nature communications

Article

https://doi.org/10.1038/s41467-025-56908-5

AI-based approach to dissect the variability of mouse stem cell-derived embryo models

Received: 30 August 2024

Accepted: 5 February 2025

Published online: 19 February 2025

Check for updates

Paolo Caldarelli[1,5], Luca Deininger[2,3,5], Shi Zhao[1], Pallavi Panda[1], Changhuei Yang[1], Ralf Mikut[2] ✉ & Magdalena Zernicka-Goetz[1,4] ✉

Recent advances in stem cell-derived embryo models have transformed developmental biology, offering insights into embryogenesis without the constraints of natural embryos. However, variability in their development challenges research standardization. To address this, we use deep learning to enhance the reproducibility of selecting stem cell-derived embryo models. Through live imaging and AI-based models, we classify 900 mouse post-implantation stem cell-derived embryo-like structures (ETiX-embryos) into normal and abnormal categories. Our best-performing model achieves 88% accuracy at 90 h post-cell seeding and 65% accuracy at the initial cell-seeding stage, forecasting developmental trajectories. Our analysis reveals that normally developed ETiX-embryos have higher cell counts and distinct morphological features such as larger size and more compact shape. Perturbation experiments increasing initial cell numbers further supported this finding by improving normal development outcomes. This study demonstrates deep learning's utility in improving embryo model selection and reveals critical features of ETiX-embryo self-organization, advancing consistency in this evolving field.

In recent years, the field of developmental biology has undergone a transformative shift with the introduction of stem cell-derived embryo models[1]. These innovative models emulate various stages of embryonic development, offering new research avenues that were previously constrained by ethical and practical limitations associated with using actual embryos. Among these, some are designed to mimic pre-implantation development[2,3], while others capture post-implantation stages of mouse (e.g., ETS[4], ETXs[5] and ETiXs[6]) and human embryogenesis[7–9]. The ability of these post-implantation models to progress to advanced developmental stages makes them invaluable tools for translational research, particularly in unraveling developmental disorders and advancing regenerative medicine[10].

Despite their considerable promise, these models face significant challenges, primarily stemming from a limited understanding of the initial stages of cell and tissue self-organization. The initial phases of development in these models are crucial, as they set the foundation for all subsequent growth and differentiation[11]. However, the lack of detailed characterization of these phases results in variability in development outcomes. Typically, within just a few days of development, there are observable differences in growth and morphology among embryo models. This variability complicates efforts to standardize experiments and limits the models' applicability for more high-throughput screens.

The ETiX-embryo model stands as one of the most advanced model systems for simulating post-implantation development in mice[12]. This model is constructed by aggregating embryonic stem cells (ESCs) with trophoblast stem cells (TSCs) derived from extra-embryonic ectoderm (ExE) precursors. Additionally, ESCs that are

[1]Division of Biology and Biological Engineering, California Institute of Technology, Pasadena, CA, USA. [2]Group for Automated Image and Data Analysis, Institute for Automation and Applied Informatics, Karlsruhe Institute of Technology, Eggenstein-Leopoldshafen, Germany. [3]Division of Pediatric Neurology and Metabolic Medicine, Department I, Center for Pediatric and Adolescent Medicine, Medical Faculty Heidelberg, Heidelberg University, Heidelberg, Germany. [4]Mammalian Embryo and Stem Cell Group, Department of Physiology, Development and Neuroscience, University of Cambridge, Cambridge, UK. [5]These authors contributed equally: Paolo Caldarelli, Luca Deininger. ✉e-mail: ralf.mikut@kit.edu; magdaz@caltech.edu

transiently induced to express the visceral endoderm master regulator GATA4, are incorporated to enhance the model's fidelity to natural development processes. The ETiX model has demonstrated potential in reaching early organogenesis stages[6,13,14]. However, by the 4th-day post-seeding, a culling process is necessary to eliminate structures that fail to meet the developmental criteria akin to the natural progression of embryogenesis at a comparable stage. Such selection is common in organoid and stem cell-derived embryo models research, where typically only the most promising samples are carried forward for further study. The selection process is inherently subjective and reliant on the individual researcher's judgment, leading to variability in outcomes across different laboratories.

To address these challenges and improve the reliability of stem cell-derived embryo models, it is essential to adopt innovative analytical approaches. One such promising solution is the integration of Artificial Intelligence (AI) techniques for image analysis, which has been proven a valuable tool in a variety of complex scenarios[15]. Particularly, a branch of AI known as deep learning revolves around training neural networks on large datasets. In the context of image classification, convolutional neural networks (CNNs) and vision transformers (ViTs) play a pivotal role. CNNs leverage hierarchical layers that perform convolution operations, allowing them to extract features from raw data, particularly suited for tasks like image classification[16–18]. Notable CNN architectures include AlexNet[16], MobileNet[17], and ResNet[18], each contributing unique innovations like pioneering deep network structures (AlexNet), mobile device optimization (MobileNet), and residual connections for deep network training (ResNet). In contrast, ViTs divide images into fixed-size patches, transforming each into a sequence of tokens processed by a transformer encoder for feature extraction. One recent ViT, in particular, designed for video classification, is the Multiscale Vision Transformer (MViT)[18], which combines hierarchical features with ViTs.

The proven success of deep learning in image and video classification has driven significant breakthroughs across various fields, including biomedicine, by enabling precise disease diagnosis, drug discovery, and personalized treatment strategies. In organoid research, deep-learning models facilitate the monitoring, tracking, and analysis of organoid morphological features over time[19,20]. In embryo research, models like EmbryoNet[21], a ResNet-based model, identified molecular defects in zebrafish embryos. These successes highlight the potential for deep learning to accurately classify and analyze stem cell-derived embryo models such as ETiX-embryos.

Here, we developed a model optimized for the cultivation and imaging of hundreds of stem cell-derived embryo models at a time and applied deep-learning-based classification across various stages of ETiX-embryo development to provide insights into embryo features predictive of future successful development and discuss how our findings can improve embryo cultivation efficiency.

Results

Detailed dynamics and expert classification of ETiX-embryos via live-imaging

We employed a custom-developed live-imaging platform to observe the development of ETiX-embryos within microwells constructed from agarose, focusing on the initial 90 h (Supplementary Fig. 1). This platform, accommodating ~320 stem cell-derived embryo-like structures per session, utilized confocal microscopy to capture multifocal images of each ETiX-embryo (Fig. 1A, B and Supplementary Movie 1). To enable detailed visualization, the cells were fluorescently labeled: ESCs were tagged with membrane-targeted RFP[6], ESC-iGata4 with membrane-targeted GFP[12], and TSCs with a membrane far-red dye (CellMask), enabling tracking of each cell type within the developing ETiX-embryo.

Employing this methodology, we created a dataset of 900 ETiX-embryos, which were subsequently annotated based on images of the last 25 h of the time-lapse, focusing on specific characteristics namely lineage segregation, the presence of the pro-amniotic cavity—a fluid-filled space forming soon after implantation through lumenogenesis of the epiblast[22]—and an overall cylindrical shape (Fig. 1C). All of these characteristics had to be met for an ETiX-embryo to be classified as normal. Specifically, ETiX-embryos were classified by an expert embryologist as *normal* if they displayed a cylindrical shape with distinct cellular compartments derived from TSCs and ESCs, enveloped by a monolayer of ESC-iGata4 cells that resemble the visceral endoderm. The formation of a well-defined pro-amniotic cavity was also a crucial indicator. These features are classical hallmarks of a properly developed mouse embryo at the early post-implantation stage, just before the onset of anterior visceral endoderm migration to specify the anterior-posterior axis and gastrulation. Of the ETiX-embryos analyzed in three independent experiments, only 23% (206) met the criteria for *normal* development throughout the observation period (Fig. 1D, F and Supplementary Movie 2), while the remaining 77% (694) were classified as *abnormal*, showing structural and developmental abnormalities (Fig. 1D, G and Supplementary Movie 3). Additionally, by carrying out the time-lapse studies, we noticed that most ETiX-embryos were not synchronized in their development, and therefore, we annotated an end time point for each ETiX-embryo at a similar developmental stage, ranging between 65 and 90 h post-cell-seeding (Fig. 1E and Supplementary Fig. 2A, B). We refer to this as the synchronized dataset, which was subsequently used to train our deep-learning model, StembryoNet.

These findings demonstrate the capability of our imaging platform to provide insights into ETiX-embryo development. While confirming the known variability in ETiX-embryogenesis, our time-lapse imaging enables tracking of each embryo's developmental history, allowing us to investigate the underlying causes of this variability.

