Accurate Machine Learning Model for Human Embryo Morphokinetic Stage Detection

Misaghi, H[1]; Cree, L[1]; Knowlton, N[1,2]*

[1] Department of Obstetrics, Gynaecology & Reproductive Sciences ,University of Auckland, , New Zealand

[2] School of Mathematical and Computational Sciences, Massey University, New Zealand

** Corresponding Author*

Abstract

Purpose

The ability to detect, monitor, and precisely time the morphokinetic stages of human embryos plays a critical role in assessing their viability and potential for successful implantation. In this context, the development and utilization of accurate and accessible tools for analysing embryo development are needed. This work introduces a highly accurate, machine learning model designed to predict 16 morphokinetic stages of pre-implantation human development, which is a significant improvement over existing models. This provides a robust tool for researchers and clinicians to use to automate the prediction of morphokinetic stage, allowing standardisation and reducing subjectivity between clinics.

Method

A computer vision model was built on a public dataset for embryo Morphokinetic stage detection containing approximately 273,438 labelled images based on Embryoscope/+© embryo images. The dataset was split 70/10/20 into training/validation/test sets. Two different deep learning architectures were trained and tested, one using efficient net V2 and the other using efficient-net V2 with the addition of post-fertilization time as input. A new postprocessing algorithm was developed to reduce the noise in predictions of the deep learning model and detect the exact time of each morphokinetic stage change.

Results

The proposed model reached an overall test accuracy of 87% across 17 morphokinetic stages on an independent test set. If only considering plus or minus one developmental stage, the accuracy rises to 97.1%.

Conclusion

The proposed model shows state-of-the-art performance (17% accuracy improvement compared to the best models on the same dataset) to detect morphokinetic stages in static embryo images as well as detecting the exact moment of stage change in a complete time-lapse video.

Keywords: embryo morphokinetics, deep learning, time-lapse imaging, Machine learning, Artificial Intelligence

Introduction

Time-Lapse Imaging (TLI) incubators were first introduced into human *in vitro* fertilisation (IVF) clinical laboratories around 2010, to facilitate embryo monitoring [1]. Monitoring embryonic development through the capture of periodic images of each embryo, taken 5-15 minutes apart, holds promise, as the timings of specific developmental events has demonstrated associations with implantation potential, reviewed in [2]. Static images taken minutes apart can be assembled into a comprehensive video, chronicling the embryo's *in vitro* progression. This offers a dynamic perspective on embryo development and growth.

TLI can therefore be regarded as an advanced approach for enhancing the assessment of embryo quality, thereby aiding the ability to select, deselect and rank embryos and potentially decreasing the time to pregnancy during IVF treatments [3] [4] [5] Numerous studies have subsequently identified an association between various morphokinetic parameters (MK)K parameters and the likelihood of implantation and ploidy [6][7][8]. [3][4][5]Additionally, the volume of high-resolution imagery generated by TLI offers opportunities for the application of Machine Learning (ML) and Artificial Intelligence (AI) techniques, including Deep Learning (DL), to increase reproducibility and decrease human effort.

Several methodologies have been proposed to provide an accurate assessment of embryo viability, operating on the premise of accurately annotated timings of morphokinetic events[9] [10] . Rubio et al. conducted the first randomised control trial to determine the efficacy of a multivariable morphokinetic model on success rates; the authors found a significant increase in implantation and ongoing pregnancy rates and a significant decrease in early pregnancy loss for the cohort utilising the integrated incubator and multivariable models based on morphokinetic timings [11]. Soon thereafter, computational combinations of morphokinetic timings were combined into a Known Implantation Determination Day3 Score (KID D3 Score) which utilises a decision tree on timings t2, t3, and t5 to predict embryo viability [12]. However, manual annotation of these events is both labour-intensive and subjective, necessitating automation to alleviate the time burden and increase reproducibility. Reliability of morphokinetic stage annotation is known to be variable, with good agreement at specific time points including t2,3 and t4 and less agreement at tPNa (time of pronuclei appearance) and t9+ [12] [13] . Therefore, the wealth of high-resolution imagery generated by TLI offers opportunities for the application of Machine Learning (ML) and Artificial Intelligence (AI) techniques, including Deep Learning (DL), to increase this reproducibility and decrease human effort.

Automatic Morphokinetic stage detection

Machine learning methodologies have been employed in different stages of embryo selection processes. Most of these models focus on directly predicting the success rate of an embryo to reach the foetal heartbeat stage [14]. Some tools reported surpassing embryologists in the accuracy of identifying viable embryos[15]. There are three main approaches to utilise TLI for embryo selection, models that work directly with the video generated from TLI images and no frame selection is needed by user [15] [16], methods that only work on single image and user need to select specific frame of the TLI images and use that image for prediction [17][18], and there is a hybrid approach where timings are extracted manually and then fed into a ML model such as Kidscore [19] [9].

While approaches such as IDAScore V1 [15] and V2 [16] utilize embryo time-lapse videos, others including ERICA [20], The Life whisperer [17], and Stork [18] rely on a static single image of the blastocyst. The pipeline for selecting the blastocyst image is not entirely automated for ERICA [20], LIFE whisperer [17], and Stork [18], necessitating manual selection. However, the automation of

morphokinetic event detection can streamline the entire process by identifying the most appropriate images and incorporating them into the pipeline as a hybrid approach, where timings are extracted and then fed into a ML model such as Kidscore [19] [9].

