Appendix 1

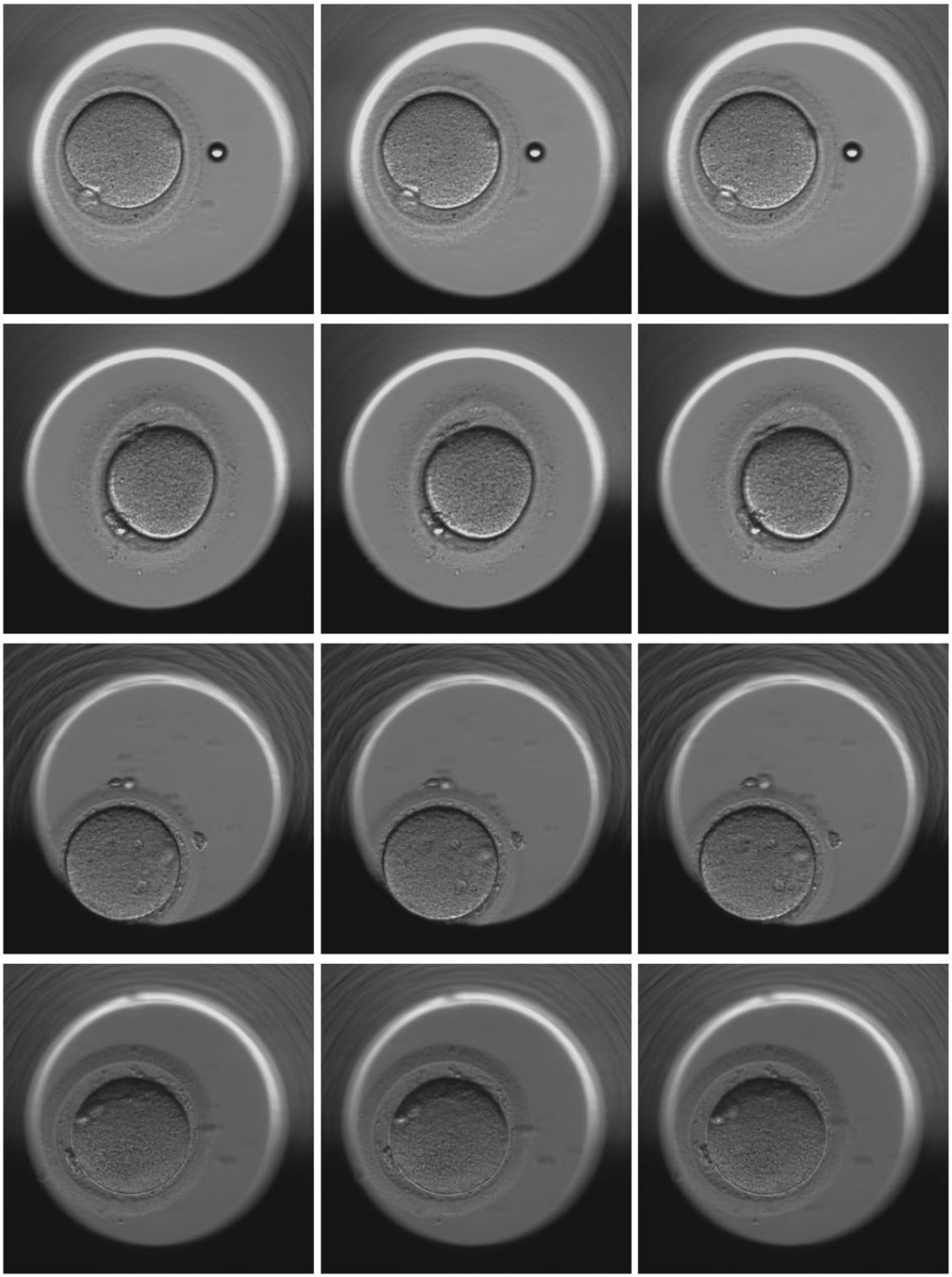

Figure 4 A random sample of the dataset for class tPNa. Each row corresponds to one embryo, and the middle image in each row corresponds to the exact frame reported in the dataset for the morphokinetic event changes. the image on the left is one frame before, and the image on the right is one frame after the labelled change. For this specific morphokinetic event, the pronuclei should be visible in the middle frame and after that.

