nature communications

Article

https://doi.org/10.1038/s41467-024-51823-7

Automatic ploidy prediction and quality assessment of human blastocysts using time-lapse imaging

Received: 3 August 2023

Accepted: 15 August 2024

Published online: 05 September 2024

Check for updates

Suraj Rajendran[1,2,3], Matthew Brendel[1,2], Josue Barnes[1,2], Qiansheng Zhan[4], Jonas E. Malmsten[4], Pantelis Zisimopoulos[1,2], Alexandros Sigaras[1,2], Kwabena Ofori-Atta[5], Marcos Meseguer[6], Kathleen A. Miller[7], David Hoffman[7], Zev Rosenwaks[4], Olivier Elemento[1,2], Nikica Zaninovic[4] & Iman Hajirasouliha[1,2]

Assessing fertilized human embryos is crucial for in vitro fertilization, a task being revolutionized by artificial intelligence. Existing models used for embryo quality assessment and ploidy detection could be significantly improved by effectively utilizing time-lapse imaging to identify critical developmental time points for maximizing prediction accuracy. Addressing this, we develop and compare various embryo ploidy status prediction models across distinct embryo development stages. We present BELA, a state-of-the-art ploidy prediction model that surpasses previous image- and video-based models without necessitating input from embryologists. BELA uses multitask learning to predict quality scores that are thereafter used to predict ploidy status. By achieving an area under the receiver operating characteristic curve of 0.76 for discriminating between euploidy and aneuploidy embryos on the Weill Cornell dataset, BELA matches the performance of models trained on embryologists' manual scores. While not a replacement for preimplantation genetic testing for aneuploidy, BELA exemplifies how such models can streamline the embryo evaluation process.

Since the advent of in vitro fertilization (IVF) in 1978, it has served as a key solution for individuals unable to conceive naturally, accounting for over 8 million successful births globally[1]. This procedure involves transvaginal transfer of laboratory-fertilized oocytes into the uterus. A critical determinant of IVF success and minimizing the risk of perilous multiple pregnancies lies in the selection of high-quality, single normal embryos, primarily influenced by their ploidy status[2,3].

Ploidy status, the chromosomal constitution of an embryo, greatly impacts pregnancy outcomes. Euploid embryos, characterized by normal chromosomal counts, typically lead to successful pregnancies, while aneuploid embryos—those with chromosomal aberrations—are associated with miscarriage, failed pregnancies, and chromosomal disorders like Down syndrome or Turner's syndrome. Embryo aneuploidy, which leads to increased miscarriage rates, correlates with advanced maternal age.

Currently, preimplantation genetic testing for aneuploidy (PGT-A) is used to ascertain embryo ploidy status. This procedure requires a biopsy of trophectoderm (TE) cells, whole genome amplification of their DNA, and testing for chromosomal copy number variations. Despite enhancing the implantation rate by aiding the selection of

[1]Institute for Computational Biomedicine, Department of Physiology and Biophysics, Weill Cornell Medicine of Cornell University, New York, NY, USA. [2]Caryl and Israel Englander Institute for Precision Medicine, The Meyer Cancer Center, Weill Cornell Medicine, New York, NY, USA. [3]Tri-Institutional Computational Biology & Medicine Program, Weill Cornell Medicine, New York, NY, USA. [4]The Ronald O. Perelman and Claudia Cohen Center for Reproductive Medicine, Weill Cornell Medicine, New York, NY, USA. [5]Weill Cornell/Rockefeller/Sloan Kettering Tri-Institutional MD-PhD Program, New York, NY, USA. [6]IVI Valencia, Health Research Institute la Fe, Valencia, Spain. [7]IVF Florida Reproductive Associates, Fort Lauderdale, Florida, USA. e-mail: imh2003@med.cornell.edu

euploid embryos, PGT-A presents several shortcomings[4]. It is costly, time-consuming, and invasive, with the potential to compromise embryo viability. Moreover, the test's accuracy can be marred by embryonic mosaicism—the co-existence of aneuploid and euploid cells within the TE—leading to false results, diminished embryo viability, and lower implantation rates[5].

The advent of computer vision in artificial intelligence, along with the accumulation of extensive IVF-related datasets—incorporating images, videos, and clinical outcomes—has spurred the development of automated embryo assessment methods via time-lapse image analysis. For instance, Khosravi et al. designed STORK, a model assessing embryo morphology and effectively predicting embryo quality aligned with successful birth outcomes[6]. Analogous algorithms can be repurposed for embryo ploidy prediction, based on the premise that embryo images may exhibit patterns indicative of chromosomal abnormalities. Chavez-Badiola et al. employed the deep-learning model ERICA to analyze 1231 embryo images to predict ploidy status, achieving a 70% accuracy, with an area under the receiver operating characteristic curve (AUC) of 74%, and a sensitivity and specificity of 54% and 86% respectively. Notably, ERICA predicted a euploid embryo in the top rank in 79% of cases. Limitations included the model's inability to distinguish between single and complex aneuploidies, the assumption of euploidy confirmation at β-HCG ≥ 20 mUI/mL on day 7, and a limited dataset potentially restricting general applicability[7]. Similarly, Barnes et al. devised a machine learning algorithm, STORK-A, to predict embryo ploidy status from a single image at 110 h post insemination (hpi), using time-lapse sequences[8]. Silver et al. speculated that the entirety of video sequences could potentially improve embryo classification accuracy, leading to the development of the UBar CNN-LSTM model, which attained an AUC of 0.82—though on a limited dataset[9]. In another recent study, Lee et al. utilized a two-stream inflated 3D model on 670 image sequences, achieving an AUC of 0.74 in differentiating euploid/mosaic and aneuploid embryos[10].

Analyzing entire time-lapse sequences of embryo development presents a challenge in predicting ploidy status, as not all developmental stages may provide pertinent information. This has led to previous studies focusing on feature extraction from specific developmental periods[11]. Campbell et al. proposed the timing and presence of blastocyst expansion on day 5 as a predictor of ploidy status[12]. However, this criterion's predictive accuracy has exhibited considerable variability across clinics, making it less reliable[13]. Analyzing full embryo development videos could bypass the need to pinpoint relevant timeframes, but the computational cost of training models on vast datasets could compromise performance due to noise. Addressing these challenges, we present BELA—a fully automated ploidy prediction model—that requires only embryo time-lapse sequences and maternal age as inputs. By removing the need for subjective manual annotation, BELA not only streamlines the ploidy prediction process but also fosters broad applicability across different clinical settings.

Table 1 | Characteristics of datasets

Dataset	WCM-Embryoscope	WCM-Embryoscope+	Spain	Florida
Sample size	1998	841	543	869
Ploidy splits	SA: 494 CxA: 588 EUP: 916	SA: 170 CxA: 261 EUP: 410	ANU: 309 EUP: 234	SA: 202 CxA: 222 EUP: 445
Clinical features	Maternal age Blastocyst score ICM score TE score Expansion score	Maternal age Blastocyst score ICM score TE score Expansion score	Maternal age	Maternal age Blastocyst score ICM score TE score Expansion score

The sample size, distribution of data across ploidy classes, and additional clinical features for each dataset are shown.