Early methods of automatic morphokinetic stage detection relied on manually designed features to identify morphokinetic events [21]. For example, Feyeux et al. used grey-level analysis of microscopic images to predict morphokinetic stages [10]. In recent years, however, modern approaches have primarily utilized deep learning techniques [14], focusing on the supervised training of convolutional neural networks to facilitate the automatic annotation of morphokinetic events. Models employing ResNet-50, Long Short-Term Memory (LSTM), and ResNet-3D type models are now commonly adopted [3]. While the models that predict Morphokinetic events show similar levels of accuracy and overall performance, they are developed on proprietary datasets that are not available. Two prominent examples are the commercial models for EmbryoScope [21] and Gerry Incubators [22]. Zabari et al. also proposed a DL methodology that leverages video frame-based initial predictions, which are further refined through Monotonic regression. This approach aims to mitigate prediction noise by limiting predictions to either the current or an immediately subsequent developmental class [21]. Recently, Gomez *et al.* have introduced a large, annotated dataset, comprising 704 videos of developing embryos, featuring 337,000 images across 16 developmental stages. This dataset provides a critical resource for model benchmarking and development. In this study two different DL models were developed and trained on Gomez et.al dataset [3]. a transformer [23] architecture will be used to fuse the input image features and information about time post-fertilisation when the image was captured. This novel, multimodal methodology incorporates both embryo images and the elapsed time since fertilisation as inputs to enhance the model's performance.

Material and Method

Dataset

Time-lapse imaging of human embryos was obtained from Gomez et al. [3]. The dataset contains 704 Embryoscope videos recorded at 7 focal lengths and annotated for 16 morphokinetic events (Vitrolife ©). This dataset is labelled by one embryologist. To label the videos, the embryologist first identified the frame in which each event occurs and assigned the label to these frames. Then, all subsequent frames until the next morphokinetic event occurs are assigned the current label [3].

Each video has on average, 8 or more events. 499 of the videos are of viable embryos, with the remaining 205 non-viable, to attempt to capture the myriad features that can cause embryonic development failure, such as high levels of fragmentation, 3PN, necrosis, etc. Only the central focal plane was used and only files that had uncorrupted jpeg images were analysed, resulting in 273,438 images with labelled events.

In order to mitigate some of the subjectivity of morphokinetic transitions effect, two images forward and backward from the recorded transition time were removed from the training set. All images were evaluated in the test set and included in performance metrics.

During the dataset quality review, it was noted that, as is standard practice, the embryos are removed for freezing or transferring on Day 5; however, the label still reflects the last event, i.e. expanded blastocyst. This mislabelling of the wells introduces ground truth errors in the dataset. To identify and relabel these images, an empty/non-empty model was developed and applied to all images(Appendix 1). The results were verified visually, and 9,734 images were labelled as empty. Additionally, it was noted that there were only 41 hatched blastocyst images (Table 1). This small number of examples makes predicting this class difficult and subject to increased variability.

Table 1 presents the data annotations along with the number of samples corresponding to each class within the new fixed dataset.

Table 1 Updated Gomez dataset definitions and number of samples for each class in each train, test, and validation set.

Annotation	Description	Number in Training set	Number in Test set	Number in Validation set
tPB2	Polar body appearance	5641	1737	880
tPNa	Pronuclei appearance	27762	8156	3839
tPNf	Pronuclei disappearance	4411	1279	643
t2	Cleavage stage 2 cells	18968	5417	2825
t3	Cleavage stage 3 cells	2825	958	570
t4	Cleavage stage 4 cells	18483	5366	2815
t5	Cleavage stage 5 cells	5009	1336	607
t6	Cleavage stage 6 cells	5154	1657	758
t7	Cleavage stage 7 cells	7094	1810	721
t8	Cleavage stage 8 cells	20133	5276	3250
t9+	Cleavage stage more than 9 cells	31241	9383	4579
tM	Morula	10402	3061	1608
tSB	Start of Blastulation	10206	2955	1709
tB	Blastocyst	4796	1096	556
tEB	Expanded blastocyst	11002	3928	1761
tHB	Hatched blastocyst	32	9	0
Empty	Empty well	7113	1847	774

Deep learning models architecture

For model 1, the backbone is an efficient net V2 large [24] fine-tuned to categorize 17 morphokinetic classes (Table 1). The input to the model was static greyscale JPG images of 380x380 resolution and the greyscale values were copied 3 times into the R,G,B channels. Weights were initialised with ImageNet dataset [24]. Figure 1 (a) illustrates the structure of this network.

