SOFTWARE TOOL ARTICLE

REVISED Automated staging of zebrafish embryos using machine learning [version 3; peer review: 1 approved, 2 approved with reservations]

Rebecca A. Jones[1,2*], Matthew J. Renshaw[3*], David J. Barry[3*], James C. Smith[1]

[1]Developmental Biology Laboratory, The Francis Crick Institute, 1 Midland Road, London, NW1 1AT, UK
[2]Department of Molecular Biology, Princeton University, Princeton, NJ, 08544, USA
[3]Crick Advanced Light Microscopy (CALM), The Francis Crick Institute, 1 Midland Road, London, NW1 1AT, UK

* Equal contributors

v3 **First published:** 09 Nov 2022, **7**:275
https://doi.org/10.12688/wellcomeopenres.18313.1
Second version: 16 Mar 2023, **7**:275
https://doi.org/10.12688/wellcomeopenres.18313.2
Latest published: 26 Apr 2023, **7**:275
https://doi.org/10.12688/wellcomeopenres.18313.3

Abstract

The zebrafish (*Danio rerio*), is an important biomedical model organism used in many disciplines, including development, disease modeling and toxicology, to better understand vertebrate biology. The phenomenon of developmental delay in zebrafish embryos has been widely reported as part of a mutant or treatment-induced phenotype, and accurate characterization of such delays is imperative. Despite this, the only way at present to identify and quantify these delays is through manual observation, which is both time-consuming and subjective. Machine learning approaches in biology are rapidly becoming part of the toolkit used by researchers to address complex questions. In this work, we introduce a machine learning-based classifier that has been trained to detect temporal developmental differences across groups of zebrafish embryos. Our classifier is capable of rapidly analyzing thousands of images, allowing comparisons of developmental temporal rates to be assessed across and between experimental groups of embryos. Finally, as our classifier uses images obtained from a standard live-imaging widefield microscope and camera set-up, we envisage it will be readily accessible to the zebrafish community, and prove to be a valuable resource.

Keywords

Zebrafish, development, machine learning, staging, developmental delay, classifier

Open Peer Review

Approval Status ✓ ? ?

	1	2	3
version 3 (revision) 26 Apr 2023			? view
version 2 (revision) 16 Mar 2023	✓ view	? view	
version 1 09 Nov 2022	? view	? view	

1. **Steffen Scholpp**, University of Exeter, Exeter, UK

2. **Amin Allalou**, Uppsala University, Uppsala, Sweden
Science for Life Laboratory BioImage Informatics Facility, Uppsala, Sweden

3. **Christian Tischer**, European Molecular Biology Laboratory, Heidelberg, Germany
Arif Khan, EMBL Heidelberg, Heidelberg, Germany

This article is included in the The Francis Crick Institute gateway.

Sebastian Gonzalez-Tirado, EMBL Heidelberg, Heidelberg, Germany

Any reports and responses or comments on the article can be found at the end of the article.

Corresponding author: David J. Barry (david.barry@crick.ac.uk)

Author roles: Jones RA: Conceptualization, Investigation, Methodology, Writing – Original Draft Preparation; **Renshaw MJ**: Investigation, Methodology, Writing – Review & Editing; **Barry DJ**: Conceptualization, Formal Analysis, Methodology, Software, Visualization, Writing – Review & Editing; **Smith JC**: Funding Acquisition, Supervision, Writing – Review & Editing

Competing interests: No competing interests were disclosed.

Grant information: This work was supported by the Francis Crick Institute which receives its core funding from Cancer Research UK (FC001-157), the UK Medical Research Council (FC001-157), and the Wellcome Trust (FC001-157).
The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

How to cite this article: Jones RA, Renshaw MJ, Barry DJ and Smith JC. **Automated staging of zebrafish embryos using machine learning [version 3; peer review: 1 approved, 2 approved with reservations]** Wellcome Open Research 2023, **7**:275 https://doi.org/10.12688/wellcomeopenres.18313.3

First published: 09 Nov 2022, **7**:275 https://doi.org/10.12688/wellcomeopenres.18313.1

REVISED **Amendments from Version 2**

We have updated our manuscript to remove the potentially confusing statement *'it was not trained to predict the actual hpf of a given embryo'* as our system is solely designed to identify developmental delay, and not the actual hpf of an individual embryo.