Results

Training and validation datasets

In our study, we utilized deep-learning techniques to predict ploidy status using time-lapse sequences of embryo development and compared various model performances across multiple clinics. Two internal datasets from Weill Cornell Medicine's Center for Reproductive Medicine (WCM) were employed: the first encompassed 1998 Embryoscope® time-lapse sequences, and the second contained 841 sequences from the Embryoscope+®. These sequences typically constituted 360–420 distinct frames, captured at 0.3-h intervals over 5 days of development. PGT-A results served as the ground truth for ploidy prediction tasks, with embryos classified as euploid (EUP) or aneuploid (ANU). Further categorization of ANU embryos identified single aneuploid (SA)—with one chromosomal abnormality—and complex aneuploid (CxA)—with multiple chromosomal abnormalities. Accompanying clinical information included blastocyst scores (BS)—derived from morphological grades and morphokinetic parameters—and maternal age at oocyte retrieval. BS encompasses three subcomponents: inner cell mass (ICM), trophectoderm (TE), and expansion score[14]. This blastocyst score formulation has been shown to be predictive of implantation success, euploidy, and live birth[14]. For additional model validation, we utilized an external dataset from IVI Valencia, Spain. Unlike the WCM datasets, this dataset only contained EUP/ANU labels without explicit SA/CxA details and BS. A second external dataset from IVF Florida provided additional detail allowing discrimination between SA and CxA embryos. Comprehensive descriptions of these datasets are detailed in Table 1, Supplementary Table 10, and further expounded in the "Methods" section.

Ploidy prediction model with model-derived blastocyst score

We introduce BELA, the Blastocyst Evaluation Learning Algorithm for ploidy prediction, a fully automated model detailed in Fig. 1. The model comprises two steps. First, BELA predicts the blastocyst score (BS) from processed day-5 time-lapse videos (96–112 hpi), a timeframe chosen based on our ablation analyses comparing embryonic development time points and image versus video inputs (Supplementary Note 1). The input video undergoes processing and transformation into feature vectors via a pre-trained spatial feature extraction model (Fig. 1, steps 1–4). To optimize performance, we used a multitasking BiLSTM model to concurrently predict ICM, TE, expansion, and blastocyst score. We evaluated the first component of BELA using the mean absolute error (MAE). In the second step, BELA uses the now 'model-derived blastocyst score' (MDBS) to predict the embryo's ploidy status, employing a logistic regression that integrates maternal age as a continuous input feature, as illustrated in Fig. 1. We trained and evaluated BELA on EUP versus CxA and EUP versus ANU splits. BELA was trained on data from the WCM-Embryoscope dataset via four-fold cross-validation. Performance was gauged using accuracy, AUC, precision, and recall across the datasets from WCM-Embryoscope, WCM-Embryoscope+, Spain, and Florida. For comparison, we trained two baseline models using the same cross-validation splits. The first baseline is a day-5 video model which exclusively uses time-lapse input from 96 to 112 hpi to directly predict ploidy status using a BiLSTM architecture. The second baseline is an embryologist-annotated model that uses only the ground-truth BS to predict ploidy status using logistic regression.

The first component of BELA predicts the blastocyst score (BS). As depicted in Supplementary Fig. 1, both the training and test sets from WCM-Embryoscope show a moderate correlation (Pearson correlation > 0.7) between the model-derived blastocyst scores (MDBS) and the embryologist BS. This moderate correlation is also evident in the

https://doi.org/10.1038/s41467-024-51823-7

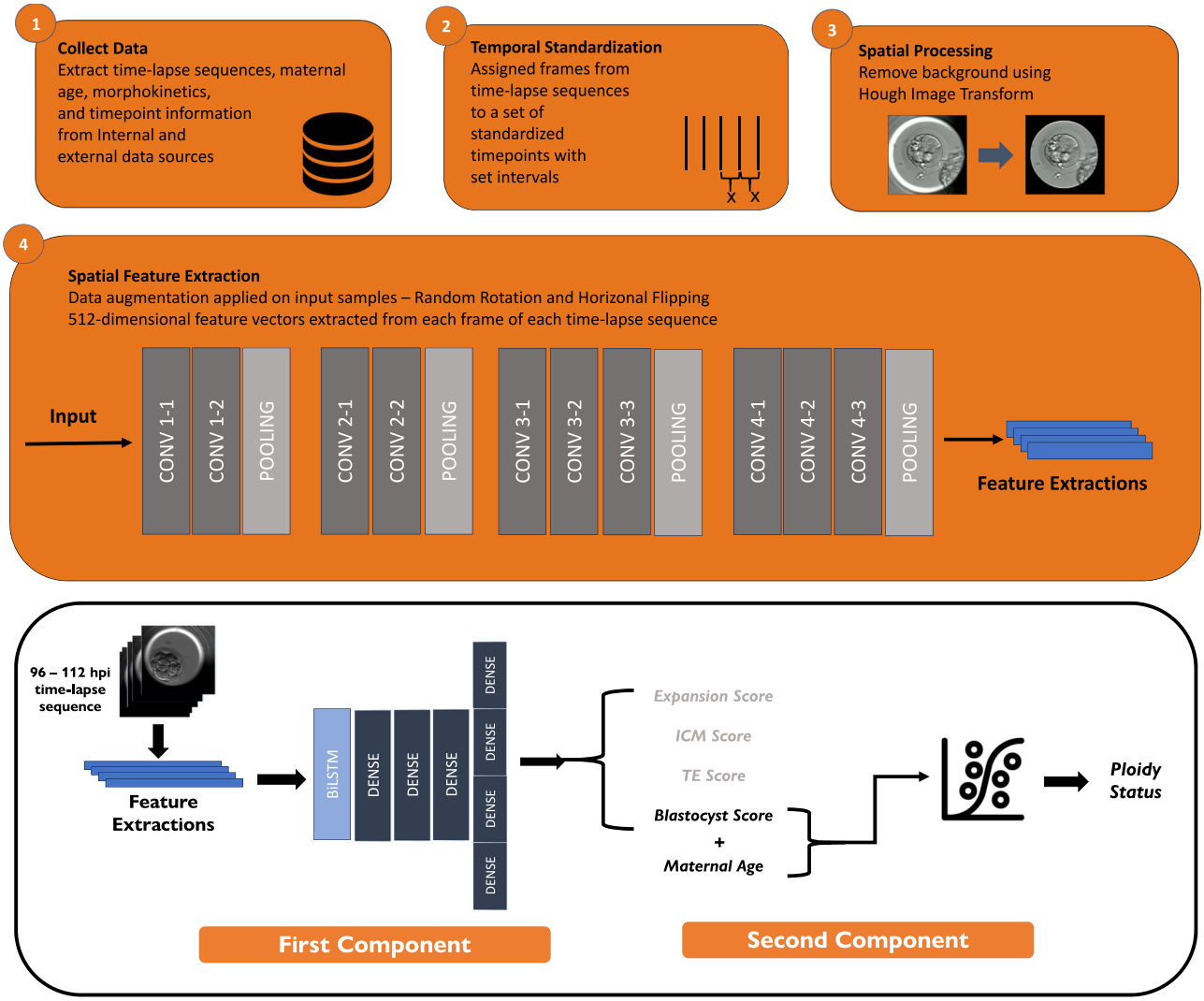

Fig. 1 | Overview of BELA development. Features are extracted from time-lapse image frames as shown in steps 1–4. Time-lapse images are both temporally and spatially processed to decrease bias. Horizontal and rotational augmentation is performed on time-lapse sequences. 512-dimensional features are extracted for each time-lapse image using a pre-trained VGG16 architecture. These features are fed into a multitask BiLSTM model which is trained to predict blastocyst score as well as other embryologist-annotated morphological scores. Predicted blastocyst scores are inputted into a logistic regression model to perform ploidy prediction.