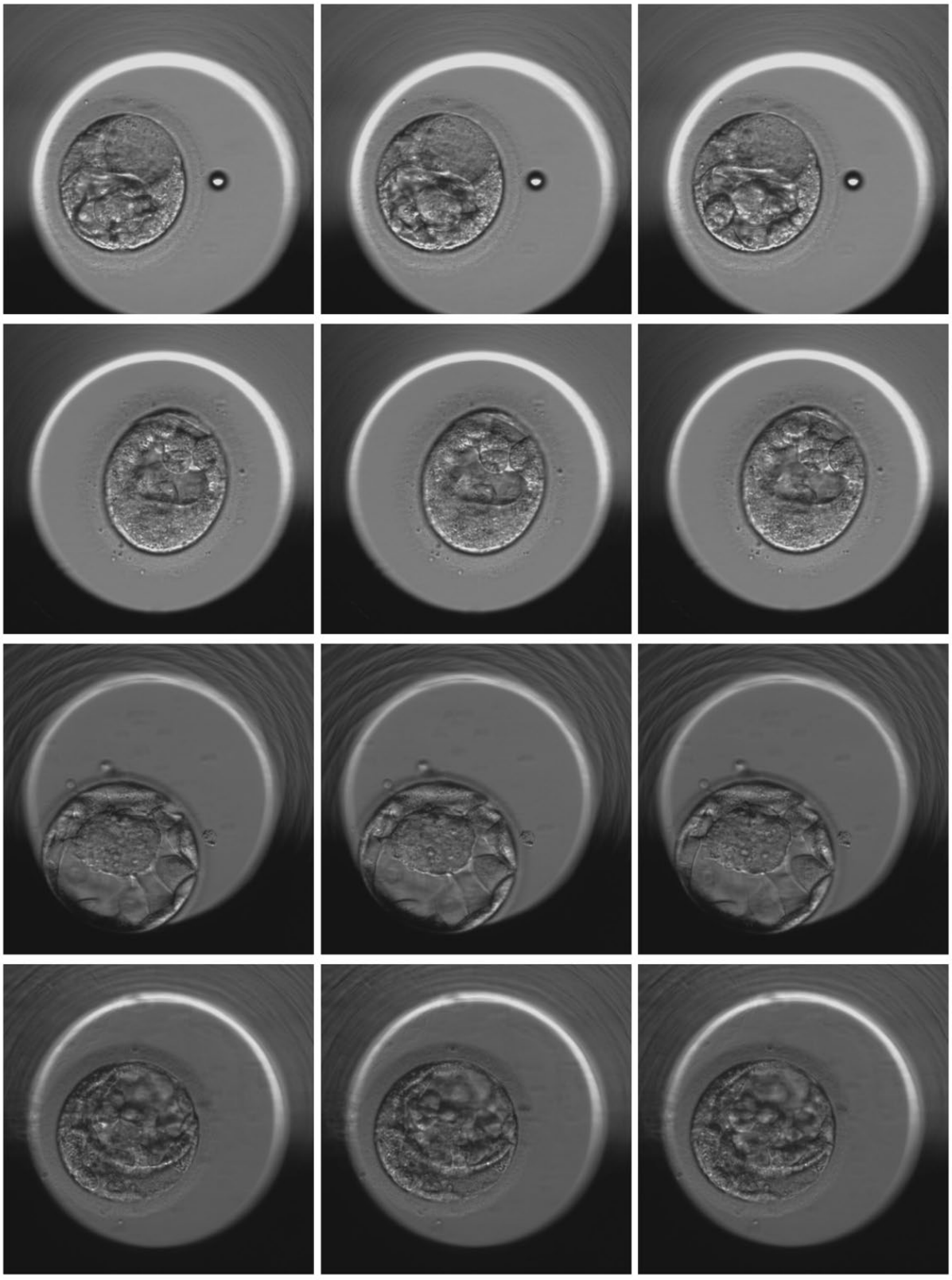

Figure 5 A random sample of the dataset for class tEB. Each row corresponds to one embryo, and the middle image in each row corresponds to the exact frame reported in the dataset for the morphokinetic event changes. the image on the left is one frame before, and the image on the right is one frame after the labelled change. This shows the subjectivity in the classes between tB and tEB. It is unclear what the criteria for the difference between the classes is.

