# Multimodal Learning for Embryo Viability Prediction in Clinical IVF

Junsik Kim[1], Zhiyi Shi[2], Davin Jeong[1], Johannes Knittel[1]
Helen Y. Yang[1], Yonghyun Song[1], Wanhua Li[1], Yicong Li[1]
Dalit Ben-Yosef[4], Daniel Needleman[1,3], and Hanspeter Pfister[1]

[1] Harvard University, Cambridge MA 02138, USA
[2] Carnegie Mellon University, PA 15213, USA
[3] Flatiron Institute, New York, NY 10010, USA
[4] Tel Aviv Sourasky Medical Center, Tel Aviv, Israel
jskim@seas.harvard.edu

**Abstract.** In clinical In-Vitro Fertilization (IVF), identifying the most viable embryo for transfer is important to increasing the likelihood of a successful pregnancy. Traditionally, this process involves embryologists manually assessing embryos' static morphological features at specific intervals using light microscopy. This manual evaluation is not only time-intensive and costly, due to the need for expert analysis, but also inherently subjective, leading to variability in the selection process. To address these challenges, we develop a multimodal model that leverages both time-lapse video data and Electronic Health Records (EHRs) to predict embryo viability. One of the primary challenges of our research is to effectively combine time-lapse video and EHR data, owing to their inherent differences in modality. We comprehensively analyze our multimodal model with various modality inputs and integration approaches. Our approach will enable fast and automated embryo viability predictions in scale for clinical IVF.

**Keywords:** Multimodal Learning · Time-lapse Video · EHR · Human Embryos · In-Vitro Fertilization.

## 1 Introduction

Infertility affects approximately one in six couples globally [7], propelling many towards assisted reproductive technologies such as In-Vitro Fertilization (IVF). IVF entails stimulating patients to produce multiple oocytes, which are then retrieved, fertilized in vitro, and the resultant embryos cultured. Selected embryos are transferred to the maternal uterus to initiate pregnancy, with surplus viable embryos cryopreserved for future attempts. Although transferring multiple embryos might increase the likelihood of conception, it simultaneously elevates the risk of multiple pregnancies, which are linked to heightened maternal and neonatal morbidity and mortality [25]. Consequently, there is a pressing need to limit embryo transfer to a single, optimally selected embryo to maximize the chances of a healthy singleton birth [20] which remains challenging [26].

The prevailing practice in embryo selection primarily relies on morphological analysis through microscopic imaging. Embryos undergo a series of developments post-fertilization, transitioning through stages from pronuclei alignment to blastocyst formation, with clinicians traditionally scoring embryos based on discrete, manually observed morphokinetic features such as cell number, cell shape, cell symmetry, the presence of cell fragments, and blastocyst appearance [8]. Nowadays, many clinics adopt time-lapse microscopy incubators to capture movies of embryos continuously without disturbing their culture conditions [3]. Despite this advancement, the analysis of these videos remains manual, which is labor-intensive and subjective.

Numerous studies have focused on predicting and analyzing the morphological features of embryos using images or videos, covering aspects like blastocyst size [15], blastocyst grade [10,16,17], cell boundaries [27,13], cell counting [14,18], and developmental stage prediction [23]. Subsequently, a comprehensive pipeline employing deep learning models was developed to predict five key morphological features of embryos [19], yielding outputs in the forms of classification, regression, and segmentation. These key morphological features are shown to be correlated to the live birth result of IVF treatments when converted to interpretable features by heuristic post-processing [32], such as the timing of stage transitions, cell symmetry index, and zona thickness. However, solely relying on the converted features may overlook more intricate and nuanced details of embryo development captured in videos. Additionally, these approaches mainly focus on visual features from time-lapse imaging and do not integrate data from Electronic Health Records (EHRs), which contain important variables such as patients' health information and treatment details.