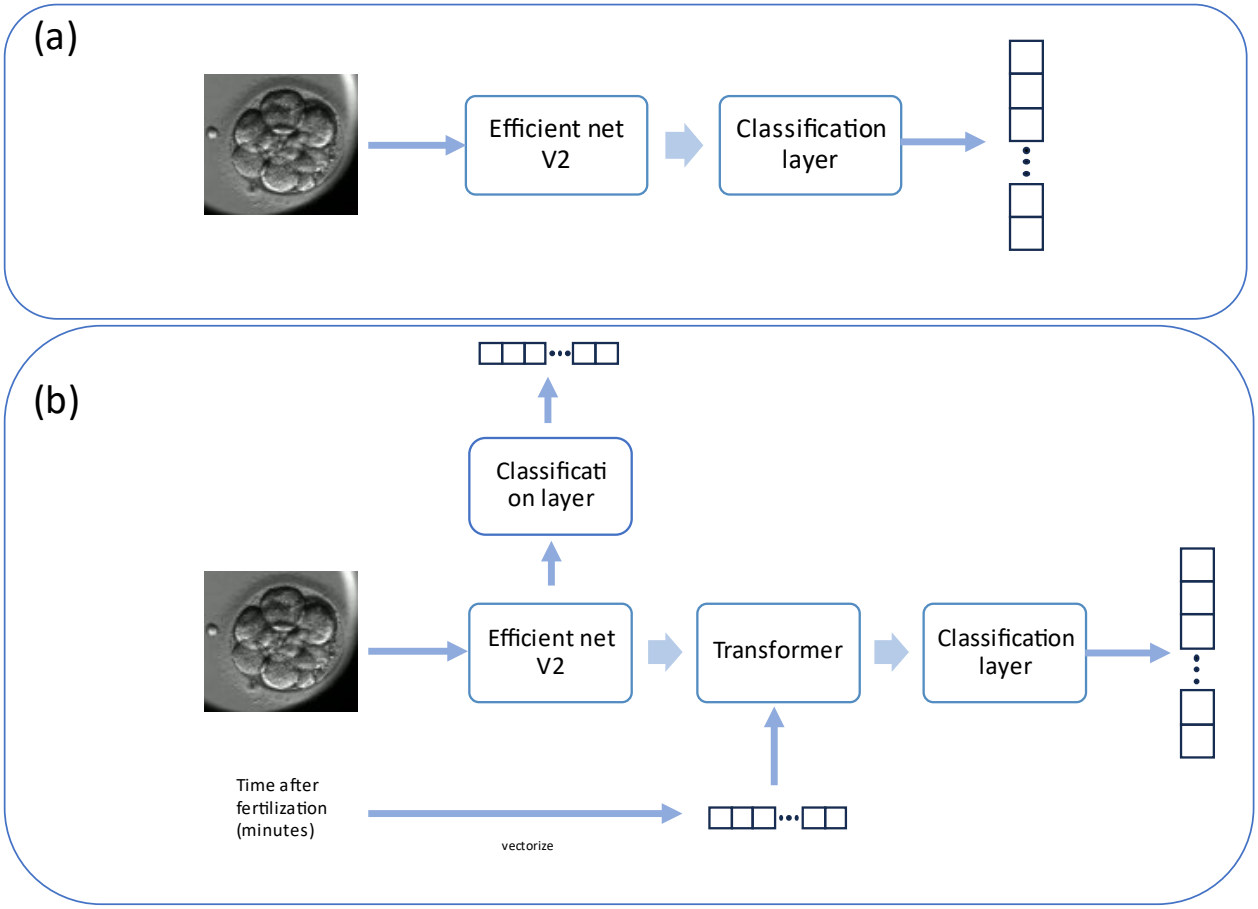

Figure 1 Architecture of the proposed models (a) Model 1, utilises efficient-net-V2-large as backbone for feature extraction and it has a fully connected layer as classifier head. (b) Model 2, has efficient net-large-V2 as feature extractor, this model has two classifier heads one after the feature extraction step and one after the transformer. The transformer in model 2 fuses the time after fertilization and extracted features from the efficient-net. The two head structure of this network ensures the proper gradient flow to backbone.

Model 2 also utilises an Efficient-net V2 L as the backbone for processing images, followed by a transformer [23] . The transformer has only one encoder layer with hidden size of 512 with 4 self-attention heads. The second input to the network is the number of minutes that have elapsed since the time of fertilization. As transformers are designed to work on sequences of vectors, the time since fertilisation is converted into a binary vector representing two-hour windows of time. This approach allows the self-attention mechanism to interact with the image features extracted earlier in the pipeline. Specifically, we assume a maximum of seven days incubation time and encode every two hour window with a one-hot vector or length resulting in a vector of length 84.

$$\text{index} = \left\lfloor \frac{\text{minutes}}{60 \times 2} \right\rfloor$$
$$\mathrm{v}_{\text{index}} = 1$$
$$\mathrm{V}_{\text{time}} = [\mathrm{v}_1, \mathrm{v}_2, \dots, \mathrm{v}_{\text{index}}, \dots, \mathrm{v}_{84}]$$

Where all v elements are 0 except of $\mathrm{v}_{\text{index}}$. For example, if the time is 150 minutes, the index will be calculated as $\left\lfloor \frac{150}{60\times2} \right\rfloor = 1$ the time vector will be $[0, 1, 0, 0, \dots, 0]$, the time vector will be the same for 160 minute, but for 240 minutes the index will be $\left\lfloor \frac{240}{60\times2} \right\rfloor = 2$ and the time vector will be $[0, 0, 1, 0, \dots, 0]$.

The time vector and the image features generated by the backbone are passed to the transformer layer (Figure 1b) layer before a final Multilayer Perceptron (MLP) layer to make class predictions for each image. By design, this model has two classification heads, one immediately after the backbone and one after the transformer layer. During the training, two cross-entropy losses are summed. This ensures the proper gradient for backbone layers and prevents the model from overfitting only on the time vector inputs.

Model 3 has the same architecture as Model 1, but was trained on the original Gomez dataset that did not include the reclassified 'empty' well labels. Model 4 was trained with Resnet-50 backbone to recreate the previous work by Gomez et al [3] to serve as a baseline model for comparison.

All models were trained using the Adam optimizer, as provided in the PyTorch 2.0.0 library [25], with an initial learning rate of 0.001 with Cross-entropy loss function. The learning rate was dynamically reduced upon observing a plateau decrease in the validation set performance. The networks underwent training for 50 epochs on the training subset of the dataset and were subsequently evaluated on a test subset, which remained unseen during the training phase.

The issue of class imbalance represents a significant challenge within this dataset, as illustrated in Table 1. Such imbalance can induce bias in the neural network towards classes with a higher proportion of samples. To mitigate this issue inverse class frequency method was used. This technique enables the model to generate regulated gradients for the classes that are represented by fewer samples.

Embryo images often exhibit significant variations in brightness, with some regions appearing overly bright and others notably dark, while the embryo is typically positioned centrally therefore proper image normalization is crucial. The image augmentation techniques employed during the training included rotating and shifting the image with a 30% probability, flipping the image with a 50% probability, and applying noise or blur with a 50% probability, followed by a final normalization step to ensure consistent image quality across the dataset. The augmentation step was conducted using the Albumentations library in Python [26]. The normalization step was performed using image contrast enhancement (CLAHE) and the normalize function, with average pixel values across RGB channels set as (0.485, 0.456, 0.406) and the standard deviation as (0.229, 0.224, 0.225) to align with the ImageNet weights.