predicted and actual scores of other embryologist metrics (Supplementary Fig. 2). The mean absolute error (MAE) between the MDBS and the ground-truth Embryoscope BS scores is 1.855 ± 0.03. Supplementary Fig. 15a shows the importance across time points that the BiLSTM attributes for predicting blastocyst score. Importances were calculated using SHapley Additive exPlanations (SHAP). The importance follows a bimodal distribution with increased importance at around 96 hpi and 112 hpi, with increased importance at 112 hpi. These importances are consistent with methods embryologists use to determine blastocyst score. First, embryologists at Weill Cornell generally use frames post 110 hpi to assign quality scores which resonate with the heightened importance the BiLSTM model puts at later time points. Second, embryologists look at the speed of blastulation (the time it takes to become a full blastocyst), which involves contrasting earlier and later time points. The bimodal increase in importance suggests that the BiLSTM model similarly contrasts earlier and later time points. The second phase of BELA involves ploidy classification. Using the WCM-Embryoscope test set, BELA, when trained to distinguish between EUP and ANU, attained an AUC of 0.66 ± 0.008, which rose to 0.76 ± 0.002 upon inclusion of maternal age. In the EUP versus CxA task, the AUC of the model was 0.708 ± 0.004 and increased to 0.826 ± 0.004 with the inclusion of maternal age. Comprehensive performance metrics of BELA are found in Supplementary Table 1. BELA's performance (in orange), compared with a day-5 Video model and the embryologist-annotated blastocyst score model, is illustrated in Fig. 2. In all tested scenarios (including or excluding age), test sets, and prediction tasks (EUP versus ANU and EUP versus CxA), BELA outperforms the day-5 video model ($p < 0.05$). Without including maternal age in ploidy prediction, the embryologist-annotated BS model surpasses BELA ($p < 0.05$) in all prediction tasks, barring EUP vs ANU on the WCM-Embryoscope test set (Fig. 2a). However, with maternal age incorporated, BELA outperforms the embryologist-annotated blastocyst score model on the WCM-Embryoscope test set ($p < 0.05$). Still, it underperforms in comparison to the embryologist-annotated blastocyst score model on the WCM-Embryoscope+ dataset. Supplementary Fig. 15b shows the feature importance of maternal age and MDBS for ANU vs EUP prediction. For both covariates, lower values are correlated with euploid predictions, consistent with embryologist decision-making.

The performance of the BELA was compared with a day-5 video model using an external dataset from Spain, consisting of 543 embryos (Fig. 3, Supplementary Table 1). As the Spanish dataset includes only

https://doi.org/10.1038/s41467-024-51823-7

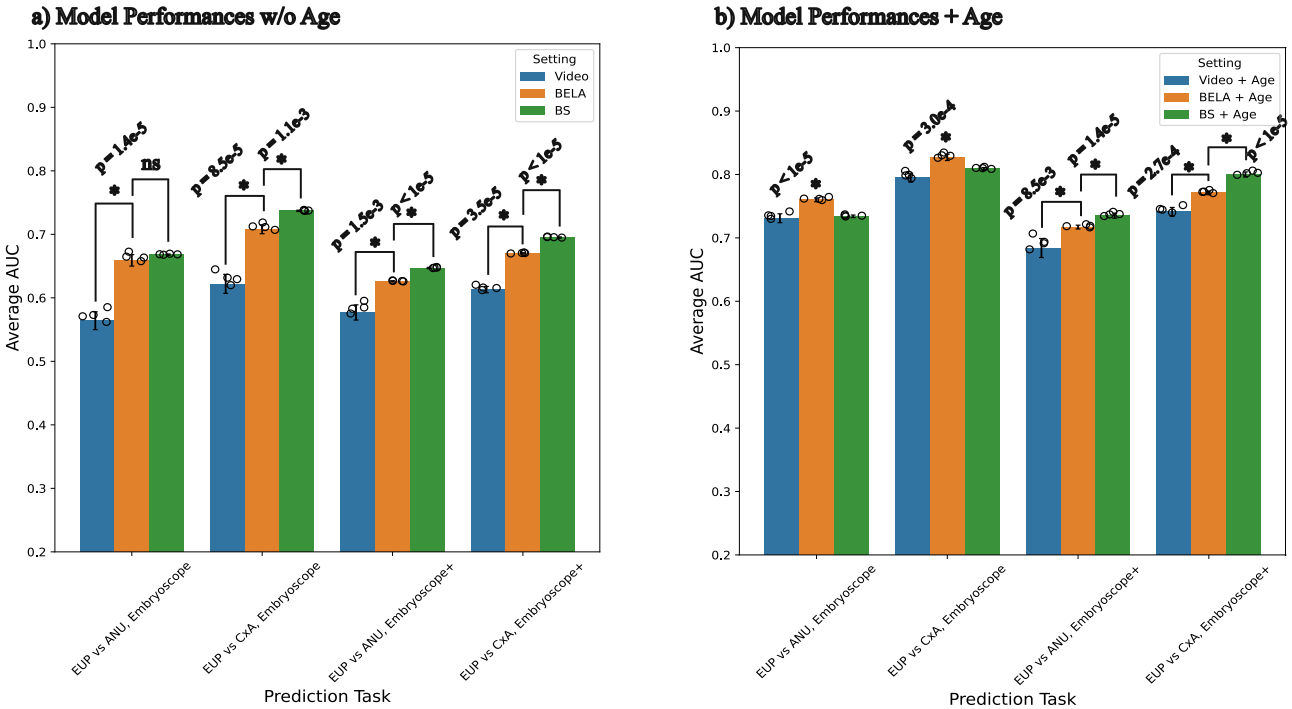

Fig. 2 | Comparison of BELA models with other models. Mean AUC scores and standard deviation for day-5 video, BELA, and embryologist-annotated BS-trained models are shown across 4 replicates (four-fold cross-validation) ($n = 4$). Performances are shown on both the WCM-Embryoscope and WCM-Embryoscope+ datasets for both EUP vs ANU and EUP vs CxA prediction tasks. **a** Performances of models without maternal age. **b** Performances of models with maternal age. Statistical significance was performed using a two-sided *t*-test, where we compared the performance of two different settings at a time. Significance (*) is shown if *p*-value < 0.05. Source data are provided as a Source Data file.

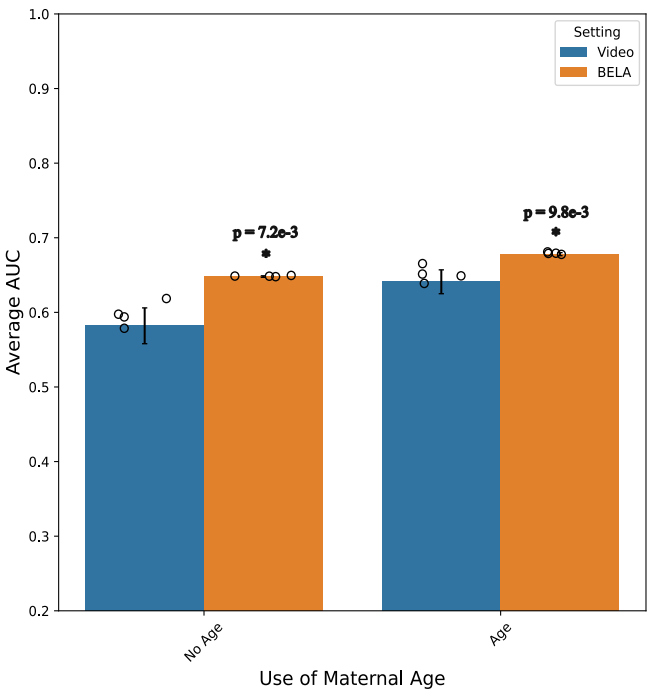

Fig. 3 | Performances of day-5 video model and BELA on the Spain dataset. Average AUC with standard errors is shown for all EUP vs ANU prediction tasks on the Spain dataset for both the day-5 video model and BELA across four replicates (four-fold cross-validation) ($n = 4$). Blue bars depict model performances of day-5 video models, whereas orange bars depict performances of BELA models. Statistical significance was performed using a two-sided *t*-test. Significance (*) is shown if *p*-value < 0.05. Source data are provided as a Source Data file.