Any further responses from the reviewers can be found at the end of the article

Introduction

The zebrafish (*Danio rerio*) is a model organism widely used in a variety of fields, including developmental biology, disease modelling, cancer biology and immunology (Choi *et al.,* 2021; Eisen, 1996; Gomes & Mostowy, 2020; Kemmler *et al.,* 2021; Zanandrea *et al.,* 2020). External fertilization, high fecundity, low cost and ease of genetic manipulation together make zebrafish a valuable model for many studies, and their transparent embryos make them particularly useful in studies of developmental biology (Nusslein-Volhard, 2012). The advent of CRISPR/Cas (Clustered Regularly Interspaced Short Palindromic Repeats/CRISPR associated (protein)) technology has meant that many studies use transgenic lines to answer important biological questions (Liu *et al.,* 2019).

Zebrafish embryos develop externally, becoming free-swimming, independently feeding larvae by five days post fertilization (dpf) (Kimmel *et al.,* 1995). Development is rapid, with gastrulation and neurulation occurring within the first 12 hours post fertilization (hpf) (Kimmel *et al.,* 1995). Many studies, and particularly developmental studies, require accurate staging of zebrafish embryos and larvae. Although the timing of fertilization can be estimated to within ~30 minutes, the numbers of hours post fertilization at the standard temperature of 28.5°C (Kimmel *et al.,* 1995) provides only an approximation of the actual developmental stage, because other factors, like population density and water quality, can affect maturation rates (Singleman & Holtzman, 2014). Even when such factors are controlled for, embryos within a clutch may develop at different rates (Parichy *et al.,* 2009). Researchers therefore use both hpf/dpf and staging guides that are based on morphological criteria to stage individual embryos (Kimmel *et al.,* 1995). These morphological features include the number of somites and the appearance of landmark structures such as the embryonic shield, tail bud and eye primordium (Kimmel *et al.,* 1995; Westerfield, 2000).

Staging of embryos is of particular importance because many studies report 'developmental delay' as part of a genetic or drug-induced phenotype. For example, transgenic lines might develop more slowly than their wild-type (WT) counterparts, (Elabd *et al.,* 2019; Giraldez *et al.,* 2005; Jia *et al.,* 2020; Li *et al.,* 2017), as might embryos injected with antisense morpholino oligonucleotides (Flinn *et al.,* 2008; Hung *et al.,* 2013; Walpita *et al.,* 2010), or those treated with drugs (Akthar *et al.,* 2019; Byrnes *et al.,* 2018; Farooq *et al.,* 2019). Significantly, zebrafish have emerged as important models in which to study the effects of environmental and aquatic toxins, with many of these treatments also resulting in a developmental delay (Aksakal & Sisman, 2020; Li *et al.,* 2020; Mesquita *et al.,* 2017). Such delays are difficult to quantify without manually staging large numbers of embryos, which is inconvenient, subjective and time-consuming, especially when assessing developmental abnormalities (Jeanray *et al.,* 2015; Teixidó *et al.,* 2019). Adding to this difficulty, the delay is often temporary, and transgenic or treated embryos 'catch up' with their WT counterparts (Elabd *et al.,* 2019; Ge *et al.,* 2019; Kamei *et al.,* 2018). It is therefore important to identify exactly when the delay is occurring to account for it in the study. Conversely, in many studies it is necessary to exclude general developmental delay, a potentially confounding variable, as the cause of either a tissue-specific phenotype or a developmental delay induced by a drug treatment or specific mutation, in order to validate the results of a given experiment (Mannucci *et al.,* 2021; Sidik *et al.,* 2021). For example, if one knocks out a gene involved in cardiac development, it is important to determine if any delay in heart formation is cardiac-specific, or part of an organism-wide developmental delay. In some studies, altered hatching rates are used as an additional proxy for developmental stage (Martinez *et al.,* 2018; Tshering *et al.,* 2021; Zhang *et al.,* 2015), yet hatching defects can be caused by hatching gland specific issues, as opposed to a more general developmental delay (Suzuki *et al.,* 2019; Trikić *et al.,* 2011). Because assessing developmental delay is such a critical part of zebrafish related work, it is imperative that we develop a more standardized and automated way to measure it: one that reduces the time and subjectivity burden inherent in manual staging.