StembryoNet: an AI model for stem cell-derived embryo classification at advanced developmental stages

For deep-learning-based ETiX-embryo classification at 90 h post-cell-seeding, we introduce StembryoNet, a deep-learning model built on a ResNet18 architecture (Fig. 2A). StembryoNet is specifically designed to be trained on synchronized data while also enabling predictions on unsynchronized data. It retains the five sequential convolutional layers and global average pooling of ResNet18 but replaces the original 1000-neuron fully connected layer with a single-neuron layer for binary ETiX classification. Additionally, we substitute the softmax function with a sigmoidal activation function to calculate class probabilities. A key feature of StembryoNet is its ability to predict outcomes from unsynchronized data. It achieves this by processing consecutive time points from the last 25 h of the same ETiX-embryo, generating individual outputs for each time point. These outputs are then concatenated, and the thresholded maximum probability across the time points is used to determine the final classification (Fig. 2A). To train StembryoNet on the phenotype of *normal* ETiX-embryos, we used the synchronized dataset, which aligns ETiX-embryo-specific time points at similar developmental stages (Supplementary Fig. 2B, D). As synchronization is not feasible for *abnormal* embryos, we sampled their synchronized time points using a normal distribution with the same mean and standard deviation as those of *normal* ETiX-embryos. To highlight the advantage of StembryoNet over state-of-the-art deep learning models, we compared its performance to a single ResNet18 trained on ETiX-embryo images captured at 90 h ($ResNet_{90h}$) and a Multiscale Vision Transformer (MViT) trained on videos of ETiX-embryo development spanning from 65 to 90 h ($MViT_{65–90\,h}$). To obtain an unbiased estimate of the models' performances, we used five-times repeated 5-fold cross-validation (Supplementary Fig. 2C). Detailed information on model training and evaluation is provided in the Methods section.

https://doi.org/10.1038/s41467-025-56908-5

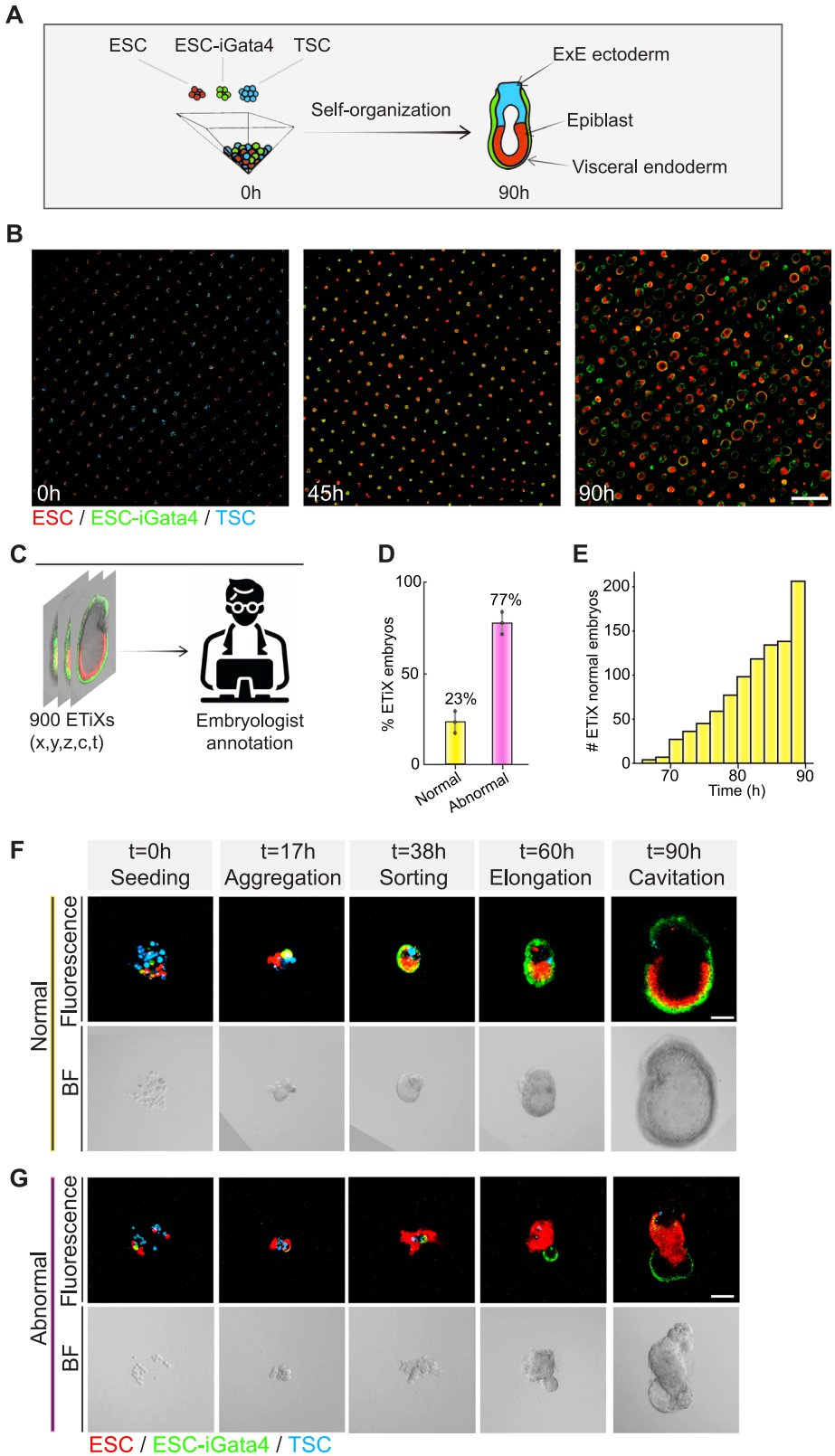

Fig. 1 | Long-term live imaging and expert annotation of mouse post-implantation stem cell-derived embryo models (ETiX). A Schematic illustrating the formation and structural organization of ETiX-embryos from 0 h to 90 h. **B** Representative time points of live imaging movies displaying around 320 ETiX-embryos at 0, 45, and 90 h after seeding in the Agarwell (n = 3, N = 900). **C** Dataset annotation of 900 ETiX-embryos by an expert embryologist based on 4D time-lapse imaging data. **D** Bar chart displaying the mean percentage of ETiX-embryos identified as *normal* or *abnormal*, with error bars representing variability across datasets (n = 3, 95% CI). **E** Cumulative histogram of expert-selected time points of *normal* ETiX-embryos. **F** Sequential fluorescence and bright field images showcasing a representative *normal* ETiX-embryo development at key time points: seeding (0 h), aggregation (17 h), sorting (38 h), elongation (60 h), and cavitation (90 h). **G** Comparative fluorescence and bright field images of a representative ETiX-embryo exhibiting *abnormal* development at corresponding stages to (**F**). ESCs (red), ESC-iGata4 (green), and TSCs (blue). Scale bars: 1000 μm for (**C**) and 100 μm for (**F**, **G**). Source data are provided as a Source Data file.

While a baseline random classifier achieves an accuracy of 50% (F1-Score = 31%), StembryoNet, $ResNet_{90h}$, and $MViT_{65\text{-}90h}$ all exceeded this baseline (Fig. 2B). Notably, StembryoNet achieved a mean accuracy of 88% (F1 = 77%), outperforming $ResNet_{90h}$ (80% accuracy, F1 = 67%) and $MViT_{65\text{-}90h}$ (81% accuracy, F1 = 68%). The confusion matrix for StembryoNet indicates a recall of 83% for *normal* ETiX-embryos at a precision of 71% (Fig. 2C). In comparison, $ResNet_{90h}$ yielded more false positives, resulting in a recall of 87% and a precision of 55% for *normal* ETiX-embryos (Fig. 2C and Supplementary Fig. 3D, E). StembryoNet consistently outperformed not only $ResNet_{90h}$ and $MViT_{65\text{-}90h}$ but also other advanced deep learning models, demonstrating significant superiority across comparisons (Supplementary Fig. 3C). All models are trained on RGB images consisting of three fluorescence channels, as adding the brightfield channel did not enhance performance (Supplementary Fig. 3A).

Testing StembryoNet on synchronized data (pre-selected time points) did not yield better performance, demonstrating that StembryoNet makes a human in the loop expendable (Supplementary Fig. 3B, E, F). To assess whether StembryoNet's training on synchronized data accounts for its superior performance compared to $ResNet_{90h}$, we tested it on ETiX-embryos at 90 h, achieving 79% accuracy ($StembryoNet_{90\,h}$, Supplementary Fig. 3B). This suggests that training on synchronized data alone does not account for the performance difference; instead, the advantage lies in StembryoNet's ability to fuse model predictions on unsynchronized data.

To evaluate StembryoNet's performance in terms of both accuracy and speed, we compared it to three embryologists. The results show that, when using two of the three embryologists as the ground truth, StembryoNet slightly outperforms the least accurate embryologist (Supplementary Fig. 3G, H). For the third embryologist, StembryoNet's performance is slightly lower than that of the least accurate embryologist (Supplementary Fig. 3I). In terms of speed, StembryoNet is 18 times faster than the embryologists (Supplementary Fig. 3J). Overall, these findings suggest that StembryoNet achieves accuracy comparable to that of human experts but with a substantial speed advantage, highlighting its practical utility for high-throughput applications.

To contextualize StembryoNet's predictions biologically, we used Grad-CAM heatmaps to visualize the model focus areas on both *normal* and *abnormal* ETiX-embryos (Fig. 2D, E). For normal ETiX-embryos, the model predominantly focused on the pro-amniotic cavity and the regions where the visceral endoderm-like (ESC-iGata4-derived tissue) encircles the epiblast-like and extraembryonic ectoderm-like (ESCs and TSCs-derived tissues), indicating proper lineage allocation during early post-implantation development (Fig. 2D). Conversely, for *abnormal* ETiX-embryos, the model's attention varied, focusing on disparate image parts such as ESC-tissue, ESC-iGata4-tissue, or a mix thereof, aligning with indicators of *abnormal* development such as lack of lumenogenesis, improper compartment formation, or incorrect lineages positioning (Fig. 2E).