embryos labeled as ANU or EUP, model performance could only be measured for the task of distinguishing between EUP and ANU. Notably, BELA significantly outperforms the day-5 video model in both scenarios—with and without the inclusion of maternal age ($p < 0.05$). Unlike the embryos from Weill Cornell Medicine (WCM-Embryoscope and WCM-Embryoscope+ datasets), those from the Spanish dataset were artificially hatched on day 3, which likely impacted later blastocyst morphology and morphokinetics. These embryos exhibit bleached zona pellucida and lack the full expansion seen in the embryos from the training set. To quantitatively verify these differences, feature encodings were extracted using the pre-trained feature extractor for each frame (between 96 hpi and 112 hpi) of each embryo. Averaging these feature encodings across frames yielded a single feature encoding for each embryo, which was further dimensionally reduced via PCA. The resulting feature encodings, categorized by dataset, can be viewed in Supplementary Fig. 3. The feature space shows a significant overlap between the datasets based in the United States, while the Spanish data clusters distinctly toward the bottom right. However, despite these noticeable differences, the performance of the model (excluding maternal age) remains comparable to that achieved with the Weill Cornell Medicine datasets. This suggests that the model might be generally applicable across various clinics, even those with practices that the training data did not account for. The models incorporating maternal age showed decreased performance relative to the Weill Cornell datasets, likely attributable to demographic differences among patients using IVF between Weill Cornell and Spain. For example, in Spain, IVF is more affordable and accessible due to different healthcare insurance policies, whereas in the United States, the high cost of IVF can limit its accessibility to individuals with the necessary financial resources[15,16]. This likely contributes to the different maternal age distributions observed within the datasets.

https://doi.org/10.1038/s41467-024-51823-7

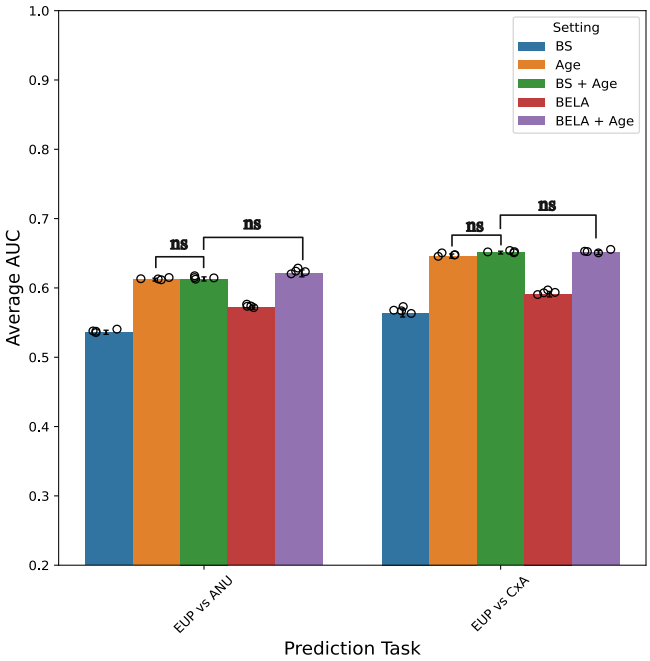

Fig. 4 | Performances of BELA and traditional machine learning models on the Florida dataset. Average AUC with standard errors is shown for EUP versus ANU and EUP versus CxA prediction tasks on the Florida dataset for BELA and logistic regression models trained only on embryologist-derived BS and/or maternal age across four replicates (four-fold cross-validation) ($n = 4$). Statistical significance was performed using a two-sided t-test, and relevant lack of statistical significance are shown via 'ns'. Source data are provided as a Source Data file.

The performance of BELA was further assessed using an external dataset from IVF Florida, comprising 869 embryos. It is important to note that the performance of BELA, as well as comparison logistic regression models trained only on maternal age and embryologist-derived blastocyst score, significantly declined in comparison to the other test sets. This decrease in performance could be attributed to the weak correlations between blastocyst score, maternal age, and ploidy within the Florida dataset (Supplementary Table 8). Maternal age is a crucial predictor of ploidy in all our models, thus, any decrease in its correlation to ploidy can significantly impact performance. Moreover, the embryologist-derived blastocyst scores in the Florida dataset were predominantly centered around a score of 7, thereby reducing the granularity that made it a potent predictor of ploidy in other datasets. This lack of granularity within the Florida dataset might be a result of different scoring practices, as IVF Florida evaluates blastocysts at 115 and 144 hpi, in contrast with the methods employed at Weill Cornell, which also utilize earlier time points for determining blastocyst score. Interestingly, the model-derived blastocyst score (MDBS) from the first module of BELA shows a stronger correlation (−0.119) with ploidy status than the embryologist-derived blastocyst score (−0.101). This finding suggests that BELA can create a score mapping that aligns better with ploidy status compared to the original embryologist-derived blastocyst scores. In order to further validate this hypothesis, an embryologist at Weill Cornell re-graded the 50 embryos within the Florida dataset where the MDBS deviated most significantly from the provided Florida blastocyst scores. The scoring method at Weill Cornell allows for greater granularity in assessing embryo quality. We observed a decrease in mean absolute error (MAE) between the MDBS versus the re-graded blastocyst score from Weill Cornell (4.16) and the MDBS versus the original Florida blastocyst score (5.02). This suggests a higher agreement between the MDBS and the Weill Cornell scoring method. This improved mapping could explain why BELA, with maternal age included, significantly outperforms the model trained on maternal age and embryologist-derived blastocyst score for the EUP vs ANU task ($p < 0.05$) (Fig. 4).

In order to make the model available for clinical use, a web-based application named STORK-V for BELA was developed (Fig. 5, Supplementary Fig. 4). This platform is designed to be user-friendly and capable of predicting an embryo's ploidy status. The required input for the prediction includes time-lapse images captured between 96 and 112 hpi, and the maternal age. Two separate logistic regression models (the second component of BELA) are incorporated to make predictions, one trained to discriminate between euploid (EUP) and aneuploid (ANU) embryos and another trained to distinguish between euploid and complex aneuploid (CxA) embryos. The output from these models includes probabilities for euploidy, aneuploidy, and complex aneuploidy. We also present the intermediary quality scores from the first component of BELA that can be leveraged for further analysis of the embryo. The STORK-V platform serves as a valuable tool for embryologists and in vitro fertilization (IVF) clinics. It offers a convenient and efficient way to assess an embryo's ploidy status, which is a crucial factor in the successful outcomes of assisted reproductive treatments. This will help medical professionals make more informed decisions regarding embryo selection and ultimately improve IVF success rates.