The use of image analysis has become increasingly popular in the life sciences, automating the quantification of microscopy images in an unbiased fashion (Meijering *et al.,* 2016). However, designing an image analysis algorithm to detect the wide range of morphological features on which staging guides depend would be a challenging endeavor. Nevertheless, the staging of embryos based on microscopy images is a task to which machine learning is well-suited. Machine learning approaches, where a computer program uses algorithms and statistical models to continuously learn and improve pattern prediction, is already used widely in biological studies (Greener *et al.,* 2022). Several labs have already made successful attempts to automate the analysis of morphological features of zebrafish embryos using machine learning. Jeanray *et al.* (2015) used a supervised machine learning approach to classify bright-field images of zebrafish embryos according to chemical treatment induced defects, with >90% concordance to manual expert classification, and various other studies have produced similar classifiers (Ishaq *et al.,* 2017; Shang *et al.,* 2020). More recently, Guglielmi *et al.* (2021) used an innovative optical projection tomography (OPT) and back-projection technique followed by semi-automated segmentation and quantitation to objectively describe the morphological features of zebrafish embryos in which BMP signaling was perturbed. In terms of developmental staging, Pond *et al.* (2021) recently developed a convolutional neural network (CNN)-based classifier to stage zebrafish tail-buds at four

discrete developmental stages, demonstrating that high accuracy can be achieved with small data sets (<100 images). These elegant systems highlight the power of machine learning approaches in the identification of morphological features and discrete developmental stages, but none of these studies extract sufficient information to enable complete temporal developmental profiles to be compared. For example, Pond *et al.* (2021) compared four developmental stages, and whilst their CNN-based classifier was able to accurately predict these stages, this is not sufficient to extract a comparable developmental profile.

Using a combination of live imaging and machine learning approaches, we have developed a classifier to quantify zebrafish embryonic development, allowing objective and meaningful relative comparisons over time. Moreover, we demonstrate our classifier's ability to stage specific developmental time-points is comparable to human experts. This work provides proof of principle that machine learning algorithms can be used to accurately stage zebrafish embryos and we hope that our classifier will become a valuable resource for the zebrafish community.

Methods

Zebrafish husbandry

All zebrafish work, including housing and husbandry, was undertaken in accordance with institutional (The Francis Crick Institute) and national (UK) ethical and animal welfare regulations, including the Crick Use of Animals in Research Policy, the Animals (Scientific Procedures) Act 1986 (ASPA) implemented by the Home Office in the UK and the Animal Welfare Act 2006. All regulated procedures were carried out at The Francis Crick Institute in accordance with UK Home Office regulations under project license PF59163DB, which underwent full ethical review and approval by The Francis Crick Institute's Animal Ethics Committee. Consideration was given to the '3Rs' in experimental design, and animals were observed on a daily basis for any signs of illness/distress. Any animals displaying evidence of suffering (physiological/behavioral changes, signs of injury) were euthanized in pH neutralized MS222 for a minimum of 30 minutes, before a second physical euthanasia method was performed. The Zirc AB line was used in all experiments. For most experiments, zebrafish embryos were obtained by tank mass-spawning using either a mating tank with a clear Perspex divider or a Mass Embryo Production (MEP) system (MBK Installations). Embryos were collected 30 minutes following divider removal or first-light respectively, and then at 30-minute intervals thereafter until spawning ceased. Embryos were maintained in plates of ~50 animals, at 28.5°C in E2 medium, prepared by The Francis Crick Institute's Media Preparation Facility. Approximately 30 minutes prior to imaging, zebrafish embryos were manually checked for correct development, then individually transferred into separate wells of a 96-well plate, containing pre-warmed (28.5°C or 25°C) E2 medium. One plate of 96 embryos was then transferred to the Crick Advanced Light Microscopy (CALM) Science Technology Platform (STP) imaging suite and mounted in the environmental chamber (see below). One 96-well plate was used for imaging each condition, as a 96-well plate set-up provided optimal conditions for individual embryo image capture. Having 96 embryos per condition also ensured that if several embryos failed to develop normally, there would still be sufficient embryos to perform both training and downstream analysis. Excess embryos were disposed of in MS222 as above.