Fully supervised methods like StembryoNet rely on labor-intensive human annotations, which makes self-supervised, annotation-free deep learning approaches an appealing alternative. DINO[22], a recent self-supervised deep learning method, trains vision transformers to learn meaningful visual representations without labeled data. Although DINO performed inferiorly to StembryoNet, it was able to cluster *normal* and *abnormal* ETiX-embryos to some extent, showcasing the potential of annotation-free approaches for the future (Supplementary Fig. 4A–C).

In conclusion, StembryoNet significantly enhances our ability to classify ETiX-embryos at advanced developmental stages, achieving superior accuracy and precision compared to state-of-the-art deep learning models. By closely aligning with embryologist perceptions and providing biological insights, StembryoNet represents a robust tool for the analysis of ETiX-embryo development.

Early identification of successful ETiX-embryos via deep learning

Deep-learning-based classification of *normal* and *abnormal* ETiX-embryos at 90 h post-cell-seeding can enhance the reproducibility of ETiX-embryo selection across laboratories. However, to advance this approach and delve deeper into the self-organization mechanism of *normal* ETiX-embryogenesis, we aimed to (1) predict ETiX-embryo development outcomes at earlier stages, and (2) identify the features most predictive of successful development. Since our goal was to forecast whether an ETiX-embryo would develop normally by 90 h, we used the embryologist's annotations at that time point as target labels (Fig. 3A). To assess model accuracy and pinpoint key classification features over time, we trained models at 5-h intervals from 0 to 90 h, generating 19 distinct model sets in total. StembryoNet was not applied for this task, as it is specifically designed for classification at later developmental stages where ETiX-embryo synchronization—annotating the time point of similar development—is feasible.

We used two types of models, a ResNet18 and a Support Vector Machine (SVM), applied to both brightfield and fluorescence images: $ResNet_{BF}$, $ResNet_{Fluor}$, SVM_{BF}, and SVM_{Fluor} (Fig. 3B). While we trained the ResNet models directly on the images, SVM models were trained on simple, pre-extracted image features of ETiX-embryos. Compared to the deep-learning-based ResNet18, SVM models offer greater explainability as their manually defined features can be analyzed for importance. The performance gap between ResNet18 and SVM provides insights into the advantages of deep-learning features in comparison to simple features. We employed five-times repeated 5-fold cross-validation to ensure unbiased estimates of model performance.

At the cell-seeding stage (t_0), both SVM_{Fluor} and $ResNet_{Fluor}$ significantly outperformed a baseline random classifier (Fig. 3C). $ResNet_{Fluor}$ achieved an accuracy of 65% (F1 = 43%), while SVM_{Fluor} performed comparably by achieving an accuracy of 64% (F1 = 42%). SVM_{Fluor} feature weights indicated that the red (ESCs) and green (ESCs-iGata4) channels were most predictive (Fig. 3D). In contrast, SVM_{BF} performed significantly worse than random and $ResNet_{BF}$ showed no improvement over random (Fig. 3C), highlighting the critical role of fluorescence information—i.e., the fluorescent labeling of ESC, ESC-iGata4, and TSC cells—in early-stage predictions.

Inherently, the channel-wise fluorescence sum (Fig. 3B) is an estimate of the initial cell number and emerged as a distinguishing feature of ETiX-embryos with *normal* and *abnormal* future development. The protocol aims for each ETiX-embryo to contain, on average, five ESCs, five ESCs-iGata4, and sixteen TSCs at the initial cell-seeding stage. However, achieving these exact cell numbers during seeding is challenging due to the stochastic nature of cell distribution in each well intrinsic to the experimental protocol. This variability could explain the differences in cell counts observed between *normal* and *abnormal* ETiX-embryos. Indeed, at the time of seeding (t0), an SVM model trained on the initial number of ESC, ESC-iGata4, and TSC cells ($SVM_{cellcount}$) performed comparably to $ResNet_{Fluor}$ (Supplementary Fig. 5A), indicating that the initial number of cells serves as a predictor for successful development. Further analysis showed that *abnormal* embryos typically had significantly fewer ESC cells (*normal*: 6.7 ± 2.9, *abnormal*: 4.5 ± 2.5, mean ± SD) and ESC-iGata4 cells (*normal*: 6.0 ± 2.7, *abnormal*: 4.1 ± 2.3, mean ± SD) on average (Supplementary Fig. 5B). In line with this finding, additional datasets were generated where the initial cell numbers for all three cell types were doubled ($Data_{2X}$, n = 306) or tripled ($Data_{3X}$, n = 276). This approach was taken to explore whether increasing the overall cell count could mitigate the variability during seeding and ensure that each ETiX-embryo receives the necessary complement of cells for normal development. By uniformly increasing the numbers of all three cell types, we hypothesized that the chances of each embryo receiving sufficient cell numbers would improve, thereby increasing the proportion of embryos that

https://doi.org/10.1038/s41467-025-56908-5

develop normally. The results confirmed this hypothesis, showing a higher proportion of *normal* ETiX-embryos as the initial cell counts increased, with $Data_{1X}$, $Data_{2X}$, and $Data_{3X}$ exhibiting *normal* ETiX-embryos proportions of 23%, 32%, and 60%, respectively (Supplementary Fig. 5C and Supplementary Movie 4). This suggests that increasing the initial number of cells across all lineages may enhance developmental outcomes by ensuring that each ETiX-embryo receives the minimum necessary amount of each cell type. Additionally, deep-learning-based brightfield embryo segmentation showed accelerated ETiX-embryo growth for $Data_{2X}$ and $Data_{3X}$, allowing them to reach similar developmental stages earlier than those in $Data_{1X}$ (Supplementary Fig. 5E). This was further supported by a negative correlation between the synchronized time points and the number of ESC cells ($r = -0.33$, Supplementary Fig. 5F) and ESC-iGata4 cells

Fig. 2 | AI-based ETiX-embryo classification at advanced stages of development. **A** Consecutive images from the final 25 h (65–90 h post-seeding) of each embryo development are input into the StembryoNet. StembryoNet comprises five convolutional layers, followed by a global average pooling (GAP) layer, fully connected (FC) layer, and sigmoidal activation function (S). The model predicts the probability of the *normal* class at each time point. These probabilities are concatenated, and the maximum probability is thresholded by a parameter θ to determine the class. **B** Performance comparison of different classifiers that are trained on single-time points ($ResNet_{90h}$) and multiple-time points: $MViT_{65–90\,h}$ and StembryoNet, over five times repeated 5-fold cross validation (CV, n = 25). $MViT_{65–90\,h}$ vs. $ResNet_{90h}$: $t(44) = 0.7$, $p = 0.48$, $d = 0.2$, 95% CI [−0.01, 0.02], StembryoNet vs. $ResNet_{90\,h}$: $t(41) = 12.5$, $p = 1.2 \times 10^{-15}$, $d = 3.5$, 95% CI [0.07, 0.09], StembryoNet vs. $MViT_{65–90\,h}$: $t(34) = 9.0$, $p = 1.4 \times 10^{-10}$, $d = 2.6$, 95% CI [0.06, 0.09], StembryoNet vs. Random: $t(81) = 76.0$, $p = 3.6 \times 10^{-77}$, $d = 11.4$, 95% CI [0.37, 0.39]. **C** Confusion matrices of random classifier, $ResNet_{90\,h}$, and StembryoNet averaged across five times repeated 5-fold CV. Grad-CAM attention maps highlight significant areas contributing to the classification decision of StembryoNet for *normal* (**D**) and *abnormal* (**E**) embryos. Box plots: center line, median; box limits, first and third quartile; whiskers, smallest/largest value no further than 1.5*IQR from the corresponding hinge. ****P < 0.0001, ns not significant, two-sided Welch's t-test. Scale bars: 100 μm. Source data are provided as Source Data file.

(r = −0.27, Supplementary Fig. 5G) in $Data_{1X}$, underscoring the need for synchronization.

As development progressed, classification accuracy improved compared to the cell-seeding stage (Fig. 3E). $ResNet_{Fluor}$ consistently outperformed SVM_{BF}, SVM_{Fluor}, and $ResNet_{BF}$, particularly during later stages (i.e., 60 to 90 h, Fig. 3E). Over time, fluorescence sum values decreased in importance while ETiX-embryo shape features gained relevance (Fig. 3F). In particular, ETiX-embryos with *normal* development followed a clear morphological trajectory (Supplementary Fig. 6B–E), displaying a higher perimeter and a lower PerimeterSurfaceRatio, indicative of a more compact shape, as opposed to the more fragmented appearance of *abnormal* ETiX-embryos likely due to tissue mispositioning (Supplementary Fig. 6A and Supplementary Movie 5). Based on the Slingshot method[23], we calculated the trajectory by connecting clusters of ETiX-embryos from distinct time points from 0 to 90 h in 5-h intervals and ordering them based on morphological similarity, using nine ETiX-embryo shape features extracted from brightfield ETiX-embryo segmentations, beginning from the initial time point of 0 h. The analysis also captured radial symmetry breaking around 65 h, where *normal* ETiX-embryos, initially spherical with radial symmetry, began to elongate, while *abnormal* ones maintained similar sphericity (Supplementary Fig. 6A and Supplementary Movie 5).