In this work, we introduce a multimodal model for predicting embryo viability, leveraging both time-lapse videos and Electronic Health Records (EHRs). Although there has been an attempt to utilize image and EHR modalities [22], their focus is not on multimodal integration, and they do not use video data. A major challenge in multimodal learning is the effective integration of diverse modal types to ensure balanced training without modality bias [31]. We explore different multimodal integration methods. Inspired by the previous works, our multimodal model not only incorporates time-lapse videos and EHR data but also includes morphological [19] and interpretable features [32] as additional inputs. Through comprehensive experiments with diverse combinations of modalities, we analyze different multimodal integrations and demonstrate the effectiveness of our multimodal model for embryo viability prediction in clinical IVF.

## 2 Dataset

We collected data from 3,695 IVF treatment cycles with 24,027 embryos imaged every 20 minutes up to the first five days of development where each image size is $500 \times 500$. This corresponds to approximately 6 million images of embryos. Additionally, electronic health record (EHR) data, including patient information, treatment information, and live birth records as a treatment outcome, are

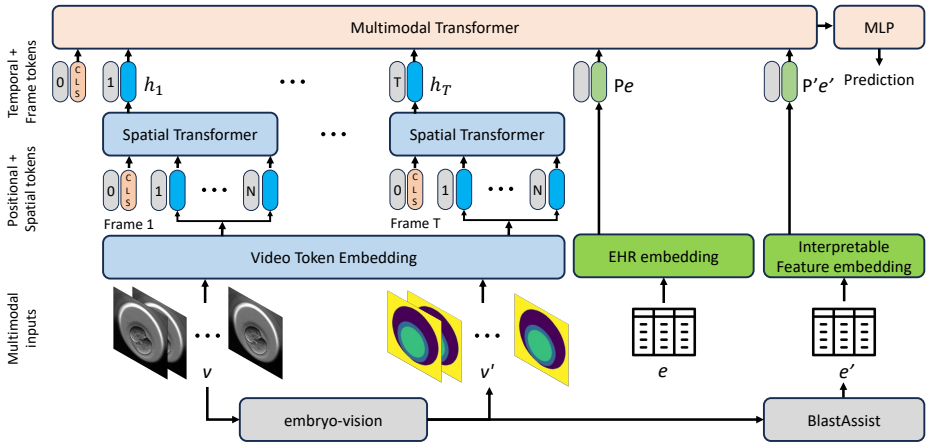


**Fig. 1.** Overview of our multimodal model. Video data is first tokenized into patches per frame. Then, the spatial transformer encodes per frame embeddings. The Multimodal transformer inputs both frame embeddings and an EHR embedding to output a multimodal feature. Lastly, the MLP head predicts embryo viability based on the multimodal feature. If additional inputs in the form of video or tabular are available, such as outputs from Embryo-vision [19] or BlastAssist [32], they are processed in a similar manner as the video input and the EHR input respectively.

collected. Among the collected data samples, we curate the multimodal dataset with embryos that have both video and EHR modalities with treatment outcomes. Our multimodal dataset consists of a total of 1700 treatment cycles with 3318 embryos. Out of 1700 treatments, 260 treatments are successful with equal or more than one live birth. It's important to note that each treatment cycle fertilizes multiple embryos, and only healthy embryos are selected for transfer. Some cycles freeze all embryos for future use rather than immediate transfer. Therefore, the number of embryos that have the treatment outcome is limited compared to the scale of the raw data collected.

## 3 Method

In this work, we explore two different directions to integrate multimodal data for embryo viability prediction. One is a transformer-based multimodal model where EHRs and videos are processed end-to-end, as shown in Fig. 1. Another approach is to take a two-stage approach where the video data is first processed to extract morphological features in tabular format using off-the-shelf methods [19,32], and then input to the tabular models with EHRs as shown in Fig. 2. Although the two-stage approach can be modeled by a single tabular modality model, it is multimodal by nature as video data is converted and included in a tabular format.