Live imaging

Zebrafish embryos were maintained at 28.5°C until shortly before four hpf as defined by both hpf and morphological criteria (sphere stage, (Kimmel *et al.,* 1995)), at which point they were transferred into U-bottomed 96-well plates (Thermo Fisher) in E2 medium as described above. Plates were covered with fluorinated ethylene propylene (FEP) membrane (1 mil Teflon FEP film, American Durafilm) to prevent condensation and allow for gas exchange. Brightfield images of embryos individually seeded in 96-well plates were acquired every 15 minutes starting at four hpf for 60 hours using a Nikon Ti2 microscope with 2X/0.1 Plan Apo objective and 1.5x intermediate magnification. A small pixel complementary metal-oxide-semiconductor (CMOS) camera (UI-3280SE, iDS) enabled a whole embryo to be captured in a single field of view at cellular resolution (pixel size 1.15 μm). Sample temperature was maintained at either 25.0 or 28.5°C using an environmental chamber enclosure (Okolab). The microscope was controlled with Micro-Manager v2.0 software (Edelstein *et al.,* 2014, RRID:SCR_000415) and the HCS Site Generator plugin was used to generate a list of positions for the 96-well plate. The workflow is summarized in Figure 1.

Machine learning pipeline configuration

Automated staging of zebrafish embryos was performed using ilastik (Berg *et al.,* 2019) and FIJI (Schindelin *et al.,* 2012), both free, open source software popular among life science researchers. All the FIJI scripts and ilastik project files needed to reproduce these steps are available to download online (Barry, 2022). This repository contains step-by-step instructions that can be used to either reproduce the raw data used to generate the plots in this manuscript, or run the classifier on new data (see README.md in Barry, 2022).

Training of machine learning model

Images of zebrafish embryos were randomly divided into training and test datasets as described in Table 1. Using an ilastik pixel classification pipeline, pixels in training data were manually labelled as belonging to one of three classes: embryo, background or embryo/background boundary (Figure 2A). These labels, together with a range of generic pixel features, were used to then train a random forest classifier using ilastik's pixel classification workflow. All training labels and pixel features used to train the pixel classifier can be viewed in the ilastik pixel classifier project file (PixelClassifier.ilp - Barry, 2022).

The probability map for the boundary class output by ilastik was then used to fully segment the embryos, using simple grey level thresholding in FIJI (Figure 2B;