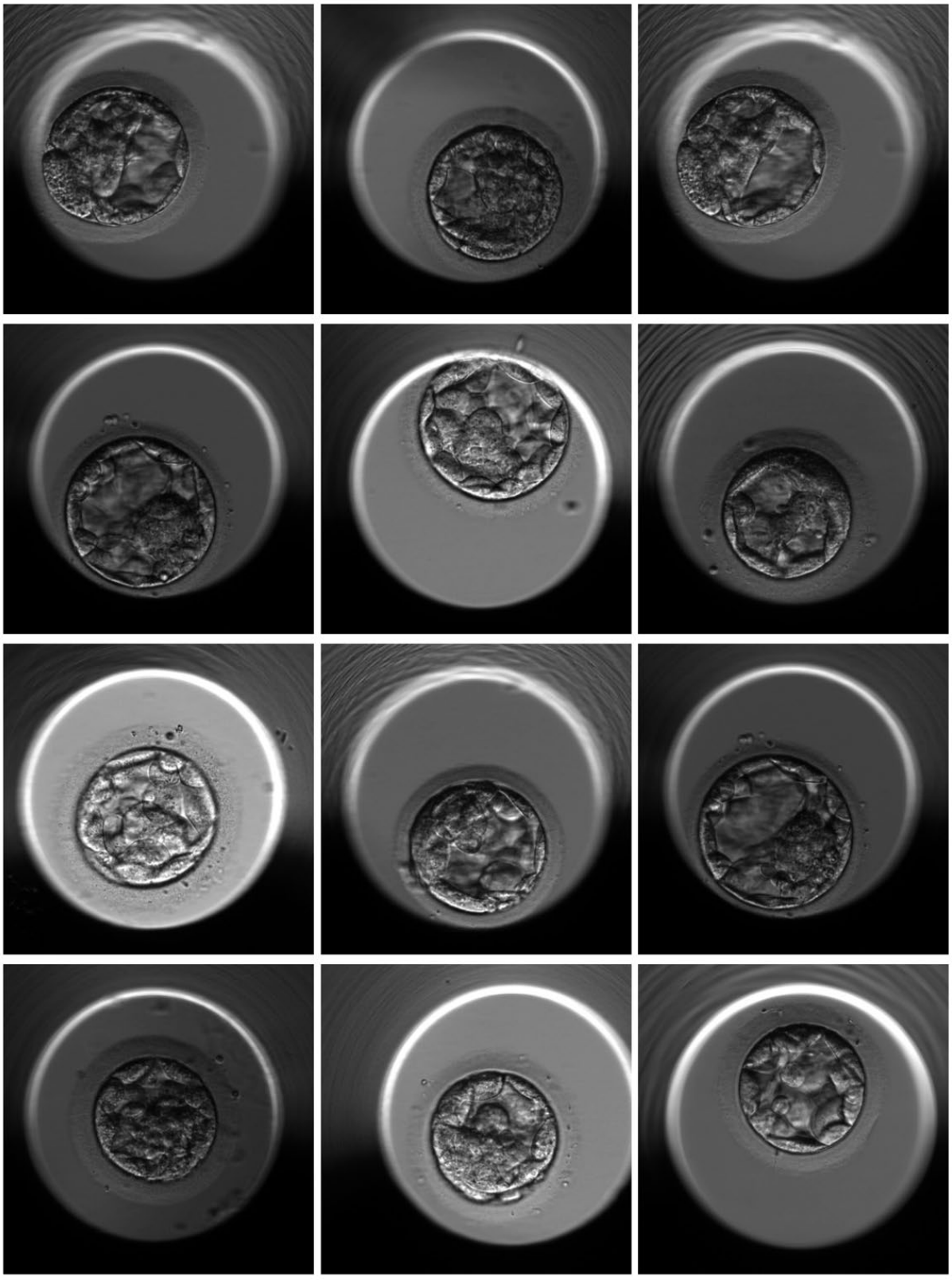

Figure 6 A random sample of Images was predicted as tEB by the network but was labelled as tB in the dataset. The misprediction between tEB and tB is among the worst classes for model 2, with 12% of the tB labelled images in the test set predicted as tEB.