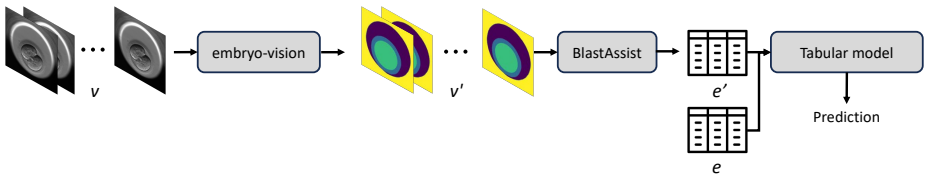


**Fig. 2.** Overview of the two-stage approach. First, morphological features $\mathbf{v}'$ are extracted from videos using [19]. Then, the extracted features $\mathbf{v}'$ are converted to interpretable features $\mathbf{e}'$ in tabular format using [32]. Lastly, the tabular model inputs EHRs $\mathbf{e}$ and interpretable features $\mathbf{e}'$ to predict embryo viability.

Although there are several multimodal transformer models [28,1,21,24] available, it is not straightforward to apply them to embryo viability prediction as they assume samples in each modality has one-to-one correspondence. In our case, videos are embryo-specific, but EHRs are treatment-specific, which is shared across embryos within the same treatment cycle. Therefore it is difficult to directly apply cross-modal correspondence or contrastive learning as in other multimodal learning approaches. To this end, we propose a multimodal transformer that is based on a video transformer architecture with modifications to allow multimodal inputs.

**Input modalities** Let $\mathcal{T}_n = \{v_n, e_n\}$ be a multimodal sample in $n$-th treatment cycle in our multimodal dataset, where $v_n^m \in \mathbb{R}^{T\times H\times W\times C}$ denotes a time-lapse video of $m$-th embryo fertilized in $n$-th treatment cycle and $e_n \in \mathbb{R}^C$ denotes an EHR containing information of the patient and treatment applied. Note that time-lapse videos are embryo-specific, but EHR data corresponds to the treatment cycle; thus, they are not embryo-specific. Our goal is to predict embryo viability formulated as $y = \frac{\text{n\_births}}{\text{n\_transferred}}$, where viability is defined as the number of births over the number of embryos transferred. The number of embryos transferred at each treatment cycle varies depending on various factors, such as the number of embryos fertilized, embryo quality examined by embryologists, or the patient's medical history.

Other than video data, we can additionally utilize morphological embryo features extracted from videos by off-the-shelf methods, *e.g.*, Embryo-vision [19], and BlastAssist [32]. Embryo-vision outputs a set of features $v_{n,t}^{'m}$ from a video frame $v_{n,t}^m$, which are zona semantic segmentation $\mathbf{s}_z$, blastomere instance segmentation $\mathbf{s}_b$, pronuclei instance segmentation $\mathbf{s}_p$, fragmentation regression $\mathbf{r}$, and stage classification $\mathbf{c}$. BlastAssist further converts the morphological features into a set of interpretable features $\mathbf{e}'$ such as zona well thickness, stage transition timing, and cell symmetry index. For more details, refer to the supplementary and [19,32].

**Video transformer** Videos are significantly larger than the size of other modalities. Directly applying spatio-temporal attention to a video would result in a large number of tokens, which require an immense amount of memory and com-

putation. Inspired by ViViT [4], we design a transformer in a factorized encoder structure where spatial attention is applied first, followed by temporal attention.