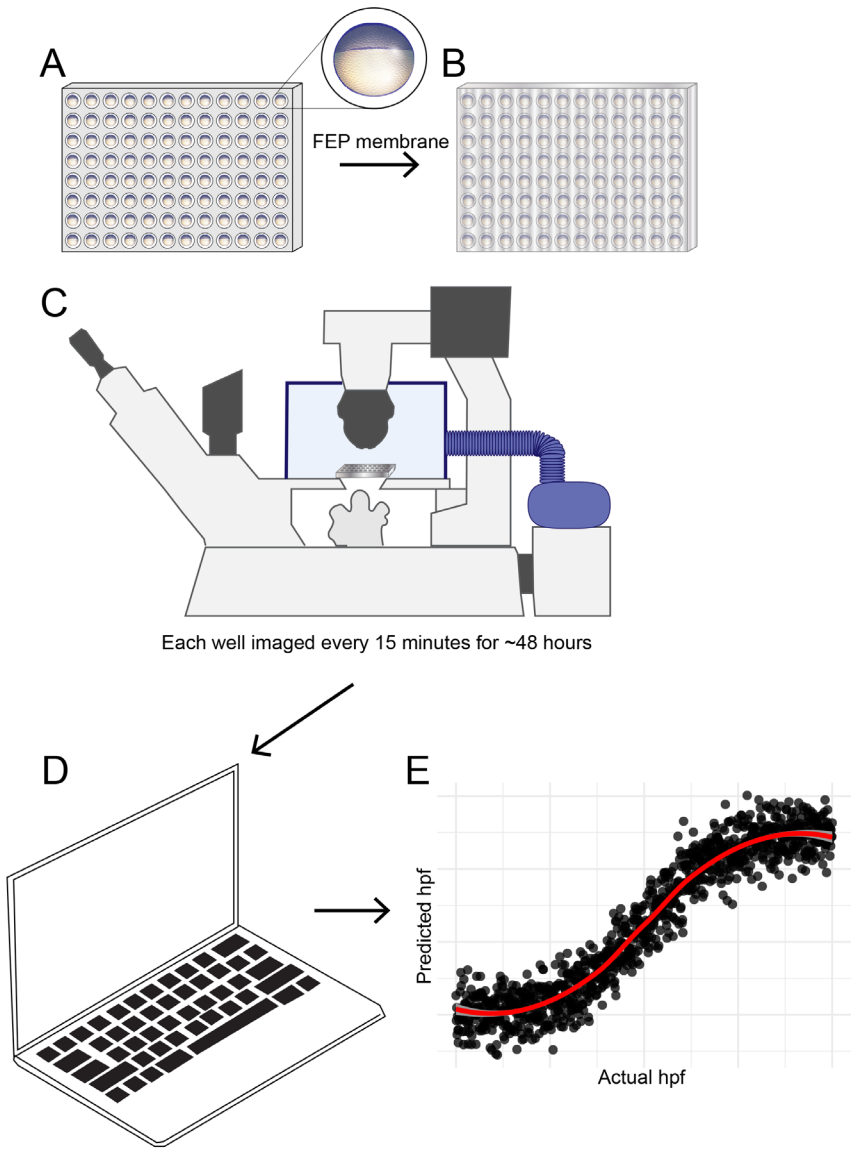

Figure 1. Schematic diagram showing developmental temporal quantification workflow. Zebrafish embryos were individually seeded into U-bottomed 96 well plates in E2 medium at around 3.5 hpf. (**A**) The plate was then sealed with a breathable FEP membrane (**B**) and transferred to an inverted microscope with motorised XY stage where the temperature was maintained at 25.0 or 28.5°C using an environmental chamber (**C**). Brightfield images were captured of each well every 15 mins from sphere stage (4hpf) until 18 hpf. Images were analysed using an ilastik object classification pipeline (**D**) to produce plots showing predicted hpf versus actual hpf, similar to the schematic plot shown (**E**).

Table 1. Overview of datasets used for training and testing.

Plate	Incubation Temperature (°C)	Wells Used for Training	Wells Used for Testing
1	28.5	13	82
2	28.5	11	84
3	25.0	0	95