Discussion

In this study, we introduced BELA, which surpasses the traditional IVF embryo classification methods that usually rely on training data from later stages of embryo development and focus only on either image or video data. Compared to previous ploidy prediction and quality estimation models like STORK and STORK-A, BELA utilizes a video sequence rather than a single static image, allowing it to capture both temporal and spatial information. We tested BELA on new additional datasets from both Weill Cornell and external clinics. BELA provides performance gains in both ploidy prediction and quality scoring across multiple additional datasets in Weill Cornell, Spain, and Florida. BELA stands out as a fully automated model that predicts blastocyst scores and utilizes these predictions as a proxy for ploidy classification. BELA's performance is competitive with a model trained on embryologist-annotated blastocyst scores and it significantly surpasses models trained exclusively on time-lapse imaging sequences without a proxy score. Remarkably, BELA only needs time-lapse images from 96 to 112 hpi and maternal age to predict an embryo's ploidy status, thereby making it effortlessly adaptable to clinical workflows without causing any disruption. Notably, BELA also offers a degree of explainability; embryologists can use the model-derived blastocyst score (MDBS) and other scores predicted via multitasking to comprehend the rationale behind a specific ploidy status classification. In terms of recall, BELA demonstrates a substantial potential for successfully selecting euploid embryos, especially for the WCM-Embryoscope+ dataset (Supplementary Table 1). While the model's performance decreases in test datasets outside Weill Cornell, BELA still outperforms models trained on maternal age and/or embryologist-derived blastocyst score. BELA also interestingly found that single aneuploid embryos were evenly predicted as either euploid (EUP) or complex aneuploid (CxA) by our EUP versus CxA BELA model, suggesting that single aneuploid embryos often resemble euploid or complex aneuploid embryos, thus making their identification more challenging. These results are further confirmed by BELA models specifically trained to discriminate between euploid and single aneuploid embryos (Supplementary Note 2). Supplementary Table 2 shows BELA's AUC performance across various age groups classified by the Society for Assisted Reproductive Technology (SART). Despite maternal age being a strong predictor, performances across SART age groups tend to be bimodal (performing best at lower and higher age groups) for the WCM-Embryoscope and WCM-Embryoscope+ datasets. Moreover, in the Spain and Florida datasets, performances across

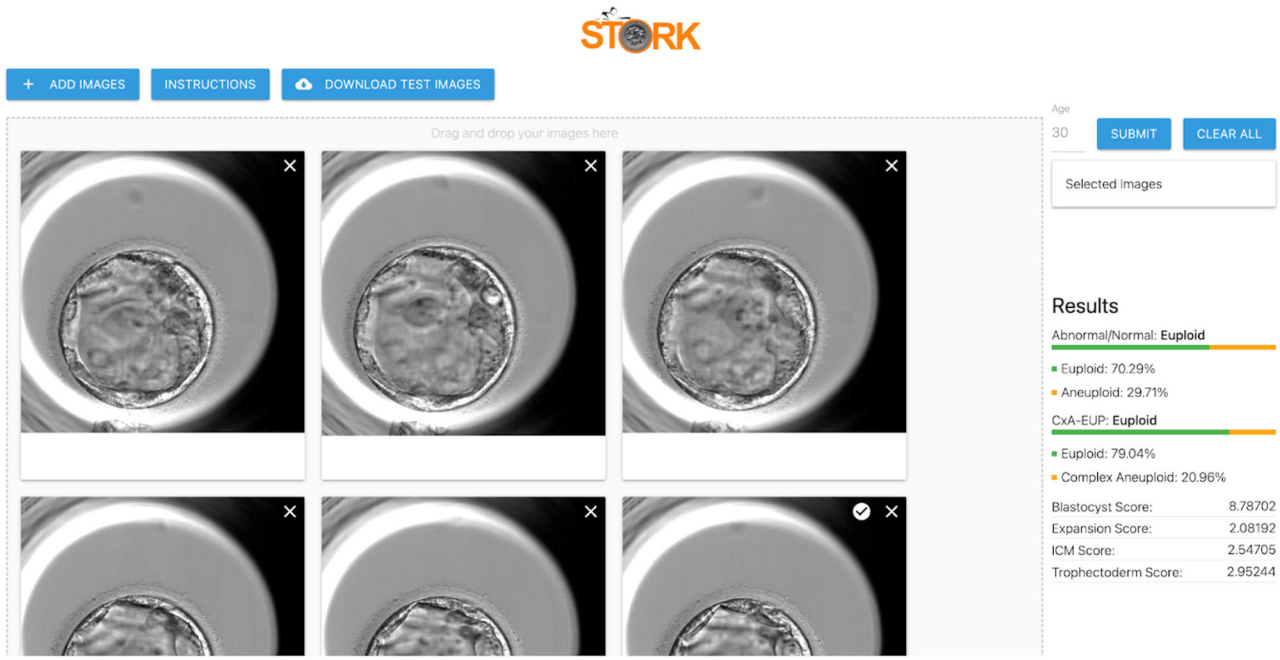

Fig. 5 | STORK-V web interface. A clinical tool that utilizes automation to assist embryologists in determining both the embryo quality score and ploidy status, providing a comprehensive assessment of the embryo.

maternal age do not follow the same distributions that were present in WCM datasets suggesting that these clinics' varying demographics may affect model performance. In conclusion, while BELA is not intended to replace PGT-A, it can provide valuable supplementary information to support decision-making by embryologists, potentially leading to improved success rates in IVF procedures.

The study has several limitations. First, video classification models, such as the one used in this study, demand substantial amounts of training data. Only ~2000 time-lapse sequences of embryo development were available for training, which restricted the ability to experiment with more computationally intensive video classification architectures like the 3D ConvNets or two-stream inflated convolutional nets. Second, despite trying multiple architectures for the feature extractor model, none performed as effectively as the ImageNet pre-trained VGG16 architecture. There could potentially be more suitable feature extractors we did not consider, which might yield information from earlier stages of embryo development. Third, we did not have access to several relevant maternal features, such as hormone levels at the time of oogenesis, demographics, and other clinically pertinent data. These could enhance the prediction of embryo ploidy status. Another limitation was the use of blastocyst scores as intermediary labels in BELA. Despite being well-documented, the blastocyst score is a manually curated label and can be subject to intra-observational bias. Nonetheless, we demonstrated that blastocyst score remains predictive of ploidy, justifying its use as an intermediary proxy value. The results might also be influenced by differing inclusion-exclusion criteria between datasets, possibly explaining some of the differences in model performance among the test datasets. After conducting an analysis (Supplementary Note 3), we have developed the BELA model to not consider mosaic embryos and as such, mosaic embryos with high implantation potential could be misclassified. Our datasets' size also limited us from exploring classifications of different types of aneuploidies beyond single and complex aneuploidy, but investigations show that BELA has the potential to already differentiate between viable and non-viable single aneuploidies (Supplementary Note 5). As it is, BELA remains a promising clinical support tool in its ability to discriminate between euploid and non-euploid embryos. Regarding the ploidy status labels, the use of different platforms for PGT-A across clinics might impact the model's accuracy and generalizability. There is significant variability in PGT-A results between labs and platforms, with no industry-wide standardization currently in place[17]. Factors like methods used for biopsy preparation and the interpretation of results by clinicians could influence PGT-A results, possibly leading to differing detection rates of single versus complex aneuploidy[18]. However, for the advancement of assistive reproductive technologies in IVF, the benchmark should be hastening the time to pregnancy and enhancing live birth outcomes. Embryo selection remains pivotal to this goal, necessitating the prioritization of embryos with high implantation potential and the deprioritization of those with low potential. While most current embryo selection methodologies, such as morphological assessments, lack standardization and are largely subjective, PGT-A offers a consistent approach. This consistency is imperative for developing universally applicable embryo selection methods. Consequently, we used PGT-A results as our model's ground-truth labels. BELA aims to deliver a standardized, non-invasive, cost-effective, and efficient embryo selection and prioritization process. Lastly, the study's model relies predominantly on data from time-lapse microscopy. Consequently, clinics lacking access to this technology will be unable to utilize the developed models.

Contrary to many prior studies that used non-viable embryos as negatives, leading to higher AUCs, the models developed in this study only consist of good biopsied embryos, making them more clinically applicable. The practical implications of these findings could significantly impact the efficiency and effectiveness of the embryo selection process. While the models developed, including BELA, do not replace PGT-A, they can help embryologists reduce the time and effort required to assess embryos. This streamlining of the workflow could allow faster decision-making, letting clinicians concentrate more on patient care and management. Using BELA could also decrease costs for patients and minimize risks associated with the biopsy process. Predictions from BELA can be used to begin ranking embryos from one patient for further downstream analysis. This is especially crucial for patients with a limited number of embryos, as it helps maximize the odds of success while minimizing potential risks and financial costs. Automated blastocyst score prediction (MDBS) is also clinically

relevant to embryologists currently manually annotating embryo scores. In situations where BELA is not used end to end to predict embryo ploidy, it could supplement manual embryo quality scoring. Additionally, BELA is a proof of concept in standardizing blastocyst scoring across clinics by providing an objective score free from embryologist subjectivity. Future iterations of models like BELA, which require no manually curated features and are fully automated from end to end, could be adopted into clinical practice.