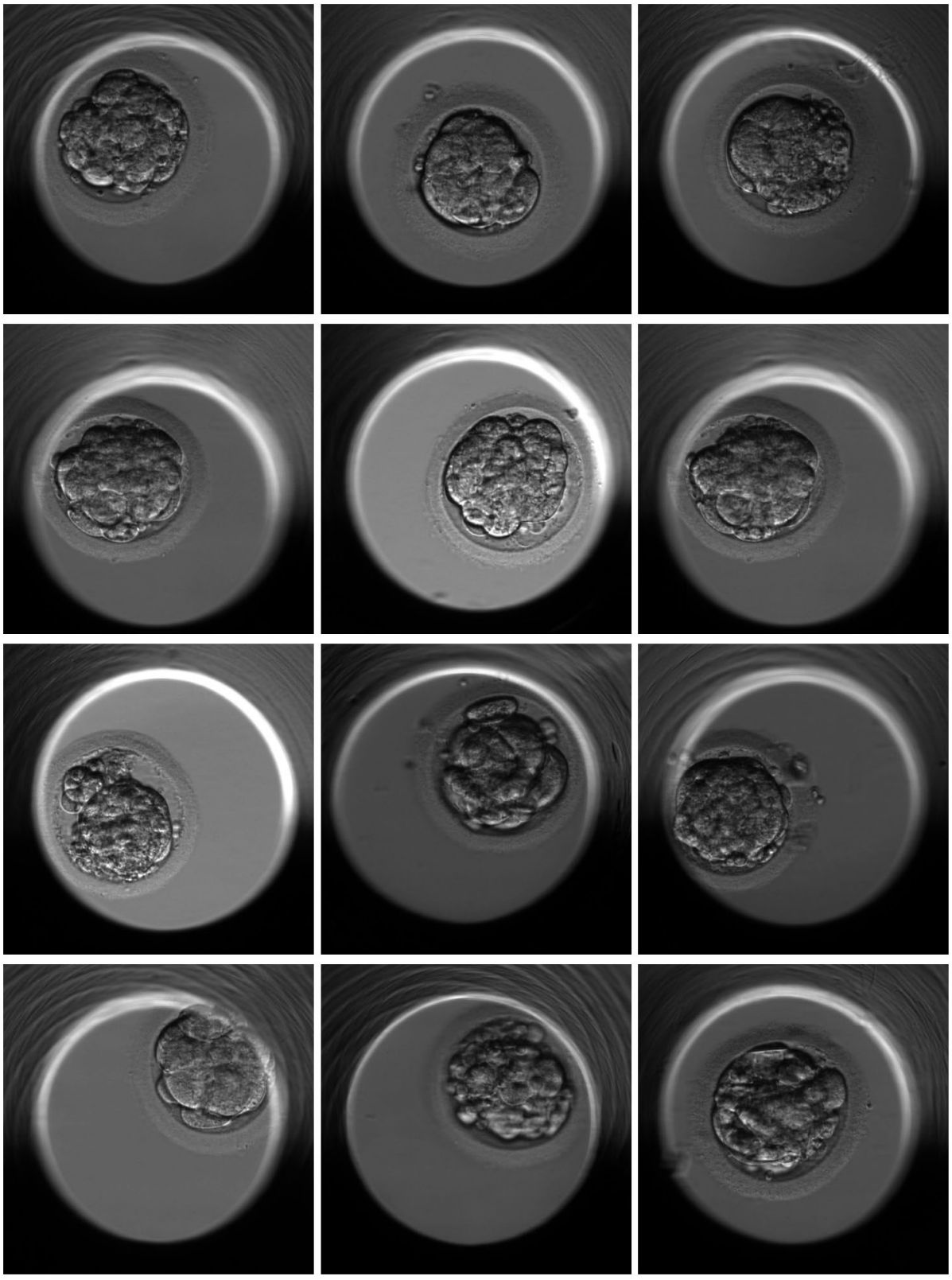

Figure 7 A random sample of Images that was predicted as tM by the network but was labelled as tSB in the dataset. The misprediction between tM and tSB is among the worst classes for model 2, with 12% of the tSB labelled images in the test set predicted as tEB.