Postprocessing algorithm

The morphokinetic annotation of videos critically depends on the consistency of predictions on all video frames. Given the dynamic nature of embryos, there are instances when the classification of a frame is ambiguous, resulting in "noisy" predictions. It is essential to accurately identify the precise moment when the morphokinetic stage of the embryo changes because these exact time values are utilized in subsequent methods that assess the embryo's viability, such as KidScore [12]. A heuristic method was developed to address this issue. This method is applied to the predictions made by networks for each frame of the embryo's timelapse video. The algorithm accommodates shifts to more advanced morphokinetic classes as well as instances of reverse cleavage. It identifies a trend within each time interval and corrects mispredictions that disrupt this trend. Mispredictions often occur due to embryo movement and the movement of cells within it, where some cells may temporarily be obscured in the image, rendering any detection of morphokinetic state changes during these moments inaccurate. Consequently, the algorithm initially amends the prediction in our models by substituting low-confidence predictions with the last morphokinetic state prediction with over 80% confidence. Subsequently, it detects changes by comparing the predicted class with the

prediction of the preceding image. Let P_I represent the value of the most probable predicted class at the index i, changes are defined as Δ :

$$\Delta(i) = \begin{cases} 1 \text{ if } & P_i \neq P_{i-1} \\ 0 & \text{otherwise} \end{cases}$$

By summing the values of changes within a sequence, various consecutive groups can be delineated as follows:

$$G_i = \sum_{k=0}^{i} \Delta(k)$$

The values of G_I can be utilized for grouping purposes, indicating that all images possessing identical G_I values belong to a contiguous group and share the same prediction values without interruption. Through the comparison of the length of each continuous group and its neighbouring groups, a determination is made regarding whether the current group should be regarded as noise or retained as a correct morphokinetic stage change. The algorithm is outlined in the following:

- Let $C = \{c1, c2, \dots, cn\}$ represent the set of unique predicted classes in the video.
- For each class c in C, create a subset Dc of all the predictions D such that all elements in Dc have prediction class equal to c. or each subset Dc, identify unique group values as $Gc = \{g1, g2, \dots, gm\}$.
- Calculate the length of each consecutive group g in Gc, denoted as Lg, where Lg is the number of elements in Dc that belong to group g.
- Determine the minimum and maximum values within G_c, denoted as g_{min} and g_{max}, respectively.
- For each potential group identifier i_g in the range $[g_{min}, g_{max}]$, identify if i_g is an interruptive group by checking if $ig \notin Gc$. For each interruptive group, calculate its length L_{ig} .
- All interruptive groups are discarded and only the main groups are kept as final labels

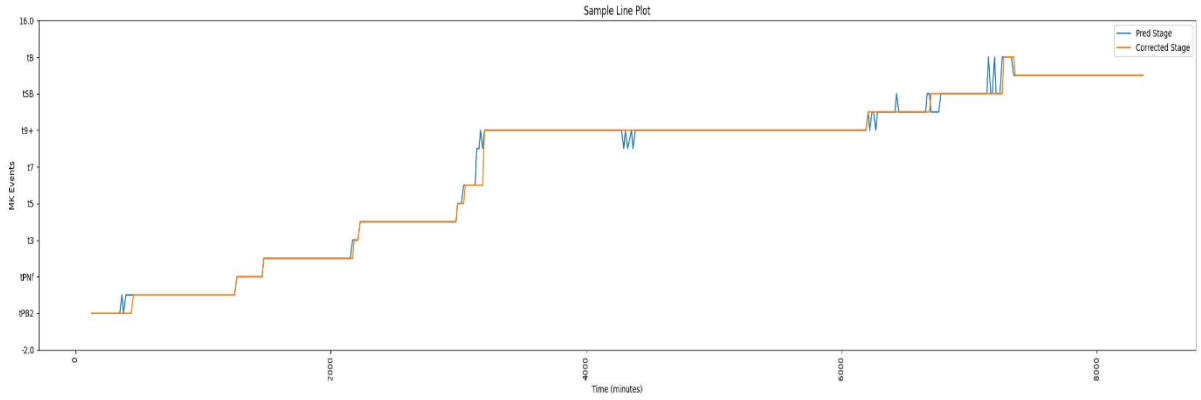

Figure 2 Example of the effect of postprocessing algorithm on predictions of Model 2. Model predictions are uncertain in some moments of the video, specifically near the moment that a morphokinetic stage change is recorded. The postprocessing algorithm ensures a clean set of predictions by assessing and ignoring noisy predictions

Results

After 50 epochs of training, Model 1 showed an accuracy of 93% while Model 2 showed an accuracy of 95%; Models 3 and 4 had an accuracy of 93%, and 96%, respectively, on the training set.

Next, each model's network was evaluated on a subset of data which remained unseen by the models during the training phase (test set). To properly assess the models and postprocessing algorithm the evaluation steps are separated, and results are presented in two sections:

1. **Single Image Processing:** results of deep learning models on single images that focuses on classification performance of the models on single images.
2. **Postprocessing:** results after applying the model on a video created by static time lapse imaging and applying the postprocessing algorithm to extract the exact time of morphokinetic stage changes.

Single Image Processing

During the testing step for the Models, standard classification metrics were calculated. All the images in the test dataset were processed by the models and the outputs compared with ground truth in the dataset. Confusion matrix, accuracy, F1-score, precision, and recall were calculated for each model. Figure 3 presents the confusion matrix of Model 2 applied to the test set of the dataset. It is evident that certain stages, such as t2 and t4, are easier to identify, whereas other stages, including tB and t5, are more challenging for the model to classify accurately. Misclassifications most commonly occur between consecutive stages, such as when images of class t5 are misclassified as either t4 or t6. It is highly unlikely for the model to confuse an image with a class significantly ahead or behind; for instance, an image labeled as t5 has less than a 1% chance of being classified as anything other than t4 or t6.