For spatial attention, we first tokenize each frame $v_{n,t}^m \in \mathbb{R}^{H\times W\times C}$ to a set of tokens by extracting non-overlapping image patches $x_i \in \mathbb{R}^{h\times w\times C}$ and then apply linear projection $\mathbf{E}$. Then, a set of embedded frame tokens and a learnable class token are added to a learnable positional embedding $\mathbf{p}$ and passed through a transformer consisting of a sequence of $L$ transformer layers to output a frame-level representation.

$$\mathbf{z} = [z_{\text{cls}}, \mathbf{E}x_1, \ldots, \mathbf{E}x_N] + \mathbf{p} \tag{1}$$

Each transformer layer $\ell$ consists of Multi-Headed Self-Attention [30], layer normalisation(LN) [5], and MLP blocks as follows:

$$\mathbf{y}^\ell = \text{MSA}(\text{LN}(\mathbf{z}^\ell)) + \mathbf{z}^\ell \tag{2}$$

$$\mathbf{z}^{\ell+1} = \text{MLP}(\text{LN}(\mathbf{y}^\ell)) + \mathbf{y}^\ell \tag{3}$$

The output token $z_{cls}^L$ embeds frame-level representation. Temporal attention is performed similarly to spatial embedding by applying $L'$ transformer layers on a set of frame tokens $\mathbf{h}$,

$$\mathbf{h} = [h_{\text{cls}}, z_{cls,1}^L, \ldots, z_{cls,T}^L] + \mathbf{t} \tag{4}$$

where $h_{cls}$ is a learnable class token in temporal attention, and $\mathbf{t}$ is a learnable temporal embedding.

**Multimodal Transformer** We modify a video transformer to allow multimodal inputs. We embed EHR data $\mathbf{e}$ by linear projection and then append to the frame tokens. If we have additional features in a tabular format, *e.g.*, interpretable features $\mathbf{e}'$, it is processed in the same way as EHR data. With EHR data tokens, the temporal attention input in Eq. (4) becomes multimodal attention input as follows,

$$\mathbf{h} = [h_{\text{cls}}, h_1, \ldots, h_T, \mathbf{P}e, \mathbf{P}'e'] + \mathbf{t} \tag{5}$$

where $h_t$ is a frame token at frame $t$, $\mathbf{P}$ and $\mathbf{P}'$ are linear projections for $\mathbf{e}$ and $\mathbf{e}'$ respectively. When only video is input to the model, a frame token $h_t$ becomes $z_{cls,t}^L$ as in Eq. (4). Additionally, we can incorporate more per-frame modality inputs from Embryo-vision to enrich the representation of a frame token $h_t$. The Embryo-vision outputs a set of morphological features $v' = \{s_z, s_b, s_p, r, c\}$ where the first three features are segmentation masks and the latter two are vectors. The mask format features are passed to the spatial attention and processed similarly to the video input. For simplicity, let's denote spatial transformer operation as $f_s : \mathbb{R}^{H\times W\times C} \to \mathbb{R}^d$. When a video is input, $f_s(v_t)$ equals $z_{cls,t}^L$ as in Eq. (4). When multiple video modalities are available, the frame token $h_t$ is formulated as a concatenation of tokens from different modalities as follows,

$$h_t = [f_s(v_t), f_s(s_{z,t}), f_s(s_{b,t}), f_s(s_{p,t}), \mathbf{E}'[r_t, c_t]] \tag{6}$$

where $\mathbf{E}'$ is a linear projection applied to the concatenation of $r_t$ and $c_t$.

**Table 1.** Number of successful and failed treatments and embryos in each split in the form of "number of embryos" / "number of treatments."

| Split | Total | Success | Fail |
|---|---|---|---|
| Train | 2617/1360 | 362/208 | 2255/1152 |
| Val | 327/170 | 54/26 | 273/144 |
| Test | 342/170 | 54/26 | 288/144 |

## 4 Experiments

**Implementation details** For spatial attention, we use the pre-trained DeiT-Ti [29] as a spatial transformer without fine-tuning. We attempted to fine-tune a spatial transformer, but it resulted in worse performance due to the limited number of labeled samples. For temporal or multimodal attention, we use 4 transformer layers. Input frames are resized from $500 \times 500$ to $224 \times 224$. Videos are clipped to have a maximum of 360 frames since this corresponds to the first 5 days of observation, where each frame is captured at 20-minute intervals. To enable memory-efficient training, we subsample every 4 frames, resulting in 90 frames per video. Flip and rotation are applied to videos and masks for augmentation. The batch size is set to 4, the learning rate is set to 1e-4, and the model is trained until the validation loss converges. MLP head consists of two fully connected layers with ReLU activation in between. Huber loss [12] is used to train the multimodal transformer. The experiments are performed using one A100 GPU.