Methods

Characteristics of datasets

The study was performed in accordance with relevant guidelines and regulations. The study was approved by the Institutional Review Board at Weill Cornell Medicine (numbers 1401014735 and 19–06020306) and by the IVI Valencia Institutional Review Board (number 1709-VLC-094-MM). IRB determined that this research meets the exemption requirements at HHS 45 CFR 46.104(d) and is secondary research for which consent is not required. A waiver of informed consent was granted from the IRB as the images were de-identified for this retrospective review of clinical data. The embryo imaging was performed as a part of the standard care procedure during the preimplantation and IVF cycle. No discarded embryos were used. In this study, information, which may include information about biospecimens, is recorded by the investigator in such a manner that the identity of the human subjects cannot readily be ascertained directly or through identifiers linked to the subjects. Moreover, the investigators do not contact the subjects, and the investigator will not re-identify the subjects. As such, informed consent was not obtained and participants did not receive compensation for the study. The research utilized multiple datasets for training and validation of the machine learning models. The first dataset, known as the WCM-Embryoscope data, was collected from the Center for Reproductive Medicine at Weill Cornell Medicine between 2018 and 2019. It comprises time-lapse images and PGT-A results for 1998 embryos, including 494 single aneuploid (SA), 588 complex aneuploid (CxA), and 916 euploid (EUP) embryos. A total of 498 patients were included in the WCM-Embryoscope data, with an average of four biopsied embryos each. We treated each sample independently, irrespective of parental origin. Accompanying the time-lapse sequences were clinical data such as embryologist-derived blastocyst score (BS), morphokinetic parameters, and maternal age at the time of oocyte retrieval. The blastocyst score is the sum of a set of scores converted from the expansion, inner cell mass (ICM), trophectoderm (TE) grades, and day of blastocyst formation[14]. The blastocyst score ranges from 3 to 14, with a lower number indicating a higher-quality embryo. The images were captured using the Embryoscope® imaging instrument. To validate the models' generalizability, we used a second dataset, referred to as the WCM-Embryoscope+ data, which was also collected from the Center for Reproductive Medicine. However, these were gathered between 2019 and 2020 and included a total of 841 embryos (170 SA, 261 CxA, and 410 EUP), using a newer Embryoscope+® instrument. Similar to the first dataset, this also contained BS, morphokinetic parameters, and maternal age for each embryo. Furthermore, two external datasets were employed for further validation. The first, referred to as the Spain dataset, came from IVI Valencia and contained 543 embryos (309 ANU and 234 EUP) with time-lapse sequences, morphokinetic parameters, and maternal age. These images were also captured using the Embryoscope instrument. The second external dataset, referred to as the Florida dataset, was collected from IVF Florida and included 869 embryos (202 SA, 222 CxA, and 445 EUP) with maternal age and blastocyst score for each embryo. These images were captured using the Embryoscope+® instrument.

Preimplantation genetic testing

Embryos from Weill Cornell were biopsied on day 5 or day 6, depending on when they reached the blastocyst stage. Biopsied cells were analyzed using next-generation sequencing (NGS) technology at the Ronald O. Perelman and Claudia Cohen Center for Reproductive Medicine (CRM). CRM uses VeriSeq technology from Illumina. The VeriSeq kit utilizes targeted DNA sequencing to detect chromosomal anomalies in embryo biopsies. Samples prepared with the VeriSeq PGS kit are sequences with the standard Illumina MiSeq system. Details about the VeriSeq kit and MiSeq system can be found on the Illumia platform[19,20]. Analyses for the Spain Dataset were done by Igenomix Spain. Embryos were subjected to assisted hatching on day 3, after cell counting, with the Hamilton-Thorne LykosVR laser. After reaching the blastocyst stage, 5–6 trophectodermal cells were biopsied and their ploidy was assessed by Thermo Fisher Scientific's NGS technology. Embryos from IVF Florida were also analyzed by Igenomix using Thermo Fisher Scientific's NGS technology. More details about PGT-A protocols can be found in García-Pascual et al.[21].

Temporal and spatial processing

Extracted time-lapse image sequences were highly variable in length, frame rate, start and end points. These variabilities resulted in numerous embryos missing information from particular time periods, and a lack of proper annotation could lead to bias in model training. To mitigate these biases, the following protocol was developed to clean and standardize all time-lapse sequences, as shown below.

1. Standardized time points are designated at 30-min intervals from 0 to 150 hpi (i.e., 0 hpi, 0.5 hpi, … 149.5 hpi, 150.0 hpi).
2. For each embryo, time-lapse images taken closest to standardized time points are assigned to each standpoint. If there is no image close enough (within 2 h) to the standardized time point, a blank frame is assigned to the standardized time point. We chose a 2-h boundary as the 'close enough' range for several reasons. First, our observations indicated that significant changes in the embryos typically occurred at intervals greater than 2 h. As a result, a 2-h window provided a balance between accurately capturing significant changes while also allowing for reasonable data standardization. This timeframe was also influenced by the overall rate of data acquisition, which sometimes varied but was generally frequent enough to capture changes within this 2-h window. However, we recognize the potential for variability, and further studies may explore the impact of different time boundaries. We also note that the rest of our analysis can be replicated with a different time window and, hence, can be modified on a case-by-case basis. At this point, each standardized time-lapse sequence has 301 frames, with each frame corresponding to a standardized test point between 0 and 150 hpi.
3. After the construction of standardized time-lapse sequences, frames can be extracted for video classification model development using three parameters: start hour, end hour, and interval. For example, a model trained on day 2 embryo development would use these parameters: start hour = 24.0 hpi, end hour = 48.0 hpi, and interval = 2 h. This results in 13 frames.
4. For image classification tasks, a time point of focus can be ascertained, and the frame assigned to that time point can be extracted.

We standardized the lengths, start, and end points of all time-lapse videos using set time points and intervals. Adjacent frames were utilized to impute missing time points. Some sequences, rendered unusable for certain prediction tasks post-standardization, were excluded from the analysis based on exclusion criteria. These criteria encompass instances where the embryo was absent from the petri dish, the embryo was less than half-visible, or the image was too dim to discern the embryo. We resized each frame from 800 × 800 to 224 × 224. To curtail background bias during model training, we implemented a circle Hough Transform for embryo segmentation in each video frame. This processing was uniformly applied across

WCM-Embryoscope, WCM-Embryoscope+, Spain, and Florida datasets. To bolster the diversity and robustness of our training data, we incorporated video augmentation techniques, including random horizontal flipping and rotations. The former yielded mirror images of original frames, effectively doubling our data and fostering diverse pattern learning. Random rotations enhanced the model's adaptability to varied embryo orientations, thereby simulating real-world scenarios. We opted for these techniques as they accurately represent potential real-world variations, fortifying our model's robustness.

