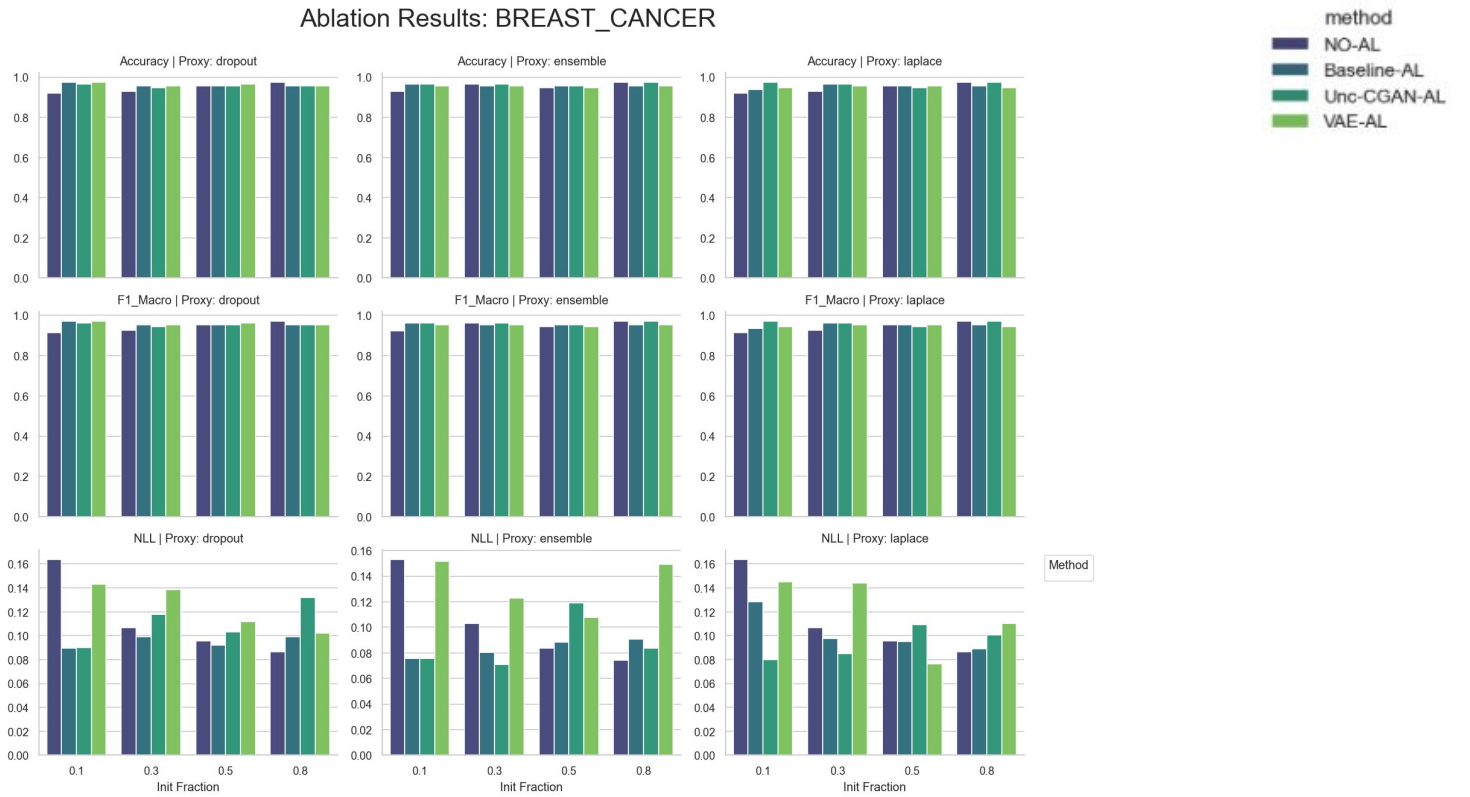
Laplace



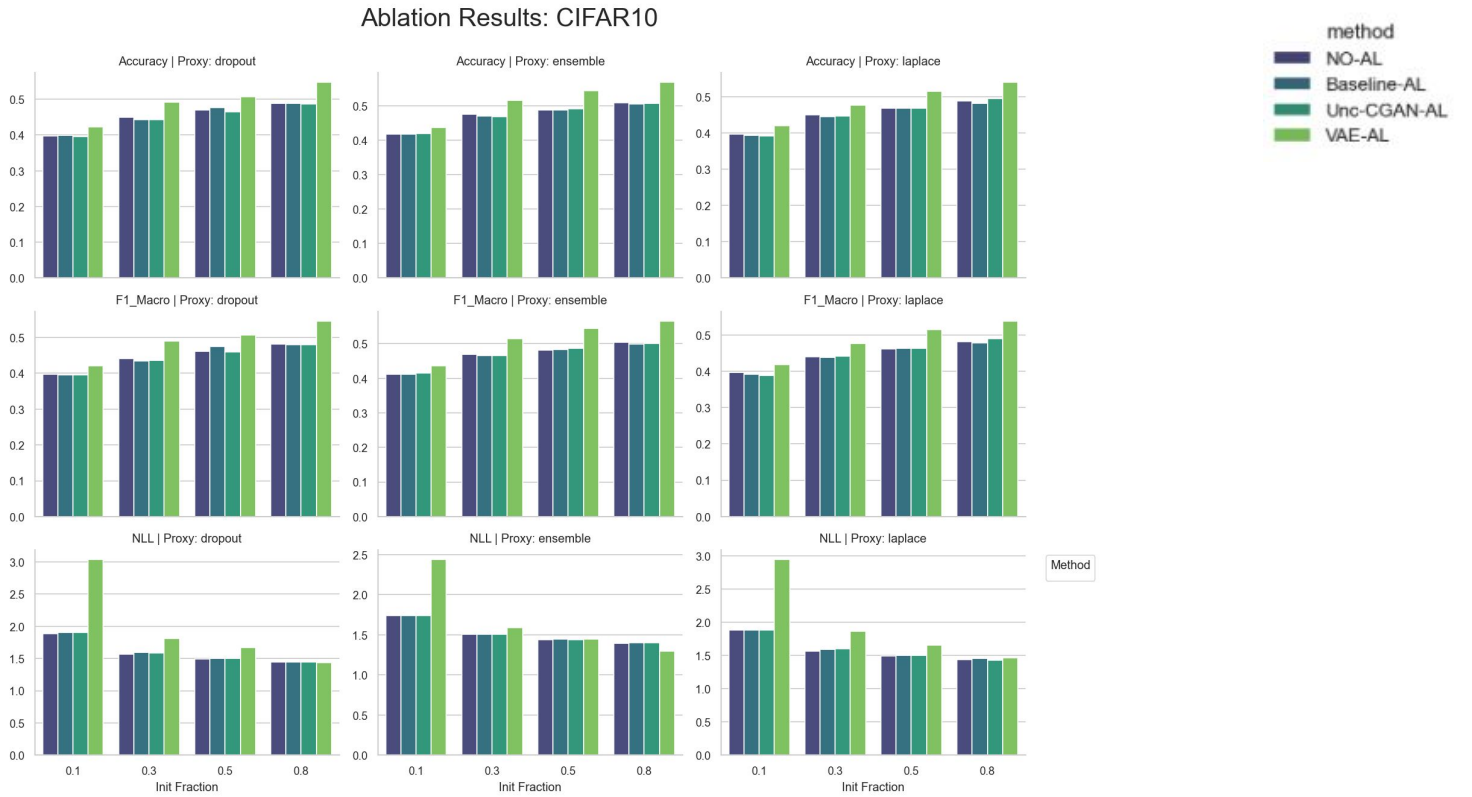




Ablation and Results

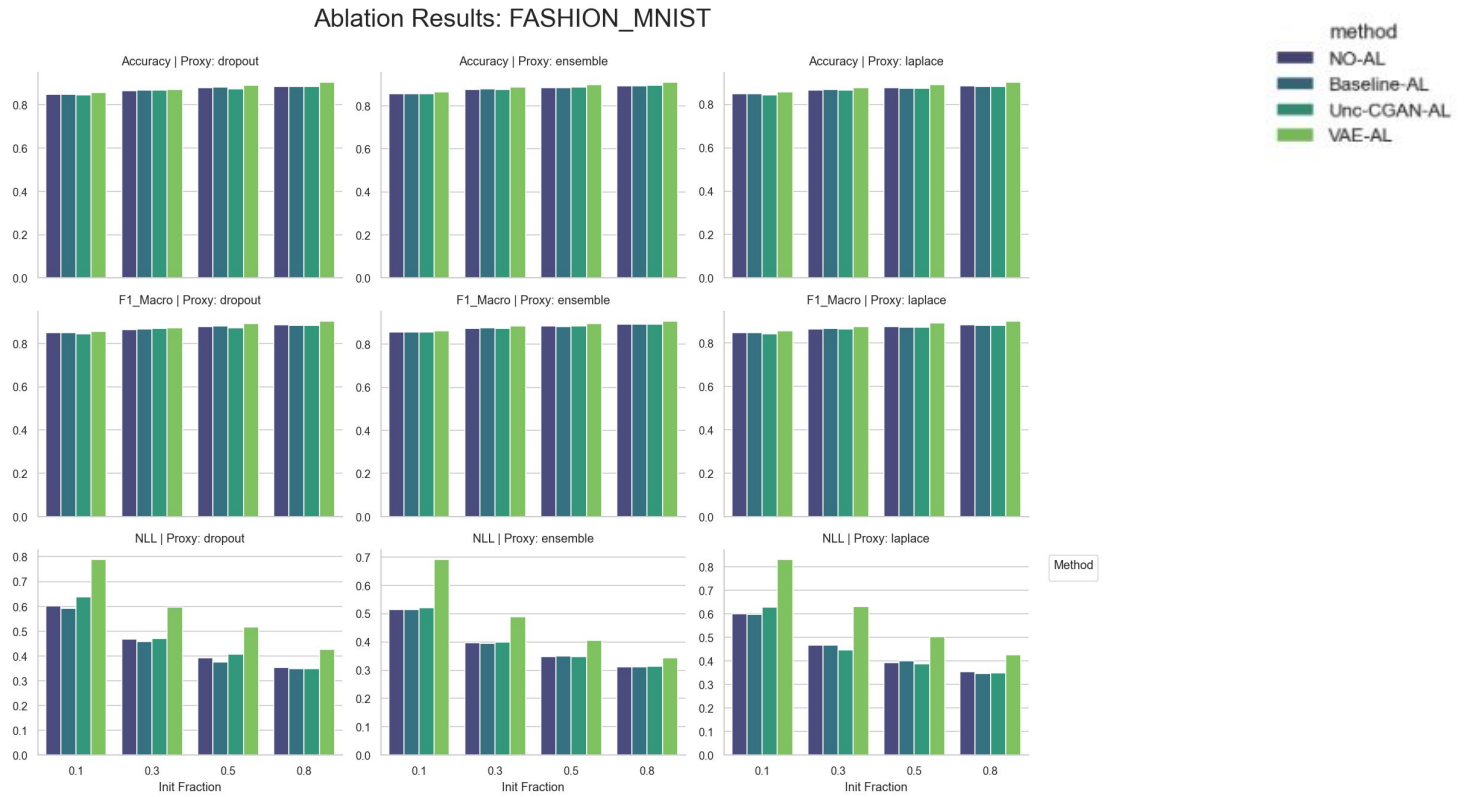


Ablation and Results

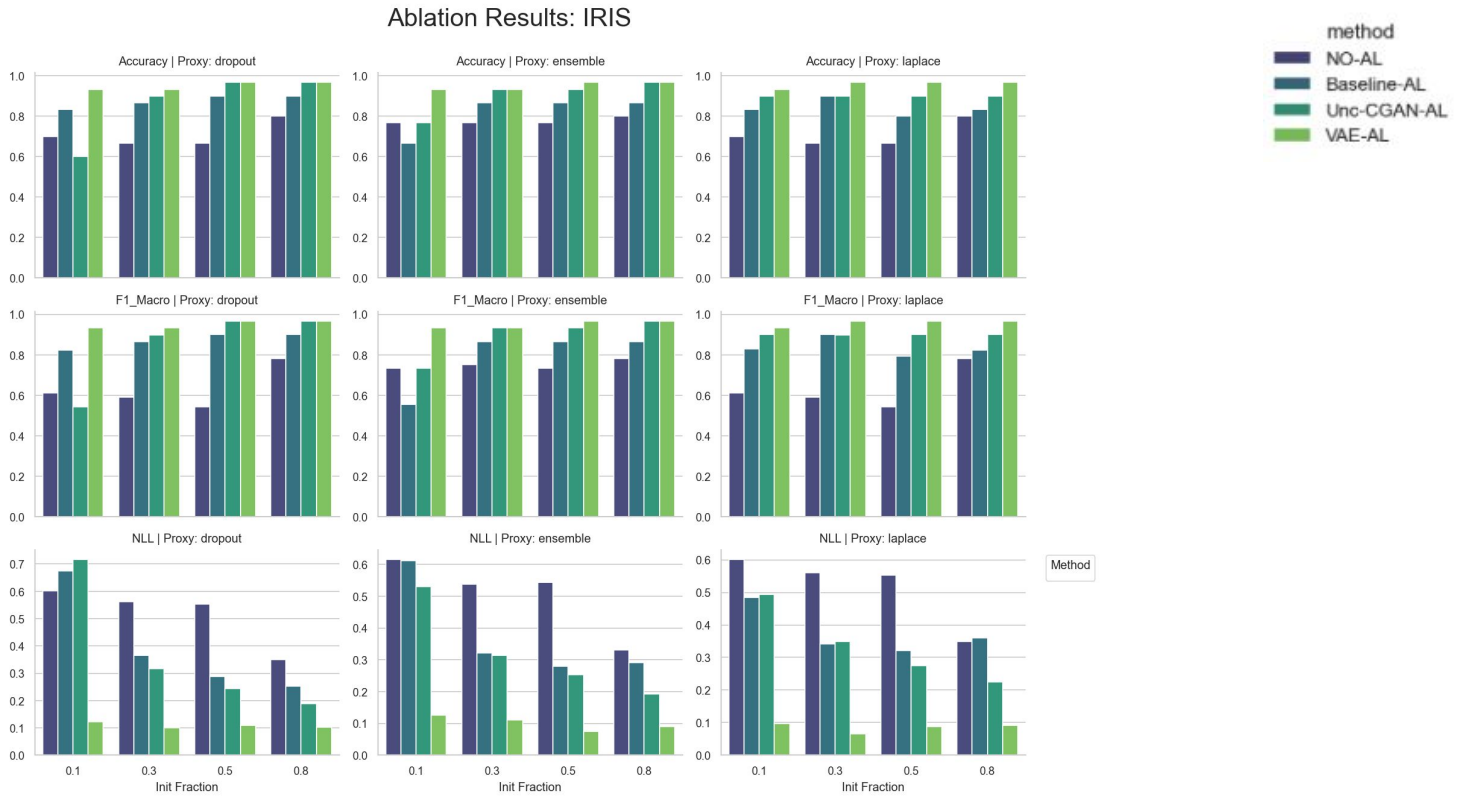


Ablation and Results