**Experiment setup** We randomly split train, validation, and test splits to an 8:1:1 ratio while preserving the success rate within each split. For evaluation, we use two performance metrics: the area under the receiver operating characteristic curve (ROCAUC) and F1-Score. We evaluate two different scenarios: embryo viability prediction and treatment success prediction. Each treatment has equal to or more than one embryo transferred. In the embryo viability prediction scenario, we set the ground truth label to '1' for all embryos transferred (instead of $\frac{\text{n\_births}}{\text{n\_transferred}}$) if the treatment is successful, then compute AUCROC and F1-Score. In treatment success prediction, we sum the viability predictions of embryos transferred together and then calculate AUCROC and F1-Score. For F1-Score measurement, we use 0.15[5] as a threshold for embryo viability prediction and 0.5[6] for treatment success prediction. F1-Score quantifies the precision of predictions at a fixed threshold, whereas AUCROC measures capability in assessing the relative quality of the samples.

**Two-stage approach** We compare our multimodal transformer with two-stage approaches using two transformer based methods: TabTransformer [11] and Tab-

[5] The treatment cycle with the highest number of embryos transferred is 5. Therefore, embryo viability values in successful treatments range from 0.2 to 1.0.

[6] Treatment success is defined as the n_births value equal to or higher than 1.

**Table 2.** Performance comparison on embryo viability prediction with different modalities using a multimodal transformer. **v** is a video modality, **v'** is an output from Embryo-vision, **e** is EHR data, and **e'** is an output from BlastAssist. The best performance is marked in bold.

| Modality | Embryo | | Treatment | |
|---|---|---|---|---|
| | AUCROC | F-1 | AUCROC | F-1 |
| **v** | 0.578 | 0.284 | 0.579 | 0.315 |
| **v+e** | 0.580 | 0.297 | 0.581 | 0.286 |
| **v+v'** | 0.676 | 0.316 | 0.675 | **0.336** |
| **v+v'+e+e'** | 0.647 | 0.296 | 0.643 | 0.310 |
| **v'** | 0.666 | 0.317 | **0.697** | 0.313 |
| **v'+e+e'** | **0.688** | **0.338** | 0.683 | 0.312 |

Net [2]. We follow the official implementation of [6] [7] to train tabular models with the best hyperparameters after performing hyperparameter search using cross-validation. For more details, refer to the supplementary.

**Table 3.** Performance comparison on embryo viability prediction with different modalities using a two-stage approach. **e** is EHR data, and **e'** is an output from BlastAssist. Confidence intervals are reported with 10 runs.

| Modality | Method | Embryo | | Treatment | |
|---|---|---|---|---|---|
| | | AUCROC | F-1 | AUCROC | F-1 |
| **e** | TabTransformer [11] | 0.586 ± 0.045 | 0.110 ± 0.068 | 0.604 ± 0.054 | 0.167 ± 0.111 |
| | TabNet [2] | 0.591 ± 0.016 | 0.240 ± 0.020 | 0.631 ± 0.017 | 0.113 ± 0.033 |
| **e+e'** | TabTransformer [11] | 0.634 ± 0.025 | 0.298 ± 0.045 | 0.681 ± 0.023 | 0.100 ± 0.031 |
| | TabNet [2] | 0.629 ± 0.025 | 0.244 ± 0.042 | 0.672 ± 0.026 | 0.188 ± 0.058 |
| **e'** | TabTransformer [11] | 0.593 ± 0.021 | 0.235 ± 0.040 | 0.624 ± 0.022 | 0.134 ± 0.030 |
| | TabNet [2] | 0.623 ± 0.012 | 0.232 ± 0.042 | 0.630 ± 0.023 | 0.146 ± 0.045 |