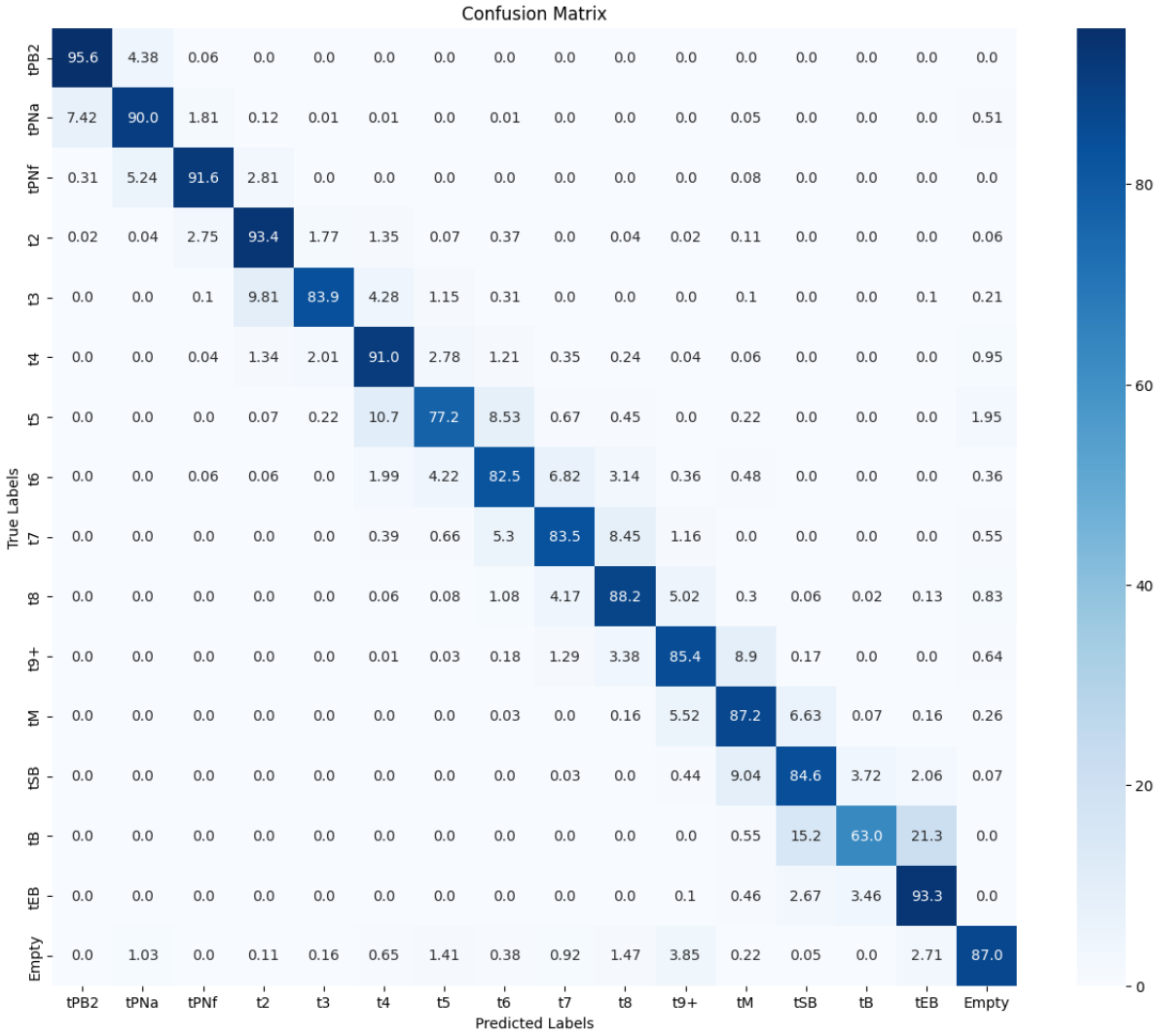

Figure 3 Model 2 confusion matrix on the test set in percentages. The blue gradient bar on the right of the matrix indicates the percentage graphically with 0% - white and 100% - dark blue. Values of 0.0 indicate that not a single miss classification occurred and perfect model prediction would be 100% down the lower right diagonal.

Standard classification metrics for the proposed models are shown in Table 2. Both exhibit high accuracy levels, with Model 2 demonstrating a slightly superior performance on all the metrics.

Table 2 Model performance of the top two models across a range of performance metrics. Models are down the rows and metrics are across the columns. Accuracy is the overall accuracy.

Model	Precision	Recall	F1-Score	Accuracy
Model 1	0.883	0.871	0.873	0.871
Model 2	0.886	0.879	0.881	0.879

Table 3 presents the accuracy of the models trained by Gomez et al. alongside those proposed in this study. Model 1 and 2 demonstrate an approximate 17% improvement in accuracy compared to the models in the original dataset paper [3]. Models 1 and 2 have the best performance, and models 3 and 4, which did not have empty labels in their dataset, have less accuracy. It can also be seen that Model 4 which was trained to represent the Gomez work is performing only slightly better than resnet-50 that was reported by Gomez et al [3].

Table 3 Comparison of accuracy of methods in Gomez et al [3] *and proposed methods trained on same dataset. Models prefixed with Gomez is reported performance on this dataset for comparison purposes. Models 1 and 2 were developed to improve discriminatory accuracy. Models 3 and 4 are comparable to the Gomez models directly, as they don't include our updated class of "empty well".*

Method	Test set Accuracy
Gomez ResNet	0.663 ± 0.041
Gomez ResNet-LSTM	0.685 ± 0.041
Gomez Resnet-3D	0.705 ± 0.036
Model 1 (Efficient net)	0.871
Model 2 (Efficient-net-transformer modality)	0.879
Model 3 (Efficient net, without empty well labels)	0.739
Model 4 (Resnet-50, without empty well labels)	0.689

To visually assess the root causes of misclassification of the models, a random sample of images that were mis-predicted by Model 2 for the classes that had highest error rate are shown in Figure 7 to Figure 10.

Postprocessing

After applying the postprocessing method the exact predicted times for each morphokinetic event is extracted. For each detected morphokinetic stage change one single time is extracted. In the analysis here only the resulting morphokinetic state is considered, for example, changes from t2 to t4, and t3 to t4 are considered as transition to t4.

While the model results were accurate, a plus or minus one developmental stage classifications was observed in a sequential set of images (Figure 3). This is to be expected due to the embryos moving around the incubation dish i.e., not fixed, exposing or obscuring different features. To combat this, a heuristic post-processing was implemented.

Upon applying the model and the subsequent postprocessing algorithm to the entire video in the test dataset, the timings for changes in the MK stage were extracted. Error is calculated as $time_{prediction} - time_{ground\ truth}$ the average and standard deviation and 25%, 50%, 75%, 80%, 85%, 90%, 95%, and 99% percentile of the errors belonging to each class is calculated and presented in Table 4. It is evident that errors are small for the majority of cases and large errors are seen in less than 5% of cases. The largest errors belong to tB class.

