Supplementary Material: A Novel Deep Learning Approach with Transformer-GRU Architecture to identify embryo kinetic events

Tables S1 S2 and S3 show the F1-score, Precision and Recall (in %) per event on Embryoscope®, GERI® and MIRI® respectively. We can see the model performs better on Embryoscope which is reasonable considering 70% of the videos in the dataset come from this TLS. Also the test size for GERI® and MIRI® are quite limited (47 and 35 videos respectively) of which a significant portion are non developing embryos (49% and 37%).

Event	**Precision**	**Recall**	**F1-score**	**n samples**
t2	95.8	96.3	96.1	190
t3	57.6	79.1	66.7	134
t4	87.1	72.7	79.3	176
t5	48.4	60.7	53.8	145
t6	69.4	80.7	74.7	135
t7	60.5	75.4	67.2	122
t8	49.7	43.7	46.5	176
tM	63.5	58.4	60.8	125
tSB	70.6	69.5	70.1	128
tB	73.1	76.3	74.7	114
Weighted average	**68.6**	**71.4**	**69.5**	**193**

Table S1: F1-score, Precision and Recall (in %) per event on Embryoscope®. $N = 193$ embryos of which $n = 5$ were non developing embryos.

Event	**Precision**	**Recall**	**F1-score**	**n samples**
t2	88.6	81.6	84.9	38
t3	58.8	60.6	59.7	33
t4	83.3	44.1	57.7	34
t5	44.4	44.4	44.4	27
t6	71	68.7	69.8	32
t7	51.9	53.8	52.8	26
t8	53.8	56	54.9	25
tM	29.6	42.1	34.8	19
tSB	0	0	0	9
tB	71.4	78.9	75	19
Weighted average	**61.9**	**57.6**	**58.9**	**47**

Table S2: F1-score, Precision and Recall (in %) per event on GERI®. $N = 47$ embryos of which $n = 23$ were non developing embryos.

Event	**Precision**	**Recall**	**F1-score**	**n samples**
t2	82.8	80	81.4	30
t3	34.4	55	42.3	20
t4	57.1	33.3	42.1	24
t5	37.9	47.8	42.3	23
t6	55.6	71.4	62.5	21
t7	37.5	52.9	43.9	17
t8	33.3	31.6	32.4	19
tM	38.1	40	39	20
tSB	94.7	90	92.3	20
tB	31.6	30	30.8	20
Weighted average	**52.1**	**54.2**	**52.3**	**35**

Table S3: F1-score, Precision and Recall (in %) per event on MIRI®. $N = 35$ embryos of which $n = 13$ were non developing embryos.

Event	**F1-score w. poor**	**F1-score w/o. poor**
t2	92.8	96.7
t3	62.7	67.2
t4	73.1	77.9
t5	51.3	55.3
t6	72.5	76.2
t7	62.5	65.3
t8	46.3	47.1
tM	54.4	55.6
tSB	70.4	70.6
tB	69.2	69.7
Weighted average	66.3	68.7

Table S4: F1-score (in %) per event after removing non-developing embryos from the test set.

Table S4 displays the F1-score per event of the BEE model on the test set before and after removing the poor embryos. Performances on earlier cleavage stages benefit the most from this removal. Performance on later events such as tM, tSB and tB do not change mainly du to the fact that these poor embryos do not reach these stages.