**Experiments with multimodal transformer** We evaluate our multimodal transformer on embryo viability prediction task using different combinations of modalities in Tab. 2. The first 4 rows in the table show the results with video modality. The model trained with only video modality performs worse than the other modality combinations. When both video and EHR modalities are used, AUCROC marginally improves. On the other hand, the model performance improves significantly when semantic features are added. This shows that directly predicting embryo viability is challenging and semantic information is important for the prediction. However, adding tabular format modalities to video modalities did not improve the prediction. We conjecture this is due to the increased complexity of multimodal data to learn given limited training samples. The per-

[7] https://github.com/kathrinse/TabSurvey

formance drop with interpretable features is noticeable with video modality, but the performance drop is not observed in other combinations of modalities.

We evaluate the multimodal model without a video input $v$ in the last 2 rows in Tab. 2. The results without a video modality perform better than those with a video modality. This may be due to the limited number of training videos to learn good representation. We deploy a pre-trained vision transformer DeiT-Ti [29] to overcome the limited training set size, but multimodal transformer layers are trained from scratch; therefore, the multimodal attention is performed in a sub-optimal way. On the contrary, a model trained with Embryo-vision outputs $v'$ performs significantly better than those with $v$. Unlike raw video, Embryo-vision outputs are in the form of segmentation masks, which are semantically meaningful and have a simple visual structure. Therefore, it is easier for the model to understand and optimize the weights to extract relevant features for the task.

**Experiments with two-stage approach** We compare the two-stage approach with different types of tabular models. Unlike the end-to-end multimodal learning method, we observed higher performance variation in two-stage methods. We conjecture this is due to the early convergence of two-stage models, which results in different solutions. Here, we report confidence intervals from 10 trials of the two-stage approaches. Among different modalities, using both EHR and interpretable features performs best for the two-stage approaches. Although visual data is not directly input to the model, interpretable features encode visual information; therefore, the tabular models show competitive performance when using both EHRs and interpretable features.

One noticeable difference to the multimodal transformer is the low F-1 score on treatment success prediction. Although tabular models are trained with regression objectives, they fail to calibrate the prediction confidence, resulting in a low F-1 score. In practice, finding the best threshold is a difficult problem. Therefore, without an appropriate threshold estimation method, a model with good confidence calibration is favored. If an optimal threshold can be found, a higher F-1 score will be achieved for both multimodal transformers and two-stage tabular models.

## 5 Discussion and Conclusion

One challenge in multimodal learning with supervision is the size of the training data. Although the collected data is not on a small scale, the embryos with a treatment outcome are very limited. This hinders the supervised training of large-scale models. We conjecture the negative effect of video modality in Tab. 2 is also due to the limited training size. One solution is to pre-train modality-specific encoders separately with pretext tasks using self-supervised learning [9] and then fine-tune the encoders with multimodal transformers by supervised learning for the downstream task. With better encoder representations by self-supervised learning, a multimodal transformer will effectively integrate modality features without performance degradation.

In this work, we explore two approaches to incorporate time-lapse videos and EHRs to build a multimodal model for embryo viability prediction. First, we build a multimodal transformer to allow different modalities to be integrated together. The multimodal inputs include not only videos and EHRs but also other morphological features extracted from off-the-shelf methods. We also explore an alternative method; a two-stage approach where the first stage is to extract and convert visual morphological features to tabular format and then combine it with EHRs for tabular models. The experiments with various modalities demonstrate the effectiveness of our multimodal model over two-stage approaches. We further analyze which modality is important in predicting embryo viability. In future research, we will explore pre-training and fine-tuning methodologies to address the challenge posed by the limited size of supervised training sets in multimodal learning.