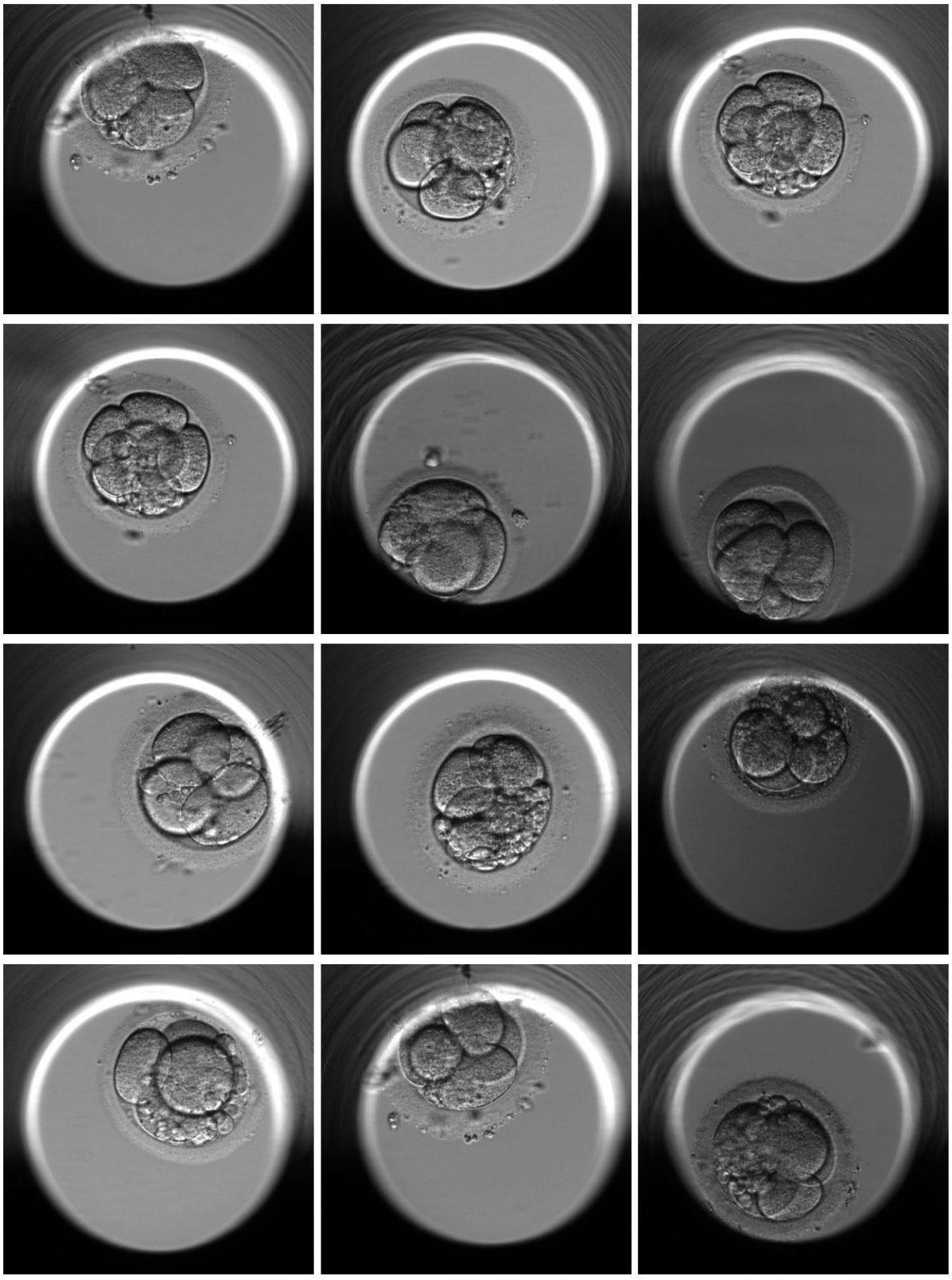

Figure 8 A random sample of Images that was predicted as t4 by the network but was labelled as t5 in the dataset. The misprediction between t4 and t5 is among the worst classes for model 2, with 16% of the t5 labelled images in the test set being predicted as t4.

