Hybrid Real–Synthetic Blastocyst Classification

Conditional GAN augmentation & SWIN transformer

Ioannis Kasionis & Nikolas Kavaklis

MSc in Artificial Intelligence – Deep Learning

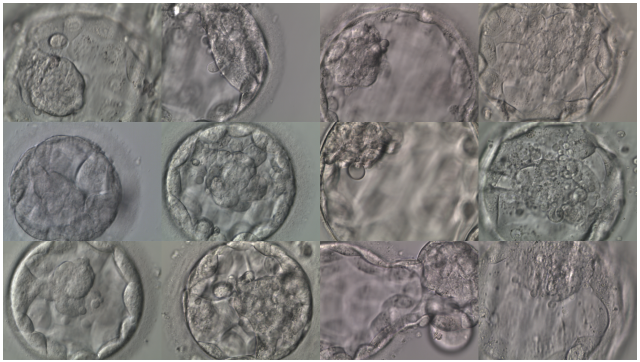
June 17, 2025

Agenda

Blastocyst Grading in IVF Clinics

Embryo selection in IVF relies heavily on morphological assessment of blastocyst images, scored by Gardner's criteria.

- Expansion(EXP): from 0 to 4
- Inner Cell Mass (ICM): from 0 to 3
- Trophectoderm (TE) Quality: from 0 to 3



Objectives

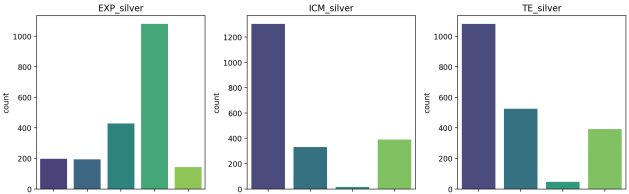
Manual scoring is subjective, time-consuming, inter-rater variance. Solution:

- Develop multiple classifiers to predict Gardner's EXP, ICM, and TE grades from single static images.
- Generate high-fidelity synthetic blastocyst images conditioned on all Gardner grades.
- Augment the real dataset with synthetic minority class samples and retrain classifiers to compare their performance against the original results.

Human Blastocyst Dataset

We used an annotated human blastocyst dataset to benchmark deep learning architectures from the paper of the same name.

- **Real silver**: 2 044 training images (512×384, single expert)
- **Gold test**: 279 images, consensus labels
- Severe class imbalance (below)



Baseline Models

From the paper that the dataset originates from, we have the following baselines:

	EXP	ICM	TE
mean±std experts	0.78±0.12	0.74±0.06	0.70±0.14
XCeption	0.78	0.69	0.62
Deit Transformer	0.82	0.63	0.58
Swin Transformer	0.85	0.65	0.62

Table: Accuracy for different methods across EXP, ICM, and TE

SWIN Transformer

SWIN is a type of Vision Transformer architecture designed for computer vision tasks like image classification, object detection, and segmentation. The Swin Transformer builds hierarchical feature maps much like Convolutional Neural Networks (CNNs).

- We used the swin base patch4 window7 224 from timm
- 12 Transformer blocks
- Patch size of 4x4
- Attention window of 7x7
- Image size of 224

Model Parameters

- **Optimizer and Scheduler**
  - We used `AdamW` to improve training stability and generalization by decoupling weight decay from the gradient update.
  - Benefit: Better generalization and more stable training.
  - Scheduler: Cosine LR Scheduler with warm-up.
- **Model and Training Parameters**
  - Model architecture: multiple SWIN sizes
  - Pretrained vs. training from scratch
  - Drop rate range
  - Model freeze strategies (freeze none, freezing embedding, freezing everything except classifier, freezing everything)
  - Loss: CrossEntropy with or without class weights
- **Training Procedure**
  - Learning rate
  - Weight decay
  - Warm-up for LR Scheduler
  - Split into training and validation sets
  - Validation used only real images

SWIN Real Data Results

For all Criteria we used the same parameters, as we found that they performed the best in all cases, our training was close to or above the baseline.

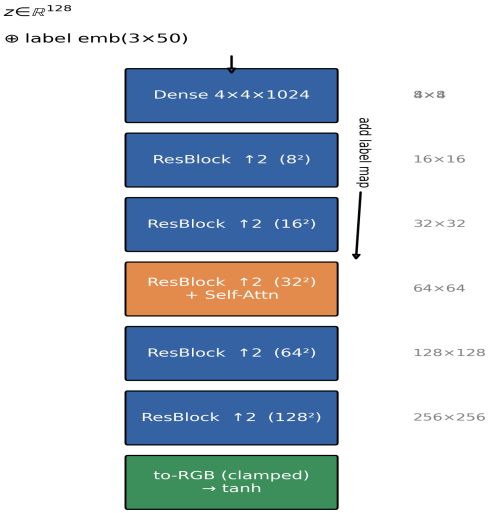
Metric (%)	EXP	ICM	TE
Accuracy	84.6	68.5	69.18
Macro F1	76.0	58.00	72.00

Table: *

Real-only performance (baseline).