**Acknowledgments.** We thank all affiliates of the Harvard Visual Computing Group for their valuable feedback. This work was supported by NIH grant R01HD104969 and Harvard Data Science Initiative Postdoctoral Fellowship.

**Disclosure of Interests.** The authors have no competing interests to declare that are relevant to the content of this article.

## References

1. Akbari, H., Yuan, L., Qian, R., Chuang, W.H., Chang, S.F., Cui, Y., Gong, B.: Vatt: Transformers for multimodal self-supervised learning from raw video, audio and text. Advances in Neural Information Processing Systems (2021)
2. Arik, S.Ö., Pfister, T.: Tabnet: Attentive interpretable tabular learning. In: Proceedings of the AAAI conference on artificial intelligence (2021)
3. Armstrong, S., Bhide, P., Jordan, V., Pacey, A., Marjoribanks, J., Farquhar, C.: Time-lapse systems for embryo incubation and assessment in assisted reproduction. Cochrane Database of Systematic Reviews (5) (2019)
4. Arnab, A., Dehghani, M., Heigold, G., Sun, C., Lučić, M., Schmid, C.: Vivit: A video vision transformer. In: IEEE International Conference on Computer Vision (2021)
5. Ba, J.L., Kiros, J.R., Hinton, G.E.: Layer normalization. arXiv preprint arXiv:1607.06450 (2016)
6. Borisov, V., Leemann, T., Seßler, K., Haug, J., Pawelczyk, M., Kasneci, G.: Deep neural networks and tabular data: A survey. IEEE Transactions on Neural Networks and Learning Systems pp. 1–21 (2022). https://doi.org/10.1109/TNNLS.2022.3229161
7. Cui, W.: Mother or nothing: the agony of infertility. World Health Organization. Bulletin of the World Health Organization **88**(12), 881 (2010)
8. Elder, K., Dale, B.: In vitro fertilization, cambridge university press (2000)
9. Ericsson, L., Gouk, H., Loy, C.C., Hospedales, T.M.: Self-supervised representation learning: Introduction, advances, and challenges. IEEE Signal Processing Magazine **39**(3), 42–62 (2022)
10. Filho, E.S., Noble, J.A., Poli, M., Griffiths, T., Emerson, G., Wells, D.: A method for semi-automatic grading of human blastocyst microscope images. Human Reproduction **27**(9), 2641–2648 (2012)