In summary, deep-learning approaches can predict ETiX-embryo success from the cell-seeding stage, with fluorescence data and initial cell numbers as key predictors. Increasing initial cell counts improves the proportion of *normal* embryos. As development progresses, prediction accuracy increases, and morphological features gain predictive importance.

Analyzing individual ETiX-embryo development identifies distinct developmental progressions

We next aimed to analyze individual ETiX-embryos to assess their developmental progression and evaluate their likelihood of becoming *normal* ETiX-embryos over the course of development. We identified four distinct patterns of ETiXs development: continuously *normal*, ab*normal* that transition to *normal*, continuously *abnormal*, and initially *normal*-looking embryos that manifested abnormalities as development progressed (Fig. 4A–D and Supplementary Movie 6). For instance, one ETiX-embryo initially showed abnormal development and apparent lineage allocation failure but eventually readjusted after 35 h post-cell-seeding, exhibiting a 20-h delay compared to a *normal* embryo (Fig. 4B, E and Supplementary Movie 6). Another ETiX-embryo, which transitioned from *normal* to *abnormal* development, initially resembled a continuously *normal* ETiX-embryo before deviating at around 35 h post-cell-seeding, likely due to a failure in tissue sorting (Fig. 4D, E and Supplementary Movie 6).

This classification revealed a spectrum of developmental outcomes, ranked from most to least frequent: continuously *abnormal*, initially *normal* then becoming *abnormal*, continuously *normal*, and initially *abnormal* then turning *normal* (Fig. 4F). Applying this categorization to $Data_{2X}$ and $Data_{3X}$ revealed higher proportions of normal EtiX-embryos, further supporting our previous findings (Supplementary Fig. 5D). This categorization deepens our understanding of ETiX-embryogenesis dynamics and underscores the importance of tracking and analyzing individual ETiX-embryo trajectories over time, rather than assuming successful development is established early on and maintained throughout the observation period. Indeed, our findings confirm the non-deterministic and highly self-organizing nature of the system, demonstrating that ETiX-embryos can readjust in response to fluctuations.

Discussion

Our study marks an advancement in the use of deep learning to classify and dissect the experimental variability of stem cell-derived embryo models, representing the first study of this kind in the field. We introduced StembryoNet, a deep-learning model specifically designed to classify ETiX-embryos at advanced developmental stages (90 h post-seeding), achieving a remarkable accuracy of 88% (F1-Score = 77%). This performance significantly surpasses that of other state-of-the-art models, such as ResNet18 and MViT ($p < 0.0001$), confirming the robustness and reliability of StembryoNet in distinguishing between *normal* and *abnormal* embryonic forms. We found that the simpler 2D ResNet performs *on par* with the more complex video classification model MViT. We speculate that MViT might require larger datasets to effectively learn which specific time points are relevant for classifying an ETiX-embryo as *normal* or *abnormal*. Given the moderate size of our dataset, the added complexity of MViT did not translate into better performance, suggesting it may be more suited to larger datasets where its advanced temporal resolution could be fully leveraged. Currently, StembryoNet is limited to classification at advanced stages, as embryo synchronization—where an embryologist annotates the time point of similar development—is only feasible at that stage. Synchronization at earlier time points may enable StembryoNet to perform early classification in the future.

At the cell-seeding stage, our best-performing deep learning model achieved a classification accuracy of 65% (F1 = 43%), significantly outperforming a random classifier ($p < 0.0001$). The critical role of fluorescence data at this stage indicates that early developmental predictions rely heavily on initial cell count. By doubling and tripling the number of ESC, ESC-iGata4, and TSC cells, we increased the ratio of *normal* ETiX-embryos, underscoring the importance of precise cell seeding techniques. However, with larger datasets and enhanced imaging techniques, such as more frequent imaging intervals and higher spatial resolution, we may uncover subtle developmental cues in brightfield images at the cell-seeding stage that correlate with successful outcomes.

As ETiX embryogenesis progresses, raw fluorescence data become less informative, and morphological characteristics gain predictive importance. *Normal* ETiX-embryos follow a distinct morphological trajectory, characterized by overall larger sizes, more compact shapes in early stages (i.e., from 15 to 60 h), and less spherical shapes post-radial symmetry-breaking (i.e., from 65 h onwards). During this period, $ResNet_{Fluor}$ outperforms $ResNet_{BF}$. We hypothesize that $ResNet_{Fluor}$ can detect the formation of properly sorted tissue compartments during this period, while $ResNet_{BF}$ is limited to overall ETiX-embryo shape.