Table 4 Quantile analysis of timing prediction errors for each morphokinetic class, count shows the number of events that a transformation to the class was detected, mean is the average error and std is standard deviation for the class. The quantiles of 25, 50, 75, 80, 85, 95, and 99 percent are also reported for each class.

event	tPNa	tPNf	t2	t3	t4	t5	t6	t7	t8	t9+	tM	tSB	tB	tEB
count	140	121	136	51	114	66	68	70	112	144	109	95	27	60
mean	0.54	-0.01	0.26	0.23	-0.74	-0.87	-0.45	-0.07	-0.78	-2.94	-1.03	0.61	1.61	-1.51
Std	3.43	0.93	1.40	2.01	5.22	3.31	1.65	1.36	3.86	7.84	3.45	2.33	6.10	5.89

25%	0.00	-0.30	0.00	0.00	-0.30	-0.45	-0.57	-0.30	-0.85	-3.08	-3.00	-0.30	-0.15	-1.43
50%	1.00	0.00	0.20	0.00	0.00	0.00	0.00	0.00	0.00	0.00	-0.80	0.20	1.00	-0.40
75%	2.00	0.20	0.30	0.30	0.30	0.20	0.30	0.60	0.70	0.85	0.20	1.10	2.60	0.70
80%	2.04	0.20	0.50	0.30	0.50	0.30	0.46	0.72	0.70	1.20	0.74	1.82	3.10	1.02
85%	2.50	0.30	0.50	0.70	0.70	0.53	0.60	1.00	1.07	1.76	1.28	3.00	7.15	1.41
90%	2.61	0.60	0.80	1.00	1.20	0.85	1.00	1.20	1.50	2.67	1.74	3.88	9.10	2.05
95%	3.21	1.70	1.78	1.85	2.24	1.43	1.63	1.50	2.84	4.29	3.46	4.56	11.40	4.03
99%	10.79	2.44	3.71	7.05	9.51	4.84	1.80	2.00	7.37	8.30	6.10	6.17	12.89	8.20

Time extraction had less average error in some classes such as tPNf with -0.01 hours average error while the error is more significant in t9+ with -2.94 average hour error. It is important to note that images are 20 minutes apart and only one frame error contributes as 0.33 error.

Due to the significant impact of labeling errors on the accuracy of stage change time extraction, we conducted an in-depth assessment of label subjectivity and identified the primary causes of model misclassification. Figures 4 and 5 provide examples of the largest errors observed, showcasing a random sample of frames from classes with the greatest time detection discrepancies. These visual examples help highlight the inconsistencies in labeling and the challenges they pose for model accuracy.

Discussion

Morphokinetic timings have been linked to predicting the development of embryos to the blastocyst stage [6], [19] and their implantation potential [6]. However, current models largely rely on subjective manual annotations, which may limit their accuracy and reproducibility. The association between morphokinetic parameters and developmental outcomes could shift if more objective and standardized timings were utilized; unfortunately, few models employing such objective measurements are available [6] [27]. Manual annotation is inherently subjective, and often only a limited set of annotations are performed, potentially omitting critical data. To advance the field of embryo morphokinetic research, it is crucial to develop models that are tested and validated on open-source datasets. While commercially available morphokinetic models like Fairtility [21] and the Gerry TLI system [22] exist, they are proprietary, and their performance has not been assessed on publicly available datasets, limiting their transparency and comparability.

Table 3 provides a comparative analysis between the models introduced in this study and those developed by Gomez et al. [3], highlighting the superior accuracy of Model 2 with respect to performance with 87.9 % accuracy on single frames. This is consistent with the levels reported by Zabari et al. [21] (94%); however, their model has several shortfalls. Firstly, they can't detect reverse cleavage or any reduction in Morphokinetic stage. Secondly, they remove the most difficult to assess stages such as 9 cell, and intra-blastocyst stages (early, expanding, hatching, etc). Both of these decisions increase their accuracy overall for a 'perfect' embryo but remove the ability to detect potentially meaningful associations with embryos of poorer quality. Finally, without evaluating models on the same dataset, the performance metrics should be interpreted as a general indication and not directly compared. In addition, their study contained fewer classes than the dataset utilized in this study, and their dataset consisted of 20,253 labelled embryos. Although the dataset used to test the model from the Zabari et al study is not publicly available, the current Gomez class definitions were altered to be aligned with Zabari study and recalculated the accuracy and

morphokinetic timing detections to be able to compare the model with their study. Using the Zabari et al. defined labels, the accuracy of our Model 2 is 91 % on the test set. This illustrates how these models are quickly converging to the Zabari model with 100 times less data. As described in Table 3, Model 3 has the same architecture and training settings as Model 1. While the model achieved 73.9% accuracy on the test set of the original Gomez dataset, by adding the empty well class, which is only 3.5% of the images in the dataset, the accuracy increased to 87% for Model 1. This shows how the mislabel affected the whole network's learning process. While the empty well images accounted for 3.5% of the dataset, fixing them improves the accuracy by almost 9%.

Misclassifications of stage are clustered around stages that are likely to cause confusion with human annotators when only shown a single focal plane image, such as tPNa, t5, t7, tM, tEB, and tHB. The inaccuracy of the network on these classes is higher and disagrees with the ground truth labels; however, there is more consistency in the model calls than amongst human labellers [28]. For example, the difference between an expanded blastocyst (tEB) and blastocyst (tB) is more subjective than the difference between a one-cell (tPNf) and a two-cell (t2) embryo. The labelling accuracy is very important for tPNa and tPNf phases because although there is a small difference in visual features, biologically one starts a biological processes and mislabelling leads to model confusion. Examples for such mislabelling in the dataset is demonstrated in Figure 4. This is also the case for other classes such as what denotes an expanded blastocyst (Figure 6). It is evident that subjectivity between tEB and tB classes and mislabelling in classes such t4 and t5 exists. Despite these "errors" most images in the dataset have correct labels, the network has learned the classes properly and generalizes well on the test set.