11. Huang, X., Khetan, A., Cvitkovic, M., Karnin, Z.: Tabtransformer: Tabular data modeling using contextual embeddings. arXiv preprint arXiv:2012.06678 (2020)
12. Huber, P.J.: Robust estimation of a location parameter. In: Breakthroughs in statistics: Methodology and distribution, pp. 492–518. Springer (1992)
13. Jang, W.D., Wei, D., Zhang, X., Leahy, B., Yang, H., Tompkin, J., Ben-Yosef, D., Needleman, D., Pfister, H.: Learning vector quantized shape code for amodal blastomere instance segmentation. In: International Symposium on Biomedical Imaging (ISBI). IEEE (2023)
14. Khan, A., Gould, S., Salzmann, M.: Deep convolutional neural networks for human embryonic cell counting. In: Computer Vision–ECCV 2016 Workshops: Amsterdam, The Netherlands, October 8-10 and 15-16, 2016, Proceedings, Part I 14. Springer (2016)
15. Kheradmand, S., Singh, A., Saeedi, P., Au, J., Havelock, J.: Inner cell mass segmentation in human hmc embryo images using fully convolutional network. In: IEEE International Conference on Image Processing. IEEE (2017)
16. Khosravi, P., Kazemi, E., Zhan, Q., Malmsten, J.E., Toschi, M., Zisimopoulos, P., Sigaras, A., Lavery, S., Cooper, L.A., Hickman, C., et al.: Deep learning enables robust assessment and selection of human blastocysts after in vitro fertilization. NPJ digital medicine **2**(1), 21 (2019)
17. Kragh, M.F., Rimestad, J., Berntsen, J., Karstoft, H.: Automatic grading of human blastocysts from time-lapse imaging. Computers in biology and medicine **115**, 103494 (2019)
18. Lau, T., Ng, N., Gingold, J., Desai, N., McAuley, J., Lipton, Z.C.: Embryo staging with weakly-supervised region selection and dynamically-decoded predictions. In: Machine Learning for Healthcare Conference. PMLR (2019)
19. Leahy, B.D., Jang, W.D., Yang, H.Y., Struyven, R., Wei, D., Sun, Z., Lee, K.R., Royston, C., Cam, L., Kalma, Y., et al.: Automated measurements of key morphological features of human embryos for ivf. In: International Conference on Medical image computing and computer-assisted intervention. Springer (2020)
20. Lee, A.M., Connell, M.T., Csokmay, J.M., Styer, A.K.: Elective single embryo transfer-the power of one. Contraception and Reproductive Medicine **1**, 1–7 (2016)
21. Lin, Y.B., Tseng, H.Y., Lee, H.Y., Lin, Y.Y., Yang, M.H.: Exploring cross-video and cross-modality signals for weakly-supervised audio-visual video parsing. Advances in Neural Information Processing Systems (2021)
22. Liu, H., Zhang, Z., Gu, Y., Dai, C., Shan, G., Song, H., Li, D., Chen, W., Lin, G., Sun, Y.: Development and evaluation of a live birth prediction model for evaluating human blastocysts from a retrospective study. Elife **12**, e83662 (2023)
23. Lukyanenko, S., Jang, W.D., Wei, D., Struyven, R., Kim, Y., Leahy, B., Yang, H., Rush, A., Ben-Yosef, D., Needleman, D., et al.: Developmental stage classification of embryos using two-stream neural network with linear-chain conditional random field. In: International Conference on Medical image computing and computer-assisted intervention. Springer (2021)
24. Nagrani, A., Yang, S., Arnab, A., Jansen, A., Schmid, C., Sun, C.: Attention bottlenecks for multimodal fusion. Advances in Neural Information Processing Systems (2021)
25. Norwitz, E.R., Edusa, V., Park, J.S.: Maternal physiology and complications of multiple pregnancy. In: Seminars in perinatology. vol. 29, pp. 338–348. Elsevier (2005)
26. Racowsky, C., Stern, J.E., Gibbons, W.E., Behr, B., Pomeroy, K.O., Biggers, J.D.: National collection of embryo morphology data into society for assisted reproductive technology clinic outcomes reporting system: associations among day 3 cell

number, fragmentation and blastomere asymmetry, and live birth rate. Fertility and sterility **95**(6), 1985–1989 (2011)
27. Rad, R.M., Saeedi, P., Au, J., Havelock, J.: A hybrid approach for multiple blastomeres identification in early human embryo images. Computers in biology and medicine **101**, 100–111 (2018)
28. Sun, C., Myers, A., Vondrick, C., Murphy, K., Schmid, C.: Videobert: A joint model for video and language representation learning. In: IEEE International Conference on Computer Vision (2019)
29. Touvron, H., Cord, M., Douze, M., Massa, F., Sablayrolles, A., Jégou, H.: Training data-efficient image transformers & distillation through attention. In: International Conference on Machine Learning. PMLR (2021)
30. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, Ł., Polosukhin, I.: Attention is all you need. Advances in Neural Information Processing Systems (2017)
31. Wang, W., Tran, D., Feiszli, M.: What makes training multi-modal classification networks hard? In: IEEE Conference on Computer Vision and Pattern Recognition (2020)
32. Yang, H.Y., Leahy, B.D., Jang, W.D., Wei, D., Kalma, Y., Rahav, R., Carmon, A., Kopel, R., Azem, F., Venturas, M., et al.: Blastassist: a deep learning pipeline to measure interpretable features of human embryos. Human Reproduction p. deae024 (2024)