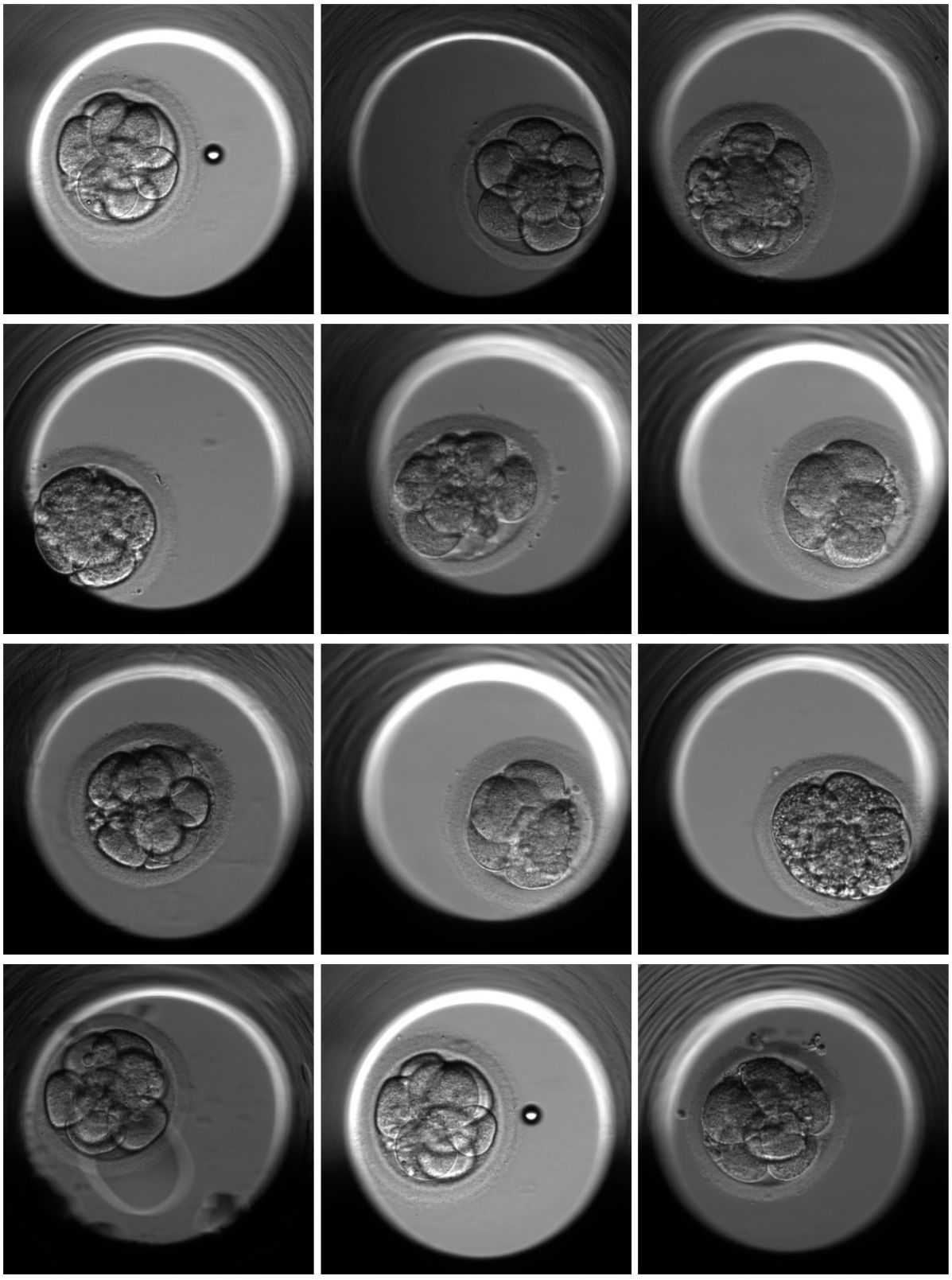

Figure 9 A random sample of Images that was predicted as t8 by the network but was labelled as t7 in the dataset. The misprediction between t8 and t7 is among the worst classes for model 2, with 13% of the t7 labelled images in the test set being predicted as t8.

Appendix 2

The model was trained on a dataset with 66,634 labelled images, the images were captured using Embryoscope device, this is the same type of device that was used in Gomez et al dataset. This dataset was labelled with different set of label definitions in comparison to Gomez et al dataset. Table 5 Shows the distribution of data and classes. This dataset includes labels for empty wells.

Table 5 Distribution of data in the second dataset.

Classes	Description	**Number of samples**
2PN	Visible pronuclei	6516
Syngamy	Syngamy	5950
2 Cell	Cleavage stage 2 cells	7992
3 Cell	Cleavage stage 3 cells	1761