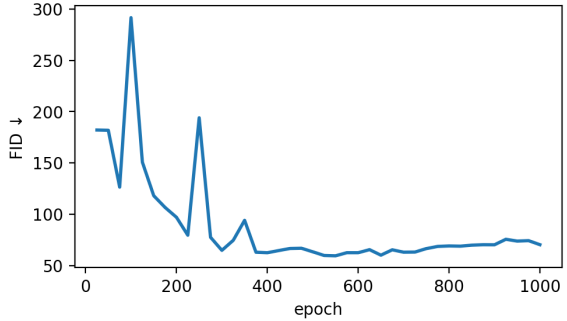
Generator Architecture

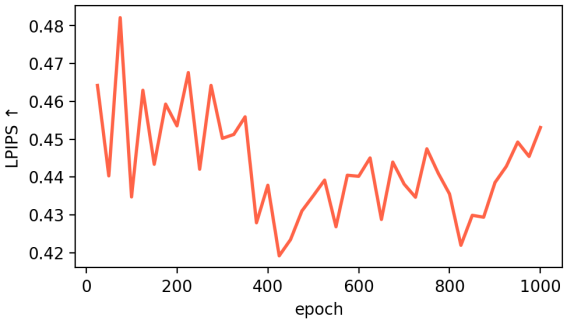
1. Concatenate $z \in \mathbb{R}^{128}$ with three 50-dim label embeddings.
2. 4×4 FC block → residual up-blocks (4→8→. . . →256).
3. Self-attention at 32×32, noise injection, batch-norm.
4. **Clamped to-RGB** (range $[0, 0.5]$).



Training Dynamics

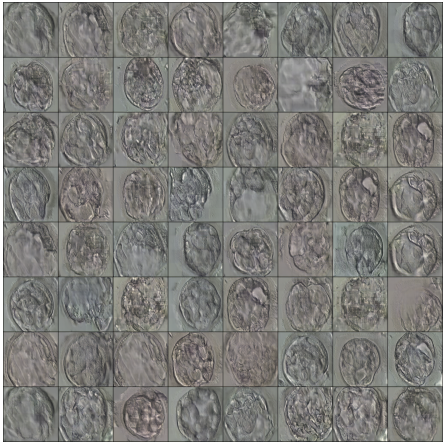
- Hinge loss, Adam: $\eta_G = 3 \cdot 10^{-4}, \eta_D = 1 \cdot 10^{-4}$.
- Spectral-norm D, 1:(1, 2, 3, 4) updates, batch 16.
- Trained 1 000 epochs $\Rightarrow$ best results at $\sim$epoch 600.





*FID↓ and LPIPS↑ convergence.

Conditioned Samples (256^2)



Synthetic Boost for Minority Classes

- Epoch 600 generator – sampled 1 000 images per label.
- 13 k synthetic embryos merged with real set.
- $\rightarrow$ near-uniform class histogram.

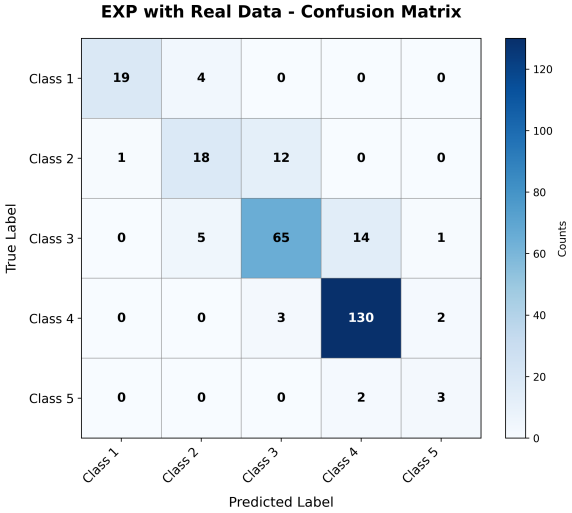
Hybrid vs Real-only

For training with synthetic images, we included 1,000 of each class and used the same exact parameters.
While we did not observe great leaps in accuracy, we did surpass the models that were trained only on real data, indicating that the information within the synthetic images helped with the training.

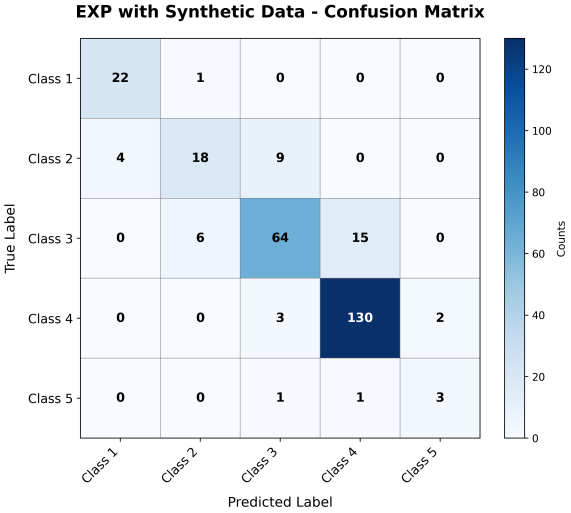
	Accuracy (%)		Macro F1	
Criterion	Real	Hybrid	Real	Hybrid
EXP	84.6	**85.0**	75.0	**77.0**
ICM	68.4	**73.1**	**55.0**	53
TE	**69.1**	66.3.0	**59.0**	53.0

↑ ICM +4.6 pp – synthetic data filled the rare class gap. EXP slight drop due to heavy re-balancing.