General study architecture

Two different prediction tasks were modeled between euploid (EUP), aneuploid (ANU), and complex aneuploid (CxA): EUP versus ANU and EUP versus CxA. Spatial features for each frame were extracted from the cleaned time-lapse images of the embryos using an ImageNet pre-trained VGG16 convolutional neural network (CNN). Time-lapse image frames from 96 hpi to 112 hpi (day 5) were processed according to the "Temporal and spatial processing" section. The features extracted from these frames were input to a multitask BiLSTM regression model (video regression task), which was primarily trained to predict embryologist-derived blastocyst scores. We investigated various dataset combinations for training the BELA models (Supplementary Note 4), ultimately using only WCM-Embryoscope data for the final models. To prevent data leakage, the WCM-Embryoscope dataset was split 70/30 for training/testing. This process exclusively utilized embryos that passed our exclusion criteria, reducing the dataset from 1998 to 1684 embryos. The BiLSTM regression model was trained only using the training slice of the dataset. Four-fold cross-validation was employed when training the BiLSTM regression models, setting aside data for monitoring validation loss. The predicted blastocyst scores for the training split embryos from the BiLSTM regression model, along with maternal age, were used to train a logistic regression model to predict embryo ploidy. A logistic regression model was trained on each of the cross-validated BiLSTM regression models, and the performance metrics of each logistic regression model were averaged. Model performance was measured using accuracy, area-under-receiver-operator-curve (AUC), precision, and recall.

Feature extraction

To extract spatial features from each frame of time-lapse images, an ImageNet pre-trained model from Tensorflow 2.7 was utilized. After experimenting with various pre-trained feature weights and extractors, we utilized a VGG16 CNN architecture to extract spatial features from images. The VGG16 architecture performs significantly better than ResNet50 and DenseNet201 ($p < 0.05$) (Supplementary Fig. 14). While not significantly better performing than the InceptionV3 architecture, a speed increase was observed with the VGG16 architecture, which further warranted its use. VGG16 architectures have been used successfully as feature extractors for other tasks pertaining to time-lapse images in IVF[22–24]. Furthermore, a survey of developments in medical image deep-learning revealed that VGG16 was among the three predominantly utilized CNN architectures, attributed to its fewer hidden layers and reduced propensity for overfitting on smaller datasets[25]. The final layer of the pre-trained architecture performed average pooling, which resulted in 512-dimensional feature vectors for each frame of each embryo.

BELA prediction models

A BiLSTM network was employed for blastocyst score regression, leveraging its capabilities in sequential data pattern recognition, thus processing temporal information from time-lapse images[26]. BiLSTM architectures have been employed in video classification and regression tasks across healthcare and broader domains[27,28]. Given that time-lapse images represent sequences of frames in which data order is pivotal, the bidirectional attributes of the architecture become essential for discerning events with distinct phases. Merging feature extraction processes, which identify spatial patterns in time-lapse images, with a BiLSTM architecture adept at interpreting temporal context, facilitates optimal utilization of the time-lapse data. Our architecture comprises a bidirectional LSTM layer and three dense layers. The BiLSTM received 512-dimensional feature vectors extracted per frame for each embryo. While attention mechanisms and multiple bidirectional LSTM layers were explored, they failed to enhance performance significantly ($p > 0.05$) across all tasks. We modified the BiLSTM architecture to perform multitasking, wherein, in addition to the blastocyst score, the model was trained to predict the expansion score, ICM score, and TE score. Multitasking has been used in previous studies to increase performance in scenarios where predicting different scenarios together may be advantageous to individual task performance. Similar tasks may have overlap in model weights required to come to accurate predictions, hence providing additional information for performing each task[29,30]. Because expansion, ICM, and TE scores make up the overall blastocyst score, we believe that multitasking can be used to improve blastocyst score prediction. The BiLSTM architecture consists of one bidirectional LSTM layer followed by two multi-unit dense layers. For each prediction task, a 1-unit dense layer is added to the model. Since all tasks of the multitask model are regression-based, we used logcosh as the loss function and Adam as the optimizer. Loss weights for each prediction task within the multitask environment were equal. Maternal age was included as a feature in the BiLSTM regression model to predict blastocyst score. Early-stopping with patience = 5 was used to ensure that the model was not overfitting to the training data by monitoring the validation loss on the cross-fold validation data. The performance of the first component of BELA was evaluated using the mean absolute error (MAE) of the predicted blastocyst score (MDBS). Multitask BELA demonstrated a lower MAE (1.855 ± 0.03) compared with a non-multitask BELA (1.877 ± 0.027) on the WCM-Embryoscope test, supporting the use of multitasking. The second part of BELA, the logistic regression model, was fed the predicted blastocyst score, sometimes in combination with maternal age, and performed a binary classification task. The logistic regression model used cross-entropy loss.

Computational resources and time requirements

Model training and inference were conducted using an Apple M1 Mac with TensorFlow Metal. Logistic regression models demonstrated an average training time of 2.5 ± 1.2 s, whereas BiLSTM models required 30.3 ± 11 min. The BELA model on the STORK-V platform was trained on a high-performance BioHPC computing cluster at Cornell, Ithaca, utilizing an NVIDIA A40 GPU and achieving a training time of 5.23 min. Inference for a single embryo on the STORK-V platform took 30 ± 5 s. The efficient use of consumer-grade hardware highlights the practicality of our models for assisted reproductive technology applications.

Statistics and reproducibility

Where relevant, we used the Student's *t*-test to compare the means between two groups. This statistical test was selected because it is well-suited for comparing the means of two samples when the data is approximately normally distributed and the variances of the two groups are similar, as is the case with our data. In addition, all experiments were adjusted for multiple testing using Bonferroni correction to control for the increased chances of observing a statistically significant result, where appropriate. Sample sizes for datasets were determined based on the maximum usable subset available after all exclusion criteria were applied to embryos. These exclusion criteria included embryos with a mosaic PGT-A status, and embryos with missing information such as blastocyst score, ploidy status, and maternal age. Randomization was introduced into experimentation through four-fold cross-validation in all relevant comparisons. The

investigators were not blinded to allocation during experiments and outcome assessment.

Reporting summary

Further information on research design is available in the Nature Portfolio Reporting Summary linked to this article.

Data availability

The embryo-imaging datasets are available under restricted access owing to reasonable privacy and security concerns. Interested researchers and institutions can fill out this form (https://forms.gle/VvvLP5zu35ZWP6UK8) to request controlled access to the de-identified training imaging and meta-data (which includes maternal age and ploidy status labels). N.Z. (nizanin@med.cornell.edu) and team will respond and review requests within a week of form submission. The training dataset was curated by Weill Cornell Medicine's Center of Reproductive Medicine (CRM) and this data access form will be reviewed by CRM and WCM. We emphasize that our proposed models are not specific to the datasets used in this study. Researchers have the flexibility to train and test our deep-learning model on relevant imaging data from their own sources or use our pre-trained models for research-only purposes. To this end, we have made our deep-learning model, BELA, available through a web-based user interface (https://stork-v.eipm-research.org/). Access to this password-protected site is granted for research purposes only and can be obtained by contacting the corresponding author. Source data are provided with this paper.

Code availability

Codes used to train and evaluate the models can be found at https://github.com/ih-lab/stork-v. We have provided documentation for the code in the repository. If the code is used, please cite it with reference number.