4 Cell	Cleavage stage 4 cells	5958
5 Cell	Cleavage stage 5 cells	2340
6-7 Cell	Cleavage stage 6 or 7 cells	4254
8 Cell	Cleavage stage 8 cells	5296
9+ Cell	Cleavage stage 9 or more cells	4577
Compacted	compacted	4023
Blastocyst12	Start of blasting and blastocyst	6077
Blast35	Expanded blastocyst and hatched blastocyst	3955
Empty	Empty well	7935

A classification network employing ResNet-50 [30] as the backbone was used to extract the features from images, then the classification is done using fully connected layers. This network used the same training methods and loss function as Model 1,2,3. For training this network the dataset was split to train and test subsets, 80 % were used for training and 20 % for testing.

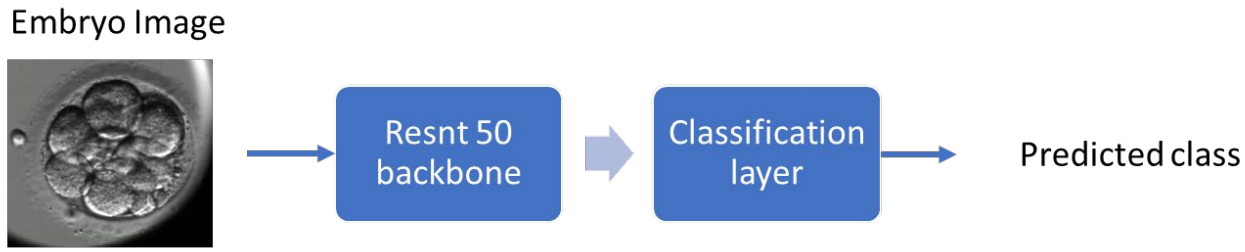

Figure 10 Model 1 Trained on Resnet-50

This model was able to predict the empty wells on the test dataset with 100% accuracy.