Ablation and Results

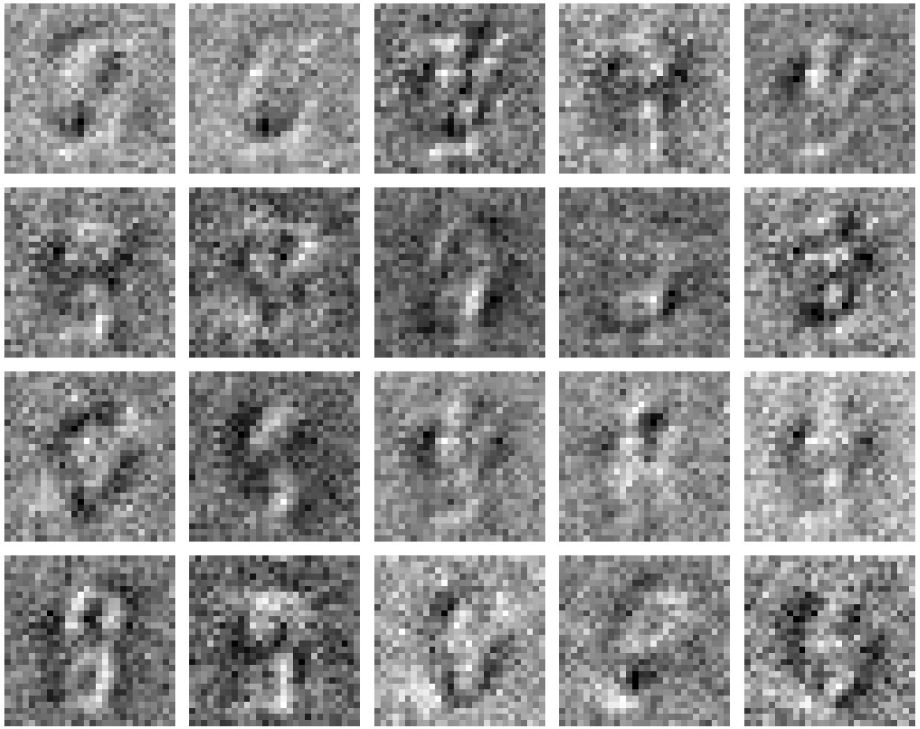


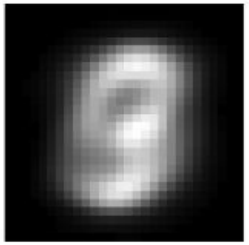
Ablation and Results



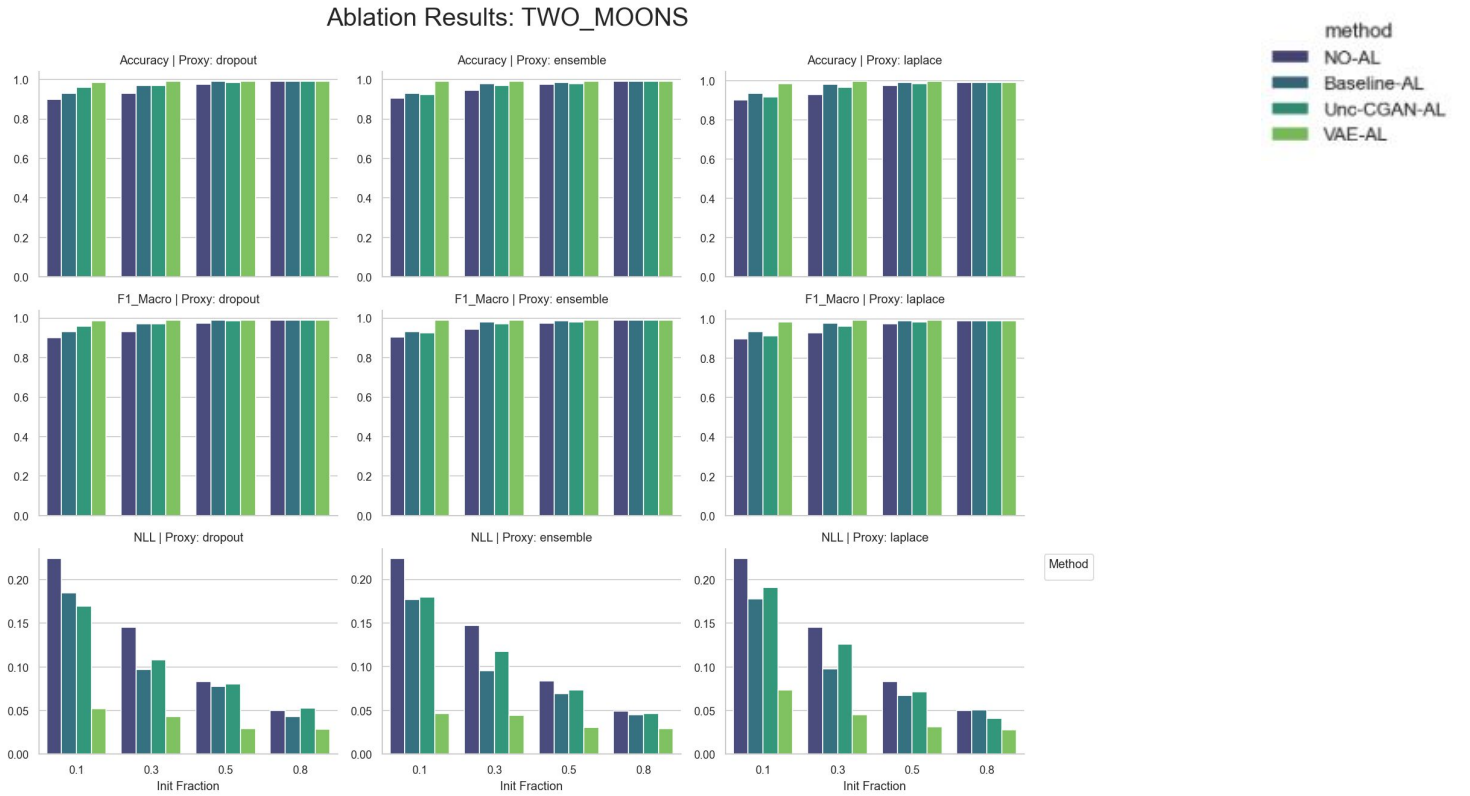
Ablation and Results

Generated samples for MNIST



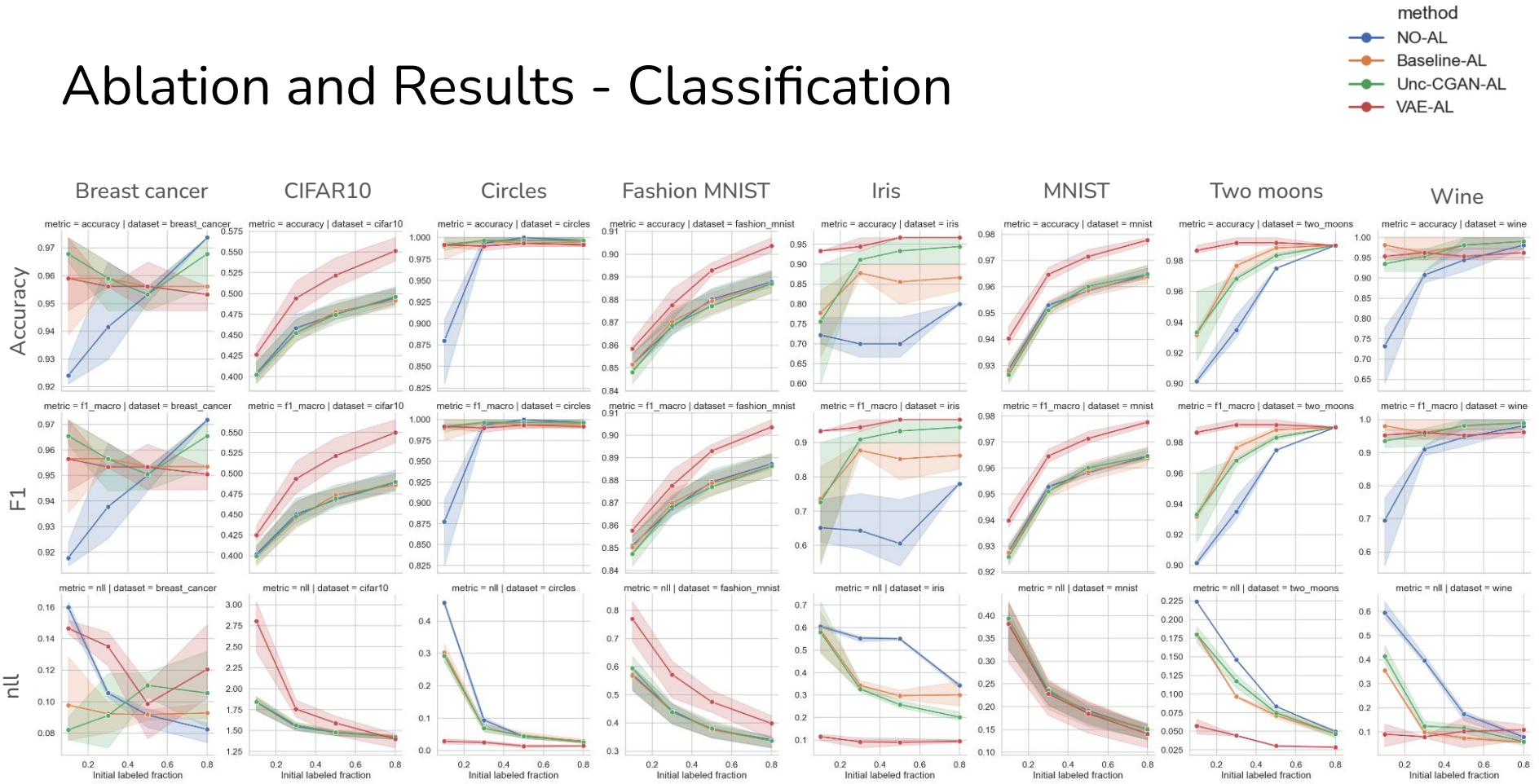