References

1. Ma, R. C. W., Ng, N. Y. H. & Cheung, L. P. Assisted reproduction technology and long-term cardiometabolic health in the offspring. *PLoS Med.* **18**, e1003724 (2021).
2. Niakan, K. et al. Human pre-implantation embryo development. *Development* **139**, 829–841 (2012).
3. Niederberger, Craig et al. Forty years of IVF. *Fertil. Steril.* **110**, 185–324.e5 (2018).
4. Greco, E. et al. Preimplantation genetic testing: where we are today. *Int. J. Mol. Sci.* **21**, 4381 (2020).
5. Zhang, Y. X. et al. The pregnancy outcome of mosaic embryo transfer: a prospective multicenter study and meta-analysis. *Genes* **11**, 973 (2020).
6. Khosravi, P. et al. Deep learning enables robust assessment and selection of human blastocysts after in vitro fertilization. *NPJ Digit. Med.* **2**, 21 (2019).
7. Chavez-Badiola, Alejandro et al. Embryo Ranking Intelligent Classification Algorithm (ERICA): artificial intelligence clinical assistant predicting embryo ploidy and implantation. *Reprod. Biomed. Online* **41**, 585–593 (2020).
8. Barnes, Josue et al. A non-invasive artificial intelligence approach for the prediction of human blastocyst ploidy: a retrospective model development and validation study. *Lancet Digit. Health* **5**, e28–e40 (2023).
9. Silver, D. H. et al. Data-driven prediction of embryo implantation probability using IVF time-lapse imaging. *Medical Imaging With Deep Learning*. 1–6 (2020).
10. Lee, C.-I. et al. End-to-end deep learning for recognition of ploidy status using time-lapse videos. *J. Assist. Reprod. Genet.* **38**, 1655–1663 (2021).
11. Gardner, D. K. & Balaban, B. 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.* **22**, 704–718 (2016).
12. Campbell, A. et al. Modelling a risk classification of aneuploidy in human embryos using non-invasive morphokinetics. *Reprod. Biomed. Online* **26**, 477–485 (2013).
13. Rienzi, L. et al. No evidence of association between blastocyst aneuploidy and morphokinetic assessment in a selected population of poor-prognosis patients: a longitudinal cohort study. *Reprod. Biomed. Online* **30**, 57–66 (2015).
14. Zhan, Q. et al. Blastocyst score, a blastocyst quality ranking tool, is a predictor of blastocyst ploidy and implantation potential. *F S Rep.* **1**, 133–141 (2020).
15. Pierce, N. & Mocanu, E. Female age and assisted reproductive technology. *Glob. Reprod. Health* **3**, e9 (2018).
16. Alon, I. & Pinilla, J. Assisted reproduction in Spain, outcome and socioeconomic determinants of access. *Int. J. Equity Health* **20**, 156 (2021).
17. Bardos, J. et al. Reproductive genetics laboratory may impact euploid blastocyst and live birth rates: a comparison of 4 national laboratories' PGT-A results from vitrified donor oocytes. *Fertil. Steril.* **119**, 29–35 (2023).
18. Munné, S. et al. Preimplantation genetic testing for aneuploidy versus morphology as selection criteria for single frozen-thawed embryo transfer in good-prognosis patients: a multicenter randomized clinical trial. *Fertil. Steril.* **112**, 1071–1079.e7 (2019).
19. VeriSeq PGS Kit. https://www.illumina.com/products/by-type/clinical-research-products/veriseq-pgs.html (2023).
20. MiSeq System. https://www.illumina.com/systems/sequencing-platforms/miseq.html (2023).
21. García-Pascual, C. M. et al. Optimized NGS approach for detection of aneuploidies and mosaicism in PGT-A and imbalances in PGT-SR. *Genes* **11**, 724 (2020).
22. Lee, H. J. et al. Six consecutive time-lapse images over 2 hours on day 3 can predict blastulation better than a single image. *Fertil. Steril.* **118**, e263 (2022).
23. Mohamed, Y. A., Yusof, U. K., Isa, I. S. & Zain, M. M. An automated blastocyst grading system using convolutional neural network and transfer learning. In *Proc. 2023 IEEE 13th International Conference on Control System, Computing and Engineering (ICCSCE)*, 202–207 (2023).
24. Lockhart, L., Saeedi, P., Au, J. & Havelock, J. Multi-label classification for automatic human blastocyst grading with severely imbalanced data. In *Proc. 2019 IEEE 21st International Workshop on Multimedia Signal Processing (MMSP)*, 1–6 (2019).
25. Wang, L. et al. Trends in the application of deep learning networks in medical image analysis: evolution between 2012 and 2020. *Eur. J. Radiol.* **146**, 110069 (2022).
26. Yousaf, K. & Nawaz, T. A deep learning-based approach for inappropriate content detection and classification of YouTube videos. *IEEE Access* **10**, 16283–16298 (2022).
27. Romeo, L., Marani, R., D'Orazio, T. & Cicirelli, G. Video based mobility monitoring of elderly people using deep learning models. *IEEE Access* **11**, 2804–2819 (2023).
28. Jamal, I. H. et al. Systematic literature review: human gait cycle model using image-temporal feature. In *Proc.* 2021 6th IEEE International Conference on Recent Advances and Innovations in Engineering (ICRAIE), 1–6 (2021).
29. Caruana, R. Multitask learning. *Mach. Learn.* **28**, 41–75 (1997).
30. Zisimopoulos, P., Sigaras, A. & ih-lab. ih-lab/stork-v: v1.0.0.patched20230130. Zenodo https://doi.org/10.5281/zenodo.11999737 (2024).

Acknowledgements

This study is supported by an NIGMS Maximizing Investigators' Research Award (MIRA) R35GM138152 to I.H. The content is solely the

responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health. I.H. is also supported by an Irma Hirschl Career Scientist Award for this project. Through allocation TG-ASC190055 to I.H., this work used the Extreme Science and Engineering Discovery Environment (XSEDE), which is supported by National Science Foundation grant number ACI-1548562. S.R. would like to acknowledge the support from the Tri-Institutional Training Program in Computational Biology and Medicine (CBM) funded by the NIH grant 1T32GM083937. S.R. would also like to acknowledge funding from the National Science Foundation Graduate Research Fellowship, application number 1000331235. K.O. is supported by a Medical Scientist Training Program grant from the NIGMS under award number T32GM007739 to the Tri-Institutional MD-PhD Program.

Author contributions

S.R., J.B., N.Z. and I.H. conceived the study. S.R., J.B., M.B., and I.H. conceived the method and designed the algorithmic techniques. S.R. wrote the codes and performed the computational analysis with input from I.H., J.B., M.B., and K.O. Q.Z., J.E.M., Z.R. and N.Z. provided the Weill Cornell datasets and labeled images. N.Z. evaluated additional embryo images. M.M. provided the Spain dataset. K.A.M. and D.H. provided the Florida dataset. S.R. drafted the manuscript with input from J.B., M.B., O.E., Q.Z. and I.H. P.Z. and A.S. designed the user interface. All the authors read the paper and suggested edits. I.H. supervised the project.

Competing interests

O.E. is a scientific adviser for, and an equity holder in, Freenome, Owkin, Volastra Therapeutics, OneThree Biotech, Genetic Intelligence, Acuamark DX, Harmonic Discovery, and Champions Oncology, and has received funding from Eli Lilly, Johnson & Johnson–Janssen, Sanofi, AstraZeneca, and Volastra. N.Z. is a paid consultant for AIVF and Fairtility, and is on the advisory board of, and has equity in, Alife Health. I.H. is a consultant and is on the advisory board of Noor Sciences. S.R., J.B., J.E.M., Z.R., O.E., N.Z. and I.H. are listed as inventors on a provisional patent filed by Cornell University (application number 63/484,177) about the technology described in this study. M.M. received speaker fees from Merck, Vitrolife, Ferring, Theramex, and Gideon Richter. P.Z. holds stocks in Pfizer and Bristol Myers Squibb. K.A.M. serves as a paid consultant and advisory board member for Fairtility and Alife Health (holding equity), and as a scientific board member for Genomic Prediction and Igenomix. The remaining authors declare no competing interests.

Additional information

Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s41467-024-51823-7.

Correspondence and requests for materials should be addressed to Iman Hajirasouliha.

Peer review information *Nature Communications* thanks the anonymous reviewers for their contribution to the peer review of this work. A peer review file is available.