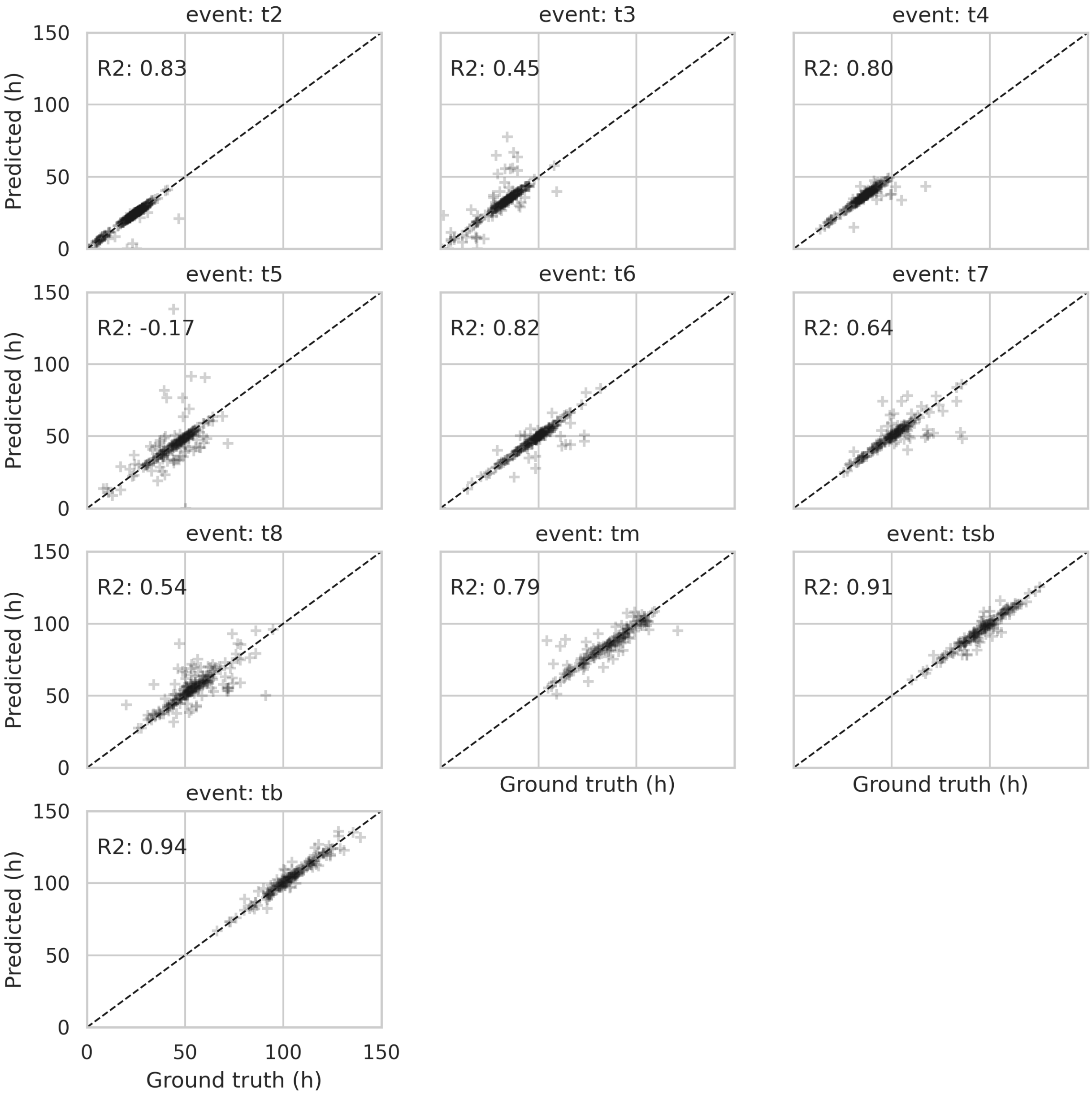

Figure S1: Scatter plot of event start in ground truth vs. predicted event start. R^2 score per event is also presented.

Figure S1 shows a scatter plot of ground truth event start versus predicted event start as well as the R^2 score per event. Some lower R^2 score are impacted by the presence of a few outliers.

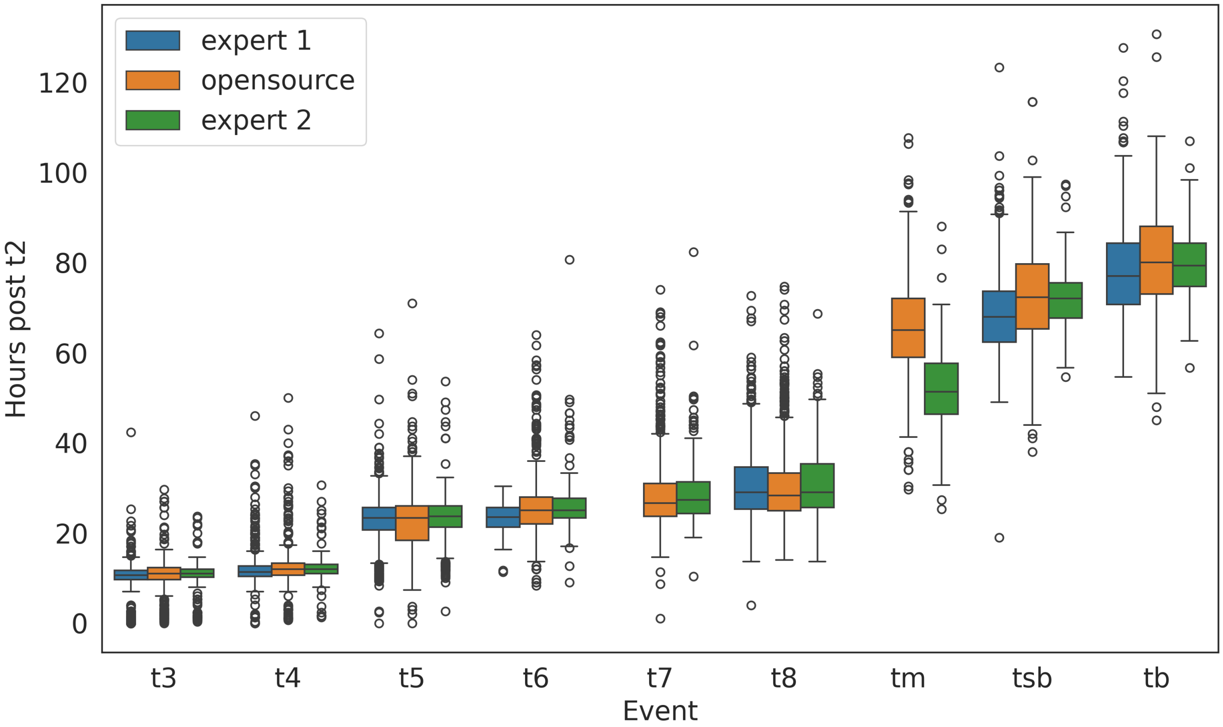

Figure S2: Boxplot for the event start, per annotator, for each event, as the number of hours post t2 start.

Figure S2 shows the boxplot of event start per annotator. To account for possible differences between ICSI and IVF methods, the timings are displayed as hours post t2. We do not have annotations from Expert 1 for t7 and tM events. We can also note the difference on tM between the open-source dataset and Expert 2, which could explain the difficulty for the model to perform well on this event.