Ablation and Results

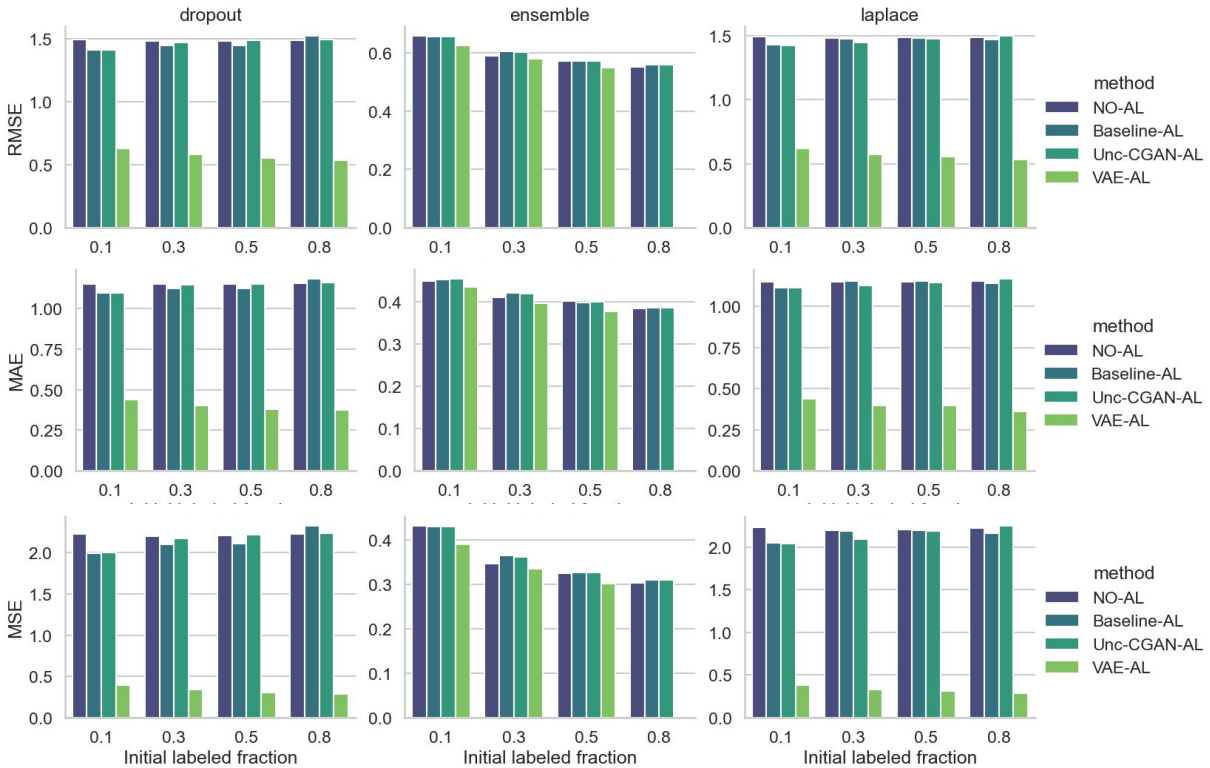


Ablation and Results

Ablation and Results - Classification



Ablation and Results - Regression



Comparison on complex datasets

Dataset	NO-AL (Acc)	Baseline-AL	Unc-CGAN-AL	**VAE-AL** (Best)