https://doi.org/10.1038/s41467-025-56908-5

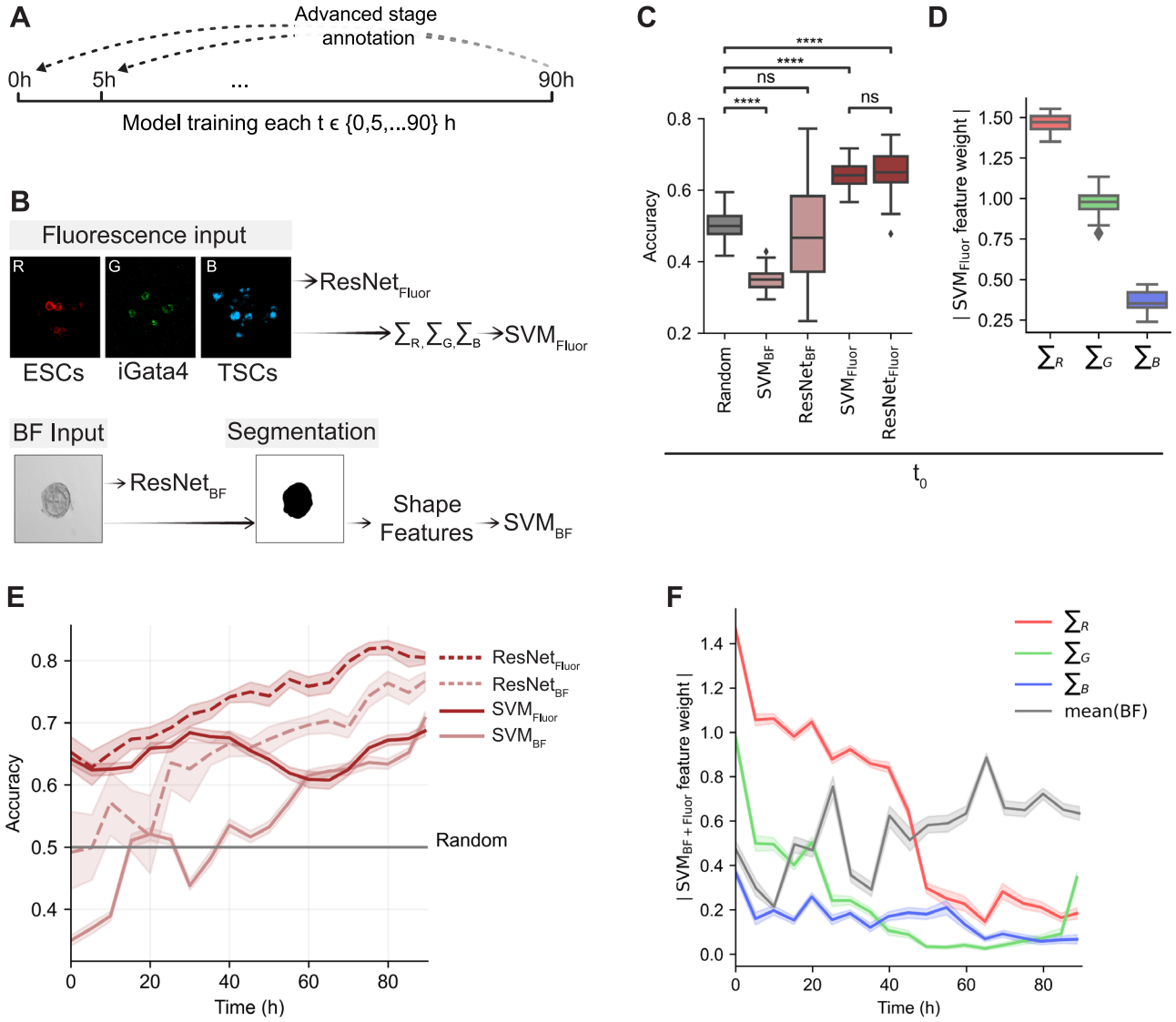

Fig. 3 | AI-based prediction of future *normal* ETiX-embryo development. **A** Diagram illustrating the approach to predicting future developmental outcomes of ETiX-embryos and identifying key classification features over time. Embryologist annotations from 65 to 90 h served as target labels. Models were trained at 5-h intervals across the 0–90-h timeframe, using 5-fold cross-validation repeated 5 times (n = 25) for each interval. **B** ResNet and Support Vector Machine (SVM) models were trained on various types of images and features for explainable classification. ResNet models included ResNet$_{BF}$ and ResNet$_{Fluor}$, trained on brightfield and fluorescence images, respectively. SVM models were based on inferred ETiX-embryo characteristics, namely fluorescent intensity for SVM$_{Fluor}$ and shape features extracted from brightfield ETiX-embryo segmentations for SVM$_{BF}$. **C** Classification performance of ResNet$_{BF}$, ResNet$_{Fluor}$, SVM$_{BF}$, and SVM$_{Fluor}$ at the time of seeding (t_0) across five times repeated 5-fold cross-validation. SVM$_{Fluor}$ vs. Random: $t(104) = 23.1$, $p = 1.1 \times 10^{-42}$, $d = 3.9$, 95% CI [0.13, 0.15], ResNet$_{Fluor}$ vs. Random: $t(27) = 10.6$, $p = 3.1 \times 10^{-11}$, $d = 3.4$, 95% CI [0.12, 0.18], SVM$_{BF}$ vs. Random: $t(112) = -27.1$, $p = 5.1 \times 10^{-51}$, $d = -4.5$, 95% CI [−0.17, −0.14], ResNet$_{BF}$ vs. Random: $t(25) = -0.4$, $p = 0.72$, $d = -0.2$, 95% CI [−0.08, 0.05], SVM$_{Fluor}$ vs. ResNet$_{Fluor}$: $t(30) = -0.7$, $p = 0.46$, $d = -0.2$, 95% CI [−0.04, 0.02]. **D** Absolute feature weights of different channels of SVM$_{Fluor}$ model at the time of seeding (t_0) **E** Classification performance of all models along complete ETiX-embryos observation time, from 0 to 90 h. **F** Absolute SVM feature weights of SVM$_{BF + Fluor}$ model throughout observation time. Feature importances of brightfield features were averaged. **E**, **F** Data are presented as mean values with error bar CI 95%. Box plots:center line, median; box limits, first and third quartile; whiskers, smallest/largest. value no further than 1.5*IQR from the corresponding hinge. ****P < 0.0001, ns not significant, two-sided Welch's t-test. Source data are provided as Source Data file.

Previous work focused on manually selected ETiX-embryos around 4–8 days after seeding[6,14] for further molecular and cellular characterization, leaving the early stage of self-organization relatively unexplored. Our AI models, trained at distinct time points from 0 to 90 h, enable the selection of embryos at earlier stages when human selection is not feasible. For example, our model trained on data at 60 h, achieving a classification accuracy of 76%, can select ETiX-embryos at the time of radial symmetry breaking for further analysis of this important developmental milestone. Furthermore, despite the predictive power of the cell count at seeding for *normal* development, ETiX-embryos with adequate initial cell setups can still fail at later stages and vice versa. Our deep learning model successfully identified such anomalies (Fig. 4). Future studies could utilize single-cell RNA sequencing to explore the mechanisms underlying these developmental deviations, which are central to the system's highly self-organizing properties. Building on this concept, our deep-learning-based automated selection of well-developed ETiX-embryos could be integrated into microscope software, utilizing photoactivable dyes to select *normal* ETiX-embryos at any stage for further characterization.

In summary, this study not only enhances our ability to reliably classify and predict ETiX-embryo development using deep learning but also deepens our understanding of the developmental dynamics

https://doi.org/10.1038/s41467-025-56908-5

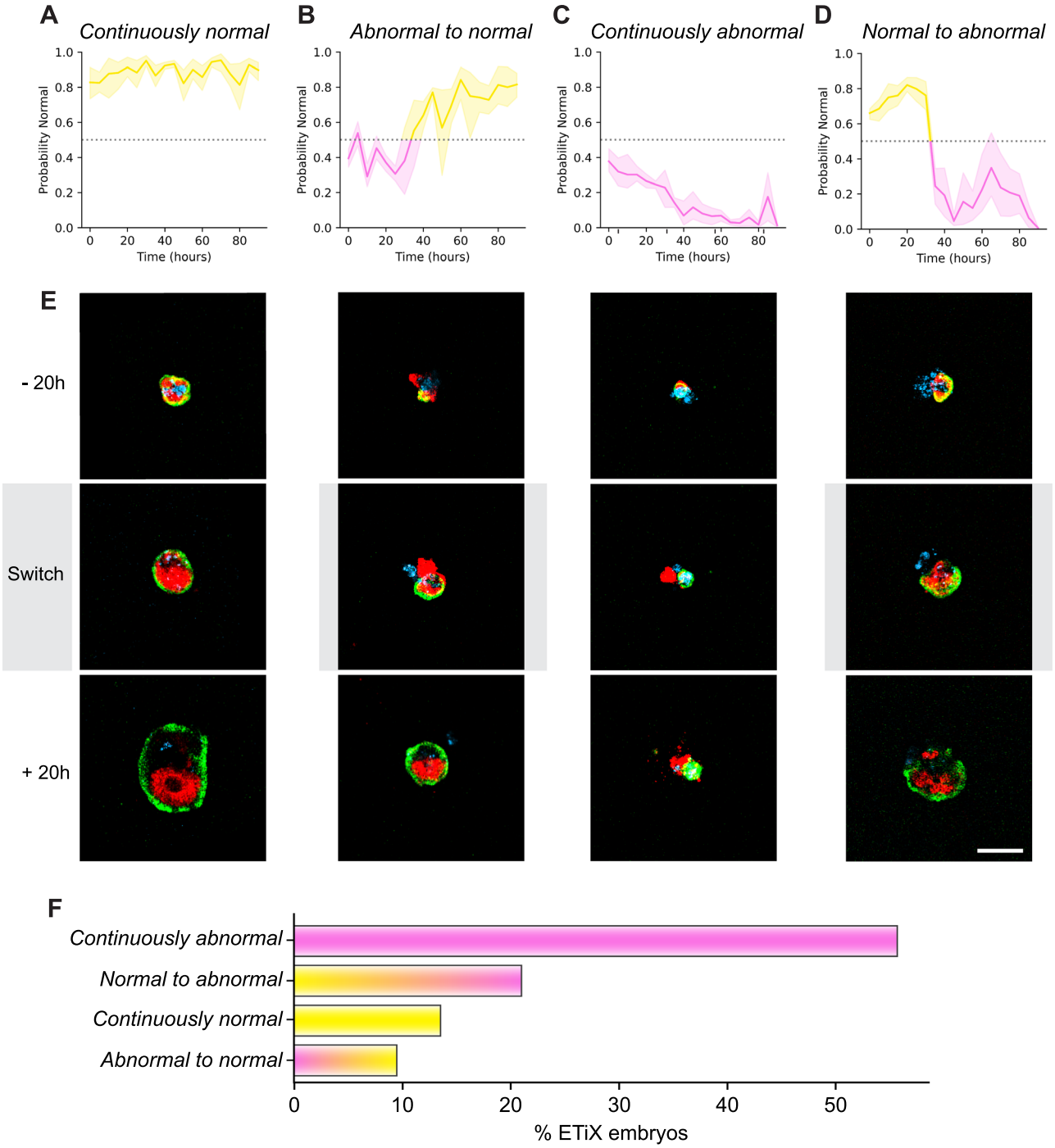

Fig. 4 | Time course AI prediction allows the identification of distinct ETiX-embryo developmental trajectories. A, **B** ResNet model predictions of the probability of being classified as *normal* for two selected *normal* ETiX-embryos plotted over the entire observation period. For this application, ResNet was trained on unsynchronized data on distinct time points along the complete observation time, from 0 to 90 h in 5-h intervals, using five repeated 5-fold cross-validation runs. (**C**, **D**) same as (**A**, **B**) but for two selected *abnormal* ETiX-embryos. **E** ETiX-embryos from (**B**, **D**) are shown at −20 h and +20 h relative to the model prediction switch, as well as at the moment of the switch. For (**A**), we show the same time point as in (**B**), and for (**C**), the same as in (**D**). **F** The proportion of embryos in different categories based on ResNet-predicted probabilities throughout the observation period. Scale bar: 100 μm. **A**–**D** Data are presented as mean values with error bar CI 95%. Source data are provided as Source Data file.

involved. These insights pave the way for a reliable selection of ETiX-embryos for further research throughout the entire observation period, contributing to the emerging field of stem cell-derived models' embryology. Finally, the methodologies and findings presented can be extended to all stem cell-derived embryo models, both in mouse and humans, broadening the impact of our work and opening new avenues for research and clinical innovations.

Methods

Cell culture and transgenic lines

Embryonic stem cells (ESCs) were maintained on plates coated with 0.1% gelatin. The culture medium used was N2B27, which is a 50:50 mixture of Dulbecco's Modified Eagle Medium/F12 (11320033) and Neurobasal-A media (10888022), supplemented with 0.5% (v/v) N2 (17502048), 1% (v/v) B27 (17504044), 100 μM β-mercaptoethanol

(31350-010), 1% (v/v) penicillin-streptomycin (15140122), and 1% (v/v) GlutaMAX (35050079). To promote an undifferentiated state, the medium was enriched with 1 µM of PD0325901 (72182), 3 µM CHIR99021 (72052), and 10 ng/ml of leukemia inhibitory factor (78056.1). The cells were cultured at 37 °C in a humidified atmosphere containing 5% CO_2 and passaged bi-daily. Weekly PCR tests were conducted to confirm the absence of Mycoplasma contamination. The ESC lines used included CAG-GFP/tetO-mCherry/tetO-Gata4 and mtmg ESCs, as established by refs. 6,12.