Confusion Matrix – EXP

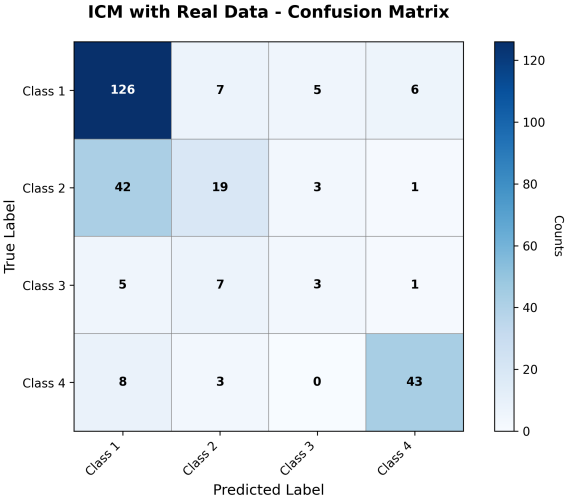


Real-only

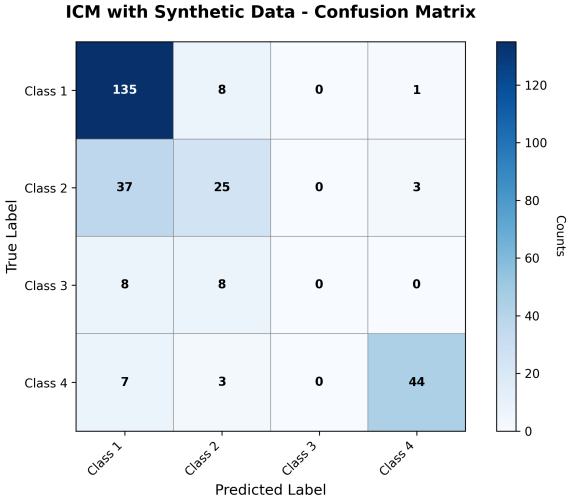


Hybrid

Confusion Matrix – ICM

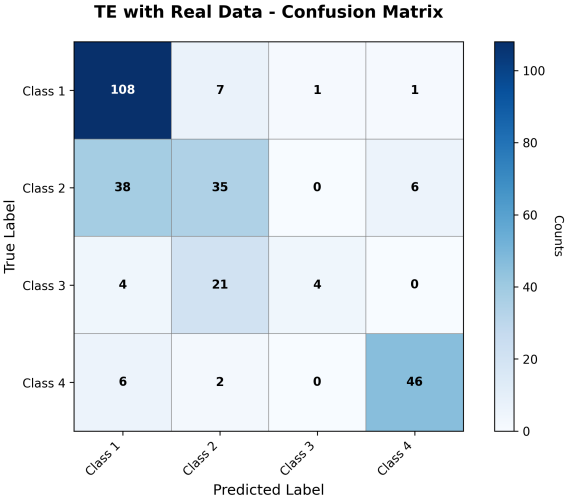


Real-only

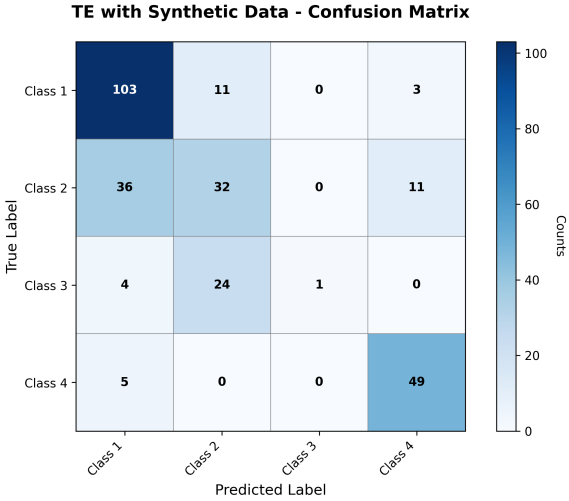


Hybrid

Confusion Matrix – TE



Real-only



Hybrid

Synthetic Images Take-away

- FID plateau $\Rightarrow$ early-stop at optimum, skip late-epoch collapse.
- LPIPS $\approx 0.45 \rightarrow$ synthetic images are *perceptually diverse*.
- EMA attempts produced grayscale artifacts; likely decay β too high.
- Future work: Diffusion models, ADA.

Classification Take-away

1. SWIN training create models close or better than Baseline.
2. Conditional GAN (FID 60) generates label-consistent embryos, close to original data.
3. Synthetic augmentation improves improves classification.

Thank you!
Questions ?