The postprocessing algorithm proposed here does not limit the predictions of the embryo stages to be monotonically increasing as opposed to both Gomez et al [3], and Zabari. [21] that only allow later time predictions to be of the same or more advanced morphokinetic stages during a time-lapse video. Thus, annotations provided by our model can be used to detect how often and study the effect reverse cleavage might have on embryo viability [29].

The difference between the performance of model 1 and model 2 is within a margin of error, as the difference in accuracy is only 0.6 %; it shows that the features existing in the images are sufficient for morphokinetic stage detection and adding the data about the post-fertilization time of each image does not improve the results. It is unclear why this is the case, given the strong relationship between time from fertilisation and developmental stage, but a likely cause is the 100s of embryos in our data set that fail to develop into blastocysts, reducing the predictive ability of time. This is a positive result and lends support to the potential for Model 1 to generalise across clinics with different media and embryo conditions that might affect developmental timings.

The precise timing of MK stages holds significant interest, and through the application of proposed models and post-processing algorithms, it was possible to extract these timings accurately; the errors predominantly fall within a 1-hour range for the worst class(tB), which is equivalent to 3 frames, given the imaging time steps of every 20 minutes. The worst classes in morphokinetic timing detections are tB, with an average error of 1.6 hours (96 minutes). While this sounds high, the "incorrect" predictions are predominantly 3.7% tSB and objectively defining the transition point to blastulation using only the centre focal plane is challenging. Current commercial software, which uses 150-400 min, suggests that this time frame of tB detection is acceptable to current lab practice [22].

Future open-source work on the Gomez dataset could help improve the annotations further, enhancing morphokinetic determinations across the field. For example, new annotations introduced here included wells that are too dark to observe the embryo and those without embryos as “empty”. Using the Gomez dataset as an objective benchmark for commercial morphokinetic software would provide a robust, repeatable metric to assess this class of model. Additionally, it would provide insight into which features, stages, or images are most difficult to predict on a model-by-model basis.

There are several limitations to the current approach. The model relies on human-annotated ground truth labels and as we showed in Fig 7 and Supp Fig 1,2,3. While we cleaned up the labels a small amount by labelling the empty and dark wells as “empty”, there is likely not a consensus embryo staging for each image, especially during stage transitions, i.e. t4-t5. Our mitigation attempt was based upon removing two images before and after the transition during training but predicting these transitions during evaluation on the test set.

Conclusion

This study introduces a highly accurate machine learning model for detecting the morphokinetic stages of human embryos, significantly advancing in vitro fertilization (IVF) technology. The inclusion of a novel post-processing algorithm, which is not constrained by monotonicity, allows for the detection of reverse cleavage events, providing a more nuanced understanding of embryo development. By automating and enhancing the annotation of morphokinetic stages—traditionally a subjective and labour-intensive process—the model improves dataset quality and performance. This model outperforms previous methodologies, potentially streamlining embryo selection in clinical practice and thus decreasing the time of pregnancy.

Acknowledgements

This work was funded by MBIE Smart Ideas UOAX2112. The authors have no competing interests to declare. Ethics for the empty model were approved by the Auckland Ethics Research Committee (AHREC #AH1033). All authors contributed to the study conception and design. Material preparation and analysis were performed by HM. The first draft of the manuscript was written by HM and all authors commented on previous versions of the manuscript. All authors read and approved the final manuscript.

References

[1] C. C. Wong *et al.*, "Non-invasive imaging of human embryos before embryonic genome activation predicts development to the blastocyst stage," *Nat. Biotechnol.*, vol. 28, no. 10, pp. 1115–1121, 2010, doi: 10.1038/nbt.1686.

[2] D. K. Gardner and B. Balaban, "Assessment of human embryo development using morphological criteria in an era of time-lapse, algorithms and 'OMICS': Is looking good still important?," *Mol. Hum. Reprod.*, vol. 22, no. 10, pp. 704–718, 2016, doi: 10.1093/molehr/gaw057.

[3] T. Gomez *et al.*, "Towards deep learning-powered IVF: A large public benchmark for morphokinetic parameter prediction," Mar. 2022, [Online]. Available: http://arxiv.org/abs/2203.00531

[4] C. Pribenszky, A. Nilselid, M. Montag, and G. Se-, "Time-lapse culture with morphokinetic embryo selection improves pregnancy and live birth chances and reduces early pregnancy loss : a meta-analysis," *Reprod. Biomed. Online*, vol. 35, no. 5, pp. 511–520, 2017, doi: 10.1016/j.rbmo.2017.06.022.

[5] R. J. Paulson *et al.*, "Time-lapse imaging : clearly useful to both laboratory personnel and patient outcomes versus just because we can doesn ' t mean we should," *Fertil. Steril.*, vol. 109, no. 4, pp. 584–591, 2018, doi: 10.1016/j.fertnstert.2018.01.042.

[6] C. Serrano-Novillo, L. Uroz, and C. Márquez, "Novel Time-Lapse Parameters Correlate with Embryo Ploidy and Suggest an Improvement in Non-Invasive Embryo Selection," *J. Clin. Med.*, vol. 12, no. 8, p. 2983, 2023, doi: 10.3390/jcm12082983.

[7] T. Bamford *et al.*, "Association between a morphokinetic ploidy prediction model risk score and miscarriage and live birth: a multicentre cohort study," *Fertil. Steril.*, pp. 108–112, 2023, doi: 10.1016/j.fertnstert.2023.06.006.

[8] D. E. Fordham *et al.*, "Embryologist agreement when assessing blastocyst implantation probability: is data-driven prediction the solution to embryo assessment subjectivity?," *Hum. Reprod.*, vol. 37, no. 10, pp. 2275–2290, 2022, doi: 10.1093/humrep/deac171.

[9] E. Gazzo *et al.*, "Original article The Kidscore TM D5 algorithm as an additional tool to morphological assessment and PGT-A in embryo selection : a time-lapse study," vol. 24, no. 1, pp. 55–60, 2020, doi: 10.5935/1518-0557.20190054.