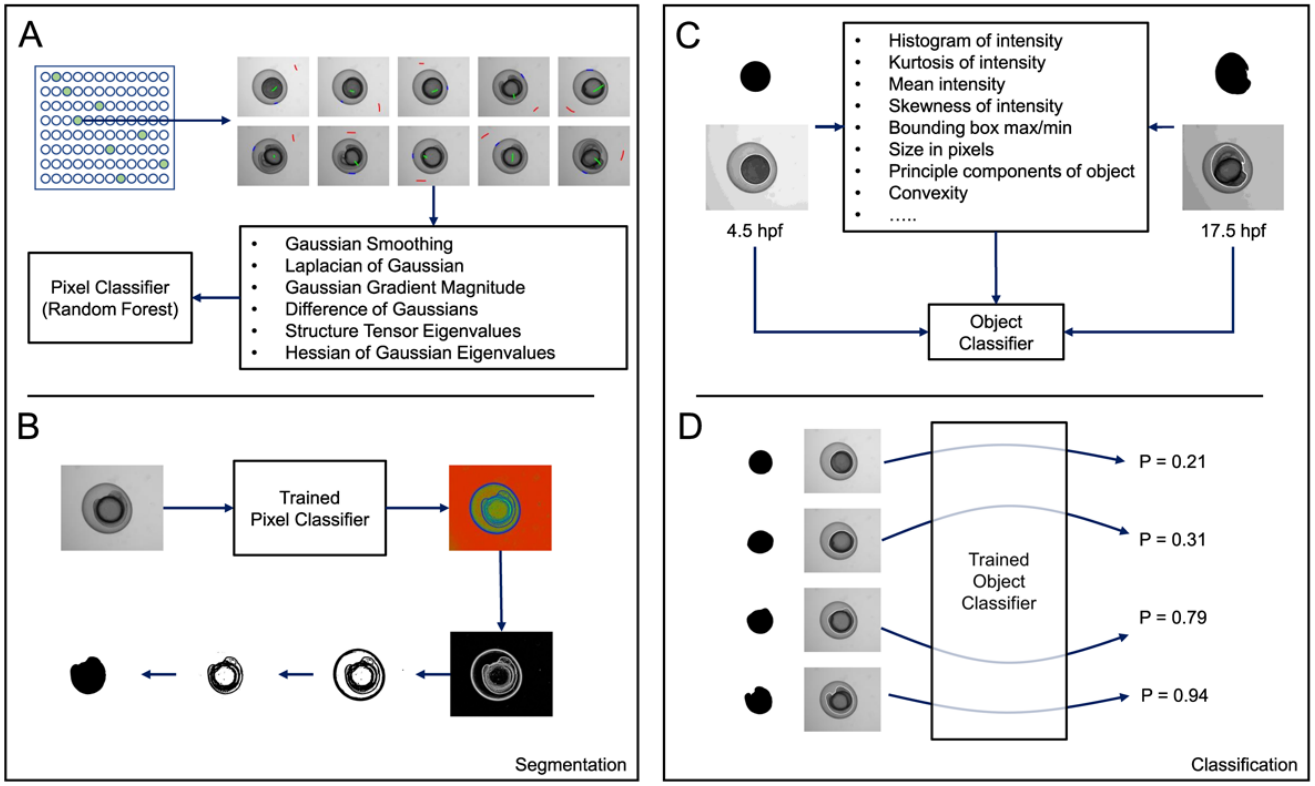

Figure 2. Overview of ilastik-based pixel and object classification pipeline. (**A**) A pixel classifier was trained to segment the embryos in each image. Using a random selection of time-points, regions in images were manually annotated as either background (red), embryo (green), or boundary (blue). Using the measures shown, calculated at various different scales, and the annotations, ilastik then trained a random-forest classifier. (**B**) When supplied with test data, the trained pixel classifier produced three probability maps, one for each of the classes listed in (**A**) (background, embryo, boundary). Each pixel in a map for a particular class gives the probability that the pixel belongs to that particular class. By thresholding the boundary probability map and performing some simple morphological processing on the resulting binary image, we could obtain a mask representing the embryo. (**C**) We then trained an object classification pipeline using ilastik. Using the mask images generated in (**B**) and the corresponding raw images, an object classifier was trained to recognise either 4.5 or 17.5 hpf embryos. (**D**) When supplied with test mask and raw images, the trained object classifier returned a probability corresponding to the likelihood that the test image represented a 4.5 (P = 0.0) or 17.5 hpf (P = 1.0) embryo.

see Segment.ijm - Barry, 2022). The resultant embryo masks were then combined with the corresponding raw embryo images to train an ilastik object classification pipeline (ObjectClassifier.ilp; Barry, 2022). In such a pipeline, ilastik calculates various morphological features based on the connected components in the masks and a series of intensity features drawn from the pixel values in the raw images within the regions delineated by the masks (see https://www.ilastik.org/documentation/objects/objects for further information). We manually labelled 4.5 and 17.5 hpf embryos in our training data (Table 1) and trained the ilastik object classifier based on these annotations (Figure 2C). All training labels and pixel features used to train the object classifier can be viewed in the ilastik object classifier project file (ObjectClassifier.ilp – Barry, 2022).

The trained object classifier, when challenged with new test data, outputs two values; one gives the probability that the embryo is 4.5 hpf ($p_{4.5}$), the other that the embryo is 17.5 hpf ($p_{17.5}$). The sum of these two probabilities is always 1.0. Given the probability of a given test embryo being 17.5 hpf, the predicted hpf is calculated as follows:

$$P_{hpf} = 4.5 + p_{17.5} \times (17.5 - 4.5)$$

Comparison with manual (human) staging

To enable comparisons to be made between the accuracy of the classifier and manual (human) staging of zebrafish embryos, three individuals were asked to stage WT zebrafish embryos in 42 still images, randomly selected from the time-lapse movies that the classifier had previously analyzed. All were provided with the standard staging guide of Kimmel *et al.* (1995). These data were then compared with the predicted hpf generated by our classifier, for the same 42 images. The range of error was then calculated as the difference between the maximum and minimum error.

Statistical analysis

LOESS (locally established scatter plot smoothing) was used to generate the line plots in Figure 3 showing the