Trophoblast stem cells (TSCs) were cultured on a layer of mouse embryonic fibroblasts (MEFs) inactivated with mitomycin C. The medium used was RPMI 1640 (11875093) supplemented with 20% fetal bovine serum, 1% (v/v) penicillin-streptomycin (15140122), and 1% (v/v) GlutaMAX (35050079) and 1 mM sodium pyruvate (11360070). Additionally, the growth factors FGF4 at 25 ng/ml and heparin at 1 µg/ml were included to support cell growth and maintenance. Cells were incubated under the same conditions as ESCs and similarly passaged every other day. Wild-type TS cells were stained with Far Red Cell Mask (C37608) for 20 min at 1:1000 before the start of the experiment.

Live imaging of ETiX

The live imaging platform was prepared using a 24-well glass-bottom plate. Each well was first lined with a 1.5% solution of hot agarose. Subsequently, a 3D-printed plunger with hydrophobic treatment and designed with indentations at the end was used to cast 1200 microwells into the solidifying agarose, creating the structure necessary for cell accommodation (we named this platform AgarWell, in assonance with the commercially available AggreWell).

ETiX-embryos were generated following the procedure described previously[24]. In summary, a cell suspension containing 6000 embryonic stem cells (ESCs), 6000 ESCs Gata4-induced embryonic stem cells (ESC-iGata4), and 19,200 trophoblast stem (TSCs) cells was prepared. The cells were resuspended in FC medium, consisting of DMEM (11995-065) with 12.5% FBS, 2 mM GlutaMax (11995-065), 0.1 mM 2-mercaptoethanol (31350-010), 0.1 mM nonessential amino acids (11140-050), 1 mM sodium pyruvate (11360-070), 1% penicillin-streptomycin (15140-122), along with 7.5 nM ROCK inhibitor (72304). This cell mixture was then carefully added to the previously prepared microwells in the imaging platform.

On the day following cell seeding (Day 1), 1 ml of the medium was removed from each well and replaced with a fresh FC medium devoid of ROCK inhibitor. This medium exchange was repeated twice to ensure the removal of the inhibitor. On Day 2, the medium was again replaced with fresh FC medium. On Day 3, the medium in each well was substituted with IVC1 medium, which consists of advanced DMEM/F12 (21331-020) supplemented with 20% (vol/vol) FBS, 2 mM GlutaMax, 1% (vol/vol) penicillin-streptomycin, 1× ITS-X (51500-056), 8 nM β-estradiol, 200 ng/ml progesterone, and 25 mM N-acetyl-L-cysteine.

Live imaging was conducted using a Zeiss LSM980 microscope equipped with a 10× objective lens. The imaging was performed at a stable temperature of 37 °C in a humidified atmosphere containing 5% CO_2, optimized for embryonic development observation. ETiXs were imaged every 35 min by collecting stacks of 4 µm z-planes. Different time series were concatenated using Fiji.

ETiXs annotation and staging at advanced stages of development

Each ETiX time-lapse underwent a detailed examination by an expert embryologist, who reviewed the last 25 h of the series of images along with every Z-plane and various imaging channels. ETiXs that exhibited a cylindrical shape and contained two distinct compartments—one from Trophoblast Stem Cells (TSCs) and the other from Embryonic Stem Cells (ESCs), all enveloped by a monolayer resembling the visceral endoderm composed of ESC-iGata4 cells—were classified as *normal* ETiX-embryos. Structures that failed to meet these criteria were deemed *abnormal*. Given the variability in developmental rates among ETiXs, a specific and consistent endpoint was established for all observations in the synchronized dataset, ensuring uniformity in data analysis.

Data preparation

To derive one image file per ETiX-embryo, ETiX-embryos were manually segmented at the last time point. ETiX-embryos that grew beyond the imaging border were excluded. Subsequently, the center point of each segmented ETiX-embryo was used to extract individual ETiX-embryo images of size 153 × 40 × 288 × 288 × 4 (T × Z × X × Y × C). As in few cases, the ETiX-embryos were growing in the imaging frame of neighboring ETiX-embryos, for each ETiX-embryo any neighboring ETiX-embryo was masked. Therefore, parts of the image belonging to neighboring ETiX-embryos were replaced with values 120 (brightfield) and 0 (fluorescence). In total, we ended up with 900 ETiX-embryos. To correct for varying brightness of the 25 (5 × 5) mosaic tiles in the brightfield channel, patches of the same z-plane were normalized by the mean intensity of all patches in this z-plane.

The dataset comprises 4 channels and 40 z-planes. However, deep-learning models for image classification are mainly designed to take 3-channel images as input. Therefore, we explored different data inputs most feasible for deep learning training:

- Brightfield in-focus images. These are obtained by selecting the z-plane where the brightfield images are most sharply focused. The in-focus brightfield plane was selected based on the z plane with the maximum Laplacian variance as commonly done for automatic focus detection[25].
- Fluorescence in-focus images. These images are derived from the z-plane where the brightfield image is in focus, selected based on the brightfield z-plane with the maximum Laplacian variance. We used the selected z-plane from the brightfield in-focus image, since we noticed that the relevant structures in the fluorescence images are likely showing at the same z plane where the brightfield image is in-focus.
- Fluorescence z-sum projection images. The idea is to combine all the slices in the z-stack by summing the pixel intensities along the *z*-axis.
- Four-channel image of brightfield and fluorescence in-focus image combined.

The results demonstrate that fluorescence information significantly surpasses brightfield information in performance, with fluorescence in-focus images showing a superior outcome compared to fluorescence z-sum projection images (Supplementary Fig. 3A). Combining the brightfield and fluorescence information did not lead to an improved performance (Supplementary Fig. 3A). After all considerations, we therefore used fluorescence in-focus images for model training.

AI-based ETiX-embryo classification at advanced stages of development

StembryoNet is a deep-learning-based architecture (Fig. 2A). StembryoNet leverages a ResNet18[18] backbone, pre-trained on ImageNet, incorporating five consecutive convolutional layers, global average pooling, a fully connected layer, and a sigmoid activation function. We selected ResNet for its rapid and reliable convergence in image classification[21]. Additionally, ResNet demonstrated the smallest standard deviation among various deep learning architectures, indicating greater stability and robustness (Supplementary Fig. 3C). StembryoNet was trained on what we refer to as synchronized data. For synchronized data, an expert embryologist annotated for each embryo the time point of similar development (Supplementary Fig. 2A, B). The StembryoNet backbone was trained on single-time points of embryos at the synchronized time point (Supplementary Fig. 2A, B) for 200 epochs with batch size 16 using optimizer Adam (learning rate = 0.001, beta1 = 0.9, beta2 = 0.999, weight decay = 0.0001) and a binary cross

entropy loss weighted by inverse class frequencies. For *abnormal* embryos, synchronization was not feasible, so we sampled their synchronized time points using a normal distribution with the same mean and standard deviation as those of *normal* embryos. We used the Python package *PyTorch Lightning* (version 2.0.4) for model training.

StembryoNet inference is based on passing consecutive embryo images of the last 25 h through the model. Then, the model-predicted probability for class *normal* of each time point is concatenated. Finally, the maximum probability is thresholded by θ to derive the prediction. θ is the threshold resulting in the highest F1-Score on the unsynchronized validation set (Supplementary Fig. 2D) and ensures balancing the trade-off between precision and recall of predicted *normal* embryos.

For comparison to StembryoNet, we trained several models using the same training procedure as for the StembryoNet backbone. These included ResNet18, MobileNet[17], ResNeXt[26], GoogleNet[27], and DenseNet[28], all trained on images at 90 h. For the Multiscale Vision Transformer (MViT)[29], we used the architecture $MViT_{base}$, pretrained on Kinetics-400, with frame length 16 and sample rate 4. We trained this model for 20 epochs with batch size 1, Adam ($beta_1 = 0.9$, $beta_2 = 0.999$, weight decay = 0.0001). To find an optimal initial learning rate, we used the PyTorch automatic learning rate finder prior to each MViT training. All together, we trained $MViT_{65–90\,h}$ which is trained on videos spanning 65–90 h. To ensure optimal use of the pretrained weights, 16 images were equally sampled from 65 to 90 h for model training and testing. For MViT training, we used *PyTorch Lightning* and *PyTorchVideo* (version 0.1.5).