[10] M. Feyeux *et al.*, "Development of automated annotation software for human embryo morphokinetics," *Hum. Reprod.*, vol. 35, no. 3, pp. 557–564, 2020, doi: 10.1093/humrep/deaa001.

[11] I. Rubio *et al.*, "Clinical validation of embryo culture and selection by morphokinetic analysis: A randomized, controlled trial of the EmbryoScope," *Fertil. Steril.*, vol. 102, no. 5, pp. 1287-1294.e5, 2014, doi: 10.1016/j.fertnstert.2014.07.738.

[12] B. M. Petersen, M. Boel, M. Montag, and D. K. Gardner, "Development of a generally applicable morphokinetic algorithm capable of predicting the implantation potential of embryos transferred on Day 3," *Hum. Reprod.*, vol. 31, no. 10, pp. 2231–2244, 2016, doi: 10.1093/humrep/dew188.

[13] E. Adolfsson and A. N. Andershed, "Morphology vs morphokinetics: A retrospective comparison of interobserver and intra-observer agreement between embryologists on blastocysts with known implantation outcome," *J. Bras. Reprod. Assist.*, vol. 22, no. 3, pp. 228–237, 2018, doi: 10.5935/1518-0557.20180042.

[14] M. Salih *et al.*, "Embryo selection through artificial intelligence versus embryologists: a systematic review," *Hum. Reprod. Open*, vol. 2023, no. 3, 2023, doi: 10.1093/hropen/hoad031.

[15] J. Berntsen, J. Rimestad, J. T. Lassen, D. Tran, and M. F. Kragh, "Robust and generalizable embryo selection based on artificial intelligence and time-lapse image sequences," *PLoS One*, vol. 17, no. 2 February, pp. 1–18, 2022, doi: 10.1371/journal.pone.0262661.

[16] J. Theilgaard Lassen, M. Fly Kragh, J. Rimestad, M. Nygård Johansen, and J. Berntsen, "Development and validation of deep learning based embryo selection across multiple days of transfer," *Sci. Rep.*, vol. 13, no. 1, pp. 1–9, 2023, doi: 10.1038/s41598-023-31136-3.

[17] M. Ver Milyea *et al.*, "Development of an artificial intelligence-based assessment model for prediction of embryo viability using static images captured by optical light microscopy during IVF," *Hum. Reprod.*, vol. 35, no. 4, pp. 770–784, 2021, doi: 10.1093/HUMREP/DEAA013.

[18] P. Khosravi *et al.*, "Deep learning enables robust assessment and selection of human blastocysts after in vitro fertilization," *npj Digit. Med.*, vol. 2, no. 1, pp. 1–9, 2019, doi: 10.1038/s41746-019-0096-y.

[19] N. Basile *et al.*, "The use of morphokinetics as a predictor of implantation: A multicentric study to define and validate an algorithmfor embryo selection," *Hum. Reprod.*, vol. 30, no. 2, pp. 276–283, 2015, doi: 10.1093/humrep/deu331.

[20] A. Chavez-Badiola, A. Flores-Saiffe-Farías, G. Mendizabal-Ruiz, A. J. Drakeley, and J. Cohen, "Embryo Ranking Intelligent Classification Algorithm (ERICA): artificial intelligence clinical assistant predicting embryo ploidy and implantation," *Reprod. Biomed. Online*, vol. 41, no. 4, pp. 585–593, 2020, doi: 10.1016/j.rbmo.2020.07.003.

[21] N. Zabari *et al.*, "Delineating the heterogeneity of embryo preimplantation development using automated and accurate morphokinetic annotation," *J. Assist. Reprod. Genet.*, vol. 40, no. 6, pp. 1391–1406, 2023, doi: 10.1007/s10815-023-02806-y.

[22] J. Vandame *et al.*, "Investigation of the reliability of semi-automatic annotation by the Geri time-lapse system," *Reprod. Biomed. Online*, vol. 45, no. 1, pp. 35–45, 2022, doi: 10.1016/j.rbmo.2022.02.012.

[23] A. Vaswani *et al.*, "Attention is all you need," *Adv. Neural Inf. Process. Syst.*, vol. 2017-Decem, no. Nips, pp. 5999–6009, 2017.

[24] M. Tan and Q. V Le, "EfficientNetV2 : Smaller Models and Faster Training," 2021.

[25] A. Paszke *et al.*, "PyTorch: An imperative style, high-performance deep learning library," *Adv. Neural Inf. Process. Syst.*, vol. 32, no. NeurIPS, 2019.

[26] A. Buslaev, V. I. Iglovikov, E. Khvedchenya, A. Parinov, M. Druzhinin, and A. A. Kalinin, "Albumentations: Fast and flexible image augmentations," *Inf.*, vol. 11, no. 2, 2020, doi: 10.3390/info11020125.

[27] K. Kato, S. Ueno, J. Berntsen, M. F. Kragh, T. Okimura, and T. Kuroda, "Does embryo categorization by existing artificial intelligence, morphokinetic or morphological embryo selection models correlate with blastocyst euploidy rates?," *Reprod. Biomed. Online*, vol. 46, no. 2, pp. 274–281, 2023, doi: 10.1016/j.rbmo.2022.09.010.

[28] C. L. Bormann *et al.*, "Consistency and objectivity of automated embryo assessments using deep neural networks," *Fertil. Steril.*, vol. 113, no. 4, pp. 781-787.e1, 2020, doi: 10.1016/j.fertnstert.2019.12.004.

[29] Y. Liu, V. Chapple, P. Roberts, and P. Matson, "Prevalence, consequence, and significance of reverse cleavage by human embryos viewed with the use of the Embryoscope time-lapse video system," *Fertil. Steril.*, vol. 102, no. 5, pp. 1295-1300.e2, 2014, doi: 10.1016/j.fertnstert.2014.07.1235.

[30] K. He and J. Sun, "Deep Residual Learning for Image Recognition," pp. 1–9.