 MNIST	0.9302	0.9313	0.9308	**0.9450**
 Fashion-MNIST	0.8559	0.8565	0.8562	**0.8624**
 CIFAR-10	0.4175	0.4166	0.4186	**0.4368**

Recovery of Synthetic Manifolds

Dataset (finit=0.1)	NO-AL (Acc)	Baseline-AL	Unc-CGAN-AL	**VAE-AL**
Circles	0.9050	0.9750	0.9900	**0.9900**
Two Moons	0.9000	0.9300	0.9600	**0.9850**

Regression Performance (RMSE)

Dataset	f_{init}	NO-AL	Baseline-AL	Unc-CGAN-AL	VAE-AL (Best)
Boston Housing	0.10	1.4928	1.4115	1.4139	**0.6314**
	0.30	1.4827	1.4476	1.4734	**0.5851**
	0.50	1.4865	1.4504	1.4883	**0.5562**
	0.80	1.4906	1.5234	1.4934	**0.5337**

Discussion & Conclusion

Superior Efficiency:

VAE-AL consistently outperformed CGAN and baseline methods, especially on high-dimensional images and regression tasks.

Data Savings:

Performance with 10% labeled data using VAE-AL often matched or exceeded a standard model using 50% data.

Manifold Recovery:

Generative samples effectively fill gaps near decision boundaries, enabling near-perfect recovery of complex synthetic structures.