All models were trained with a binary cross entropy loss weighted by inverse class frequencies. On-the-fly image augmentations comprised random rotation (0–360°) and color jittering (using PyTorch's ColorJitter), which involved adjusting the saturation (factor = 0.1) and contrast (factor = 0.1). The images were normalized according to ImageNet default normalization. Model training and evaluation comprised five-times repeated 5-fold cross-validation, stratified by ETiX-embryo class (Supplementary Fig. 2C). For each split, 75% of the training set is reserved for model training, and 25% is reserved for model validation. The validation set has two purposes. First, it is used to select the best model according to the lowest validation loss. Second, it is used to determine the best threshold θ for late classification using StembryoNet (Supplementary Fig. 2D). θ is the threshold resulting in the highest F1-Score on the validation set. Two-sided Welch's t-tests were used to compare model performances, ensuring the normality assumption was met. For instance, StembryoNet's accuracy distribution across five times repeated 5-fold cross-validation yielded a Shapiro-Wilk p-value of 0.56, indicating normality.

To generate attention heatmaps, we used Gradient-weighted Class Activation Mapping (GradCAM)[30]. We generated the heatmaps for StembryoNet for selected samples at the time point with the highest probability of the embryo being *normal*. Our implementation used the last feature extraction layer as the model target layer and backpropagated the ETiX-embryo ground truth label.

For additional validation, we compared StembryoNet's accuracy and efficiency against three embryologists (A1, A2, and A3). StembryoNet was trained using each embryologist's annotations individually. A1–A3 were included in this multi-annotator comparison, A1's annotations were used in the analysis presented in the main text. To benchmark annotation speed, we measured the time A2 and A3 required to annotate the entire dataset of 900 ETiX-embryos and compared this to the time taken by StembryoNet to predict labels for the same dataset across five repeated 5-fold cross-validation runs, using an NVIDIA A100-40 GPU.

Random classifier

We simulated a random classifier by drawing a value from a uniform distribution between −5 and 5, which was then transformed using a sigmoidal activation function to derive predicted probabilities for each embryo in the respective test set. This random classifier was evaluated with twenty-times repeated 5-fold cross-validation to benchmark against our models.

Model evaluation metric

To assess the model performance, we used Accuracy and F1-Score (F1). Accuracy is calculated as the number of correct predictions divided by the total number of predictions:

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \times 100 \quad (1)$$

with TP = true positives, TN = true negatives, FP = false positives, and FN = false negatives. The F1 is the harmonic mean of precision and recall, providing a single score that balances both aspects:

$$F1 = 2 \times \frac{precision \times recall}{precision + recall} \times 100 \quad (2)$$

$$precision = \frac{TP}{TP + FP} \quad (3)$$

$$recall = \frac{TP}{TP + FN} \quad (4)$$

A perfect classification achieves an Accuracy and F1 of 100%.

AI-based prediction of future *normal* ETiX-embryo development

For AI-based early classification, we employed two types of models, a ResNet18 and a Support Vector Machine (SVM), utilizing both brightfield and fluorescence images: $ResNet_{BF}$, $ResNet_{Fluor}$, SVM_{BF}, SVM_{Fluor}, and $SVM_{BF+Fluor}$. All models were trained on 0–90 h in steps of 5 h. The ResNet18 was trained with the same number of epochs, batch size, optimizer, loss, and image augmentation as the StembryoNet. $ResNet_{BF}$ and $ResNet_{Fluor}$ are trained with the same configuration as explained above. $ResNet_{BF}$ used the in-focus z plane from the brightfield channel as input.

SVM_{BF}, SVM_{Fluor}, and $SVM_{BF+Fluor}$ use a linear kernel for simplicity and interpretability of feature weights. Using a polynomial kernel instead of a linear kernel for the SVM resulted in an overfitted SVM only predicting the majority class and, therefore did not yield improved performance. As the ResNet, the SVM is trained and evaluated with five times repeated 5-fold cross-validation stratified by ETiX-embryo class and weighted classes based on inverse class frequencies. Z-score normalization was applied to the input data. To derive the SVM predictions, the predicted SVM values ('SVC.decision_function()') were thresholded using a sigmoidal activation function. This yielded improved results compared to using 'SVC.predict_proba()'. We used the $SVM_{BF+Fluor}$ coefficients (also termed 'weights') to determine feature importance. As the coefficient magnitude indicates the strength of the feature influence, we calculated the absolute value of each coefficient to extract feature importance. For SVM training and evaluation, we utilized the Python package *scikit-learn* (version 0.24.2).

Input to SVM_{BF} is ETiX-embryo shape features based on brightfield ETiX-embryo segmentations. For brightfield ETiX-embryo segmentation, we trained a SegFormer[31] model on 450 selected brightfield images in total. For SegFormer training, we used the GitHub repository *mmsegmentation* (version 0.30.0). SegFormer: MiT-B0 architecture, pre-trained on ADE20k, a combination of Dice Loss and Cross Entropy Loss (weighted 10:1), AdamW (lr = 0.0001, $beta_1 = 0.9$, $beta_2 = 0.999$, weight decay = 0.1), 1000 training iterations. On-the-fly image augmentations included three steps: random flip (p = 0.5), adding Gaussian Noise (variance range: 0.01–0.1), and z-score normalization. These images were sampled every 20-time points from 60 randomly selected ETiX-embryos and manually annotated. 80% of the images were used

for SegFormer training, and 20% were used for model testing. The split was made on the ETiX-embryo level to avoid information leakage between the training and test set. The trained model was used to infer the ETiX-embryo segmentations for the complete dataset. We filled the binary holes in the resulting segmentations using the binary_fill_holes function of *scikit-learn* (version 0.24.2). 0.2% of all segmentations showed no segmented embryo, mostly early on when the single cells had not form compacted tissue yet; those values were filled with the mean feature value at the respective time point in the following step. To quantify ETiX-embryo morphology and shape, we used *PyRadiomics*[32] (version 3.1.0) to extract the following 2D features: Elongation, MajorAxisLength, MaximumDiameter, MeshSurface, MinorAxisLength, Perimeter, PerimeterSurfaceRatio, PixelSurface, and Sphericity. SVM_{Fluor} takes three values as input: for each channel the sum of fluorescence intensities in the fluorescence in-focus image.

To count ESC, ESC-iGata4, and TSC cells of each ETiX-embryo at the cell-seeding time point, we applied the cyto3 model from the Python package CellPose[33] (version 2.2.3), using a cell pixel diameter of 7 on the corresponding fluorescence channel of the in-focus images. Each channel, representing a specific cell type, was used for a separate prediction. Afterward, the cell counts were manually reviewed and corrected by an embryologist.

Morphological trajectory

We performed morphological trajectory analysis for *normal* and *abnormal* ETiX-embryos. For each, the first step was to perform a principal component analysis on the nine *PyRadiomics* calculated ETiX-embryo shape features, calculated from 0 to 90 h in steps of 5 h, as described above. We performed a common PCA on *normal* and *abnormal* ETiX-embryos and conducted subsequent trajectory analysis using Slingshot[23] by class separately. A Python installation of Slingshot was used (https://github.com/mossjacob/pyslingshot/) with two epochs for trajectory fitting.

Self-supervised learning (DINO)

To investigate the potential of self-supervised learning for ETiX-embryo classification, we trained a DINO[34] model on fluorescence images ($DINO_{Fluor}$) using the ViT_{base} architecture for 200 epochs with batch size 32. For the remaining parameters, we used the default values from the DINO GitHub repository (https://github.com/facebookresearch/dino). For hierarchical clustering of $DINO_{Fluor}$ embeddings, we used average linkage clustering based on the Euclidean distance. To visualize the effects of random hierarchical clustering, we shuffled the labels of the ETiX-embryos, providing a baseline comparison for evaluating the DINO clustering performance. For ETiX-embryo classification using DINO embeddings, we trained XGBoost for downstream classification on embeddings from the time period 65–90 h. The XGBoost models were trained and evaluated using ten-time repeated 5-fold cross-validation and the classes weighted according to inverse class frequencies.

Improving cultivation efficiency

To determine the number of predicted *normal* and *abnormal* ETiX-embryos for the datasets with increased initial cell count, $Data_{2X}$ (twofold) and $Data_{3X}$ (threefold), we used the predictions of the 25 StembryoNet models trained using five-time repeated 5-fold cross-validation on ETiX-embryos from $Data_{2X}$ and $Data_{3X}$ at 35 to 64 h. Due to faster ETiX-embryo development for $Data_{2X}$ and $Data_{3X}$, we selected 64 h as the final time point.

Identification and categorization of individual ETiX-embryo developments

To analyze ETiX-embryo developments, we investigated the model-predicted probability for class *normal* for individual ETiX-embryos throughout the observation time. Therefore, we trained a ResNet using five times 5-fold cross-validation on images from every 5 h. For quantification, we subsequently counted the number of ETiX-embryos falling into the categories (1) continuously *normal*, (2) *abnormal* to *normal*, (3) continuously *abnormal*, and (4) *normal* to *abnormal*. This quantification is based on the predicted probability at 0 h and 90 h, and the slope of a fitted linear regression over all time points according to the Table below. To ensure accurate embryo categorization and as we observed that model predictions were more confident towards later observation times, i.e., closer to 0 or 1 rather than near the decision threshold of 0.5, we included the slope of a linear regression in our criteria.

ETiX-embryo category	P(Normal) at 0 h	P(Normal) at 90 h	Regression slope
Continuously normal	≥0.5	≥0.5	≥0
Abnormal to normal	<0.5	≥0.5	≥0
Continuously abnormal	<0.5	<0.5	≤0
Normal to abnormal	≥0.5	<0.5	≤0

To categorize $Data_{2X}$ and $Data_{3X}$ (Supplementary Fig. 5D), we employed the same ResNet models trained on $Data_{1X}$, using 5-h intervals from 0 to 90 h. Given the faster development of ETiX embryos in $Data_{2X}$ and $Data_{3X}$, we adjusted the prediction time points to match equivalent embryo sizes, ensuring comparability (Table below, Supplementary Fig. 5E).

	Hour																	
Training ($Data_{1X}$)	0	5	10	15	20	25	30	35	40	45	50	55	60	65	70	75	80	85
Prediction ($Data_{2X/3X}$)	0	5	10	15	15	20	25	29	35	41	42	44	50	52	55	57	59	61

Reporting summary

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

Data availability

The data generated in this study have been deposited in the Zenodo database available at https://doi.org/10.5281/zenodo.14605093[35]. Source data are provided with this paper.

Code availability

The custom code developed for this study is publicly accessible on GitHub at https://github.com/deiluca/StembryoNet and co-deposited on Zenodo at https://doi.org/10.5281/zenodo.14605177[36].

References

1. Bao, M., Cornwall-Scoones, J. & Zernicka-Goetz, M. Stem-cell-based human and mouse embryo models. *Curr. Opin. Genet. Dev.* **76**, 101970 (2022).
2. Rivron, N. C. et al. Blastocyst-like structures generated solely from stem cells. *Nature* **557**, 106–111 (2018).
3. Kagawa, H. et al. Human blastoids model blastocyst development and implantation. *Nature* **601**, 600–605 (2022).
4. Harrison, S. E., Sozen, B., Christodoulou, N., Kyprianou, C. & Zernicka-Goetz, M. Assembly of embryonic and extraembryonic stem cells to mimic embryogenesis in vitro. *Science* **356**, eaal1810 (2017).
5. Sozen, B. et al. Self-assembly of embryonic and two extra-embryonic stem cell types into gastrulating embryo-like structures. *Nat. Cell Biol.* **20**, 979–989 (2018).
6. Amadei, G. et al. Embryo model completes gastrulation to neurulation and organogenesis. *Nature* **610**, 143–153 (2022).
7. Weatherbee, B. A. T. et al. Pluripotent stem cell-derived model of the post-implantation human embryo. *Nature* **622**, 584–593 (2023).
8. Oldak, B. et al. Complete human day 14 post-implantation embryo models from naive ES cells. *Nature* **622**, 562–573 (2023).

9. Liu, L. et al. Modeling post-implantation stages of human development into early organogenesis with stem-cell-derived peri-gastruloids. *Cell* **186**, 3776–3792.e16 (2023).
10. Fu, J., Warmflash, A. & Lutolf, M. P. Stem-cell-based embryo models for fundamental research and translation. *Nat. Mater.* **20**, 132–144 (2021).
11. Bao, M. et al. Stem cell-derived synthetic embryos self-assemble by exploiting cadherin codes and cortical tension. *Nat. Cell Biol.* **24**, 1341–1349 (2022).
12. Amadei, G. et al. Inducible stem-cell-derived embryos capture mouse morphogenetic events in vitro. *Dev. Cell* **56**, 366–382.e9 (2021).
13. Lau, K. Y. C. et al. Mouse embryo model derived exclusively from embryonic stem cells undergoes neurulation and heart development. *Cell Stem Cell* **29**, 1445–1458.e8 (2022).
14. Tarazi, S. et al. Post-gastrulation synthetic embryos generated ex utero from mouse naive ESCs. *Cell* **185**, 3290–3306.e25 (2022).
15. Egger, J. et al. Medical deep learning—A systematic meta-review. *Comput. Methods Programs Biomed.* **221**, 106874 (2022).
16. Krizhevsky, A., Sutskever, I. & Hinton, G. E. ImageNet classification with deep convolutional neural networks. in *Proc. 25th International Conference on Neural Information Processing Systems - Volume 1* 1097–1105 (Curran Associates Inc., 2012).
17. Howard, A. G. et al. MobileNets: Efficient Convolutional Neural Networks for Mobile Vision Applications. Preprint at https://doi.org/10.48550/arXiv.1704.04861 (2017).
18. He, K., Zhang, X., Ren, S. & Sun, J. Deep Residual Learning for Image Recognition. in 770–778 (IEEE Computer Society, 2016). https://doi.org/10.1109/CVPR.2016.90.
19. Matthews, J. M. et al. OrganoID: a versatile deep learning platform for tracking and analysis of single-organoid dynamics. *PLoS Comput. Biol.* **18**, e1010584 (2022).
20. Deininger, L. et al. An AI-based segmentation and analysis pipeline for high-field MR monitoring of cerebral organoids. *Sci. Rep.* **13**, 21231 (2023).
21. Čapek, D. et al. EmbryoNet: using deep learning to link embryonic phenotypes to signaling pathways. *Nat. Methods* **20**, 815–823 (2023).
22. Bedzhov, I. & Zernicka-Goetz, M. Self-organizing properties of mouse pluripotent cells initiate morphogenesis upon implantation. *Cell* **156**, 1032–1044 (2014).
23. Street, K. et al. Slingshot: cell lineage and pseudotime inference for single-cell transcriptomics. *BMC Genomics* **19**, 477 (2018).
24. Stem cell-derived mouse embryos develop within an extra-embryonic yolk sac to form anterior brain regions and a beating heart-like structure. https://www.researchsquare.com (2022) https://doi.org/10.21203/rs.3.pex-2006/v1.
25. Groen, F. C. A., Young, I. T. & Ligthart, G. A comparison of different focus functions for use in autofocus algorithms. *Cytometry* **6**, 81–91 (1985).
26. Xie, S., Girshick, R., Dollár, P., Tu, Z. & He, K. Aggregated Residual transformations for deep neural networks. Preprint at https://doi.org/10.48550/arXiv.1611.05431 (2017).
27. Szegedy, C. et al. Going deeper with convolutions. In *Proc. 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* 1–9 (2015). https://doi.org/10.1109/CVPR.2015.7298594.
28. Huang, G., Liu, Z., Maaten, L., Weinberger, K. Q. Densely Connected Convolutional Networks. Preprint at https://doi.org/10.48550/arXiv.1608.06993 (2018).
29. Fan, H. et al. Multiscale Vision Transformers. Preprint at https://doi.org/10.48550/arXiv.2104.11227 (2021).
30. Selvaraju, R. R. et al. Grad-CAM: Visual Explanations from Deep Networks via Gradient-Based Localization. *Int. J. Comput. Vis.* **128**, 336–359 (2020).
31. Xie, E. et al. SegFormer: Simple and efficient design for semantic segmentation with transformers. Preprint at https://doi.org/10.48550/arXiv.2105.15203 (2021).
32. van Griethuysen, J. J. M. et al. Computational radiomics system to decode the radiographic phenotype. *Cancer Res.* **77**, e104–e107 (2017).
33. Stringer, C., Wang, T., Michaelos, M. & Pachitariu, M. Cellpose: a generalist algorithm for cellular segmentation. *Nat. Methods* **18**, 100–106 (2021).
34. Caron, M. et al. Emerging properties in self-supervised vision transformers. Preprint at https://doi.org/10.48550/arXiv.2104.14294 (2021).
35. Caldarelli, P. & Deininger, L. AI-based approach to dissect the variability of mouse stem cell-derived embryo models: data. *Zenodo* https://doi.org/10.5281/zenodo.14605093 (2025).
36. Deininger, L. deiluca/StembryoNet: AI-based approach to dissect the variability of mouse stem cell-derived embryo models: code. *Zenodo* https://doi.org/10.5281/zenodo.14605177 (2025).

Acknowledgements

We thank Wenqi Hu for the annotation of the dataset. This work was supported by the MZG.PIONEER.1.NIHP (HD104575A to M.Z.-G.), NOMIS Foundation (12540449 to M.Z.-G.), the Wellcome Trust (207415/Z/17/Z to M.Z.-G.), the Open Philanthropy Project (to M.Z.-G.), the Helmholtz Association under the joint research school "HIDSS4Health" – Helmholtz Information and Data Science School for Health to L.D., and the Helmholtz program NACIP to R.M. and L.D. This work is supported by the Helmholtz Association Initiative and Networking Fund on the HAICORE@KIT partition.

Author contributions

The project was originally conceptualized by P.C. and M.Z.-G and developed by P.C., M.Z.-G, L.D., R.M., C.Y. and P.C. designed and performed experiments with help from P.P. and L.D. developed the deep learning model with help from S.Z., P.C. and L.D. analyzed the data. P.C., L.D., and M.Z.-G. wrote the manuscript with comments from the other co-authors. M.Z.-G., R.M., and C.Y. supervised the project and acquired funding.

Competing interests

The authors declare no competing interests.

Additional information

Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s41467-025-56908-5.

Correspondence and requests for materials should be addressed to Ralf Mikut or Magdalena Zernicka-Goetz.

Peer review information *Nature Communications* thanks the anonymous, reviewer(s) for their contribution to the peer review of this work. A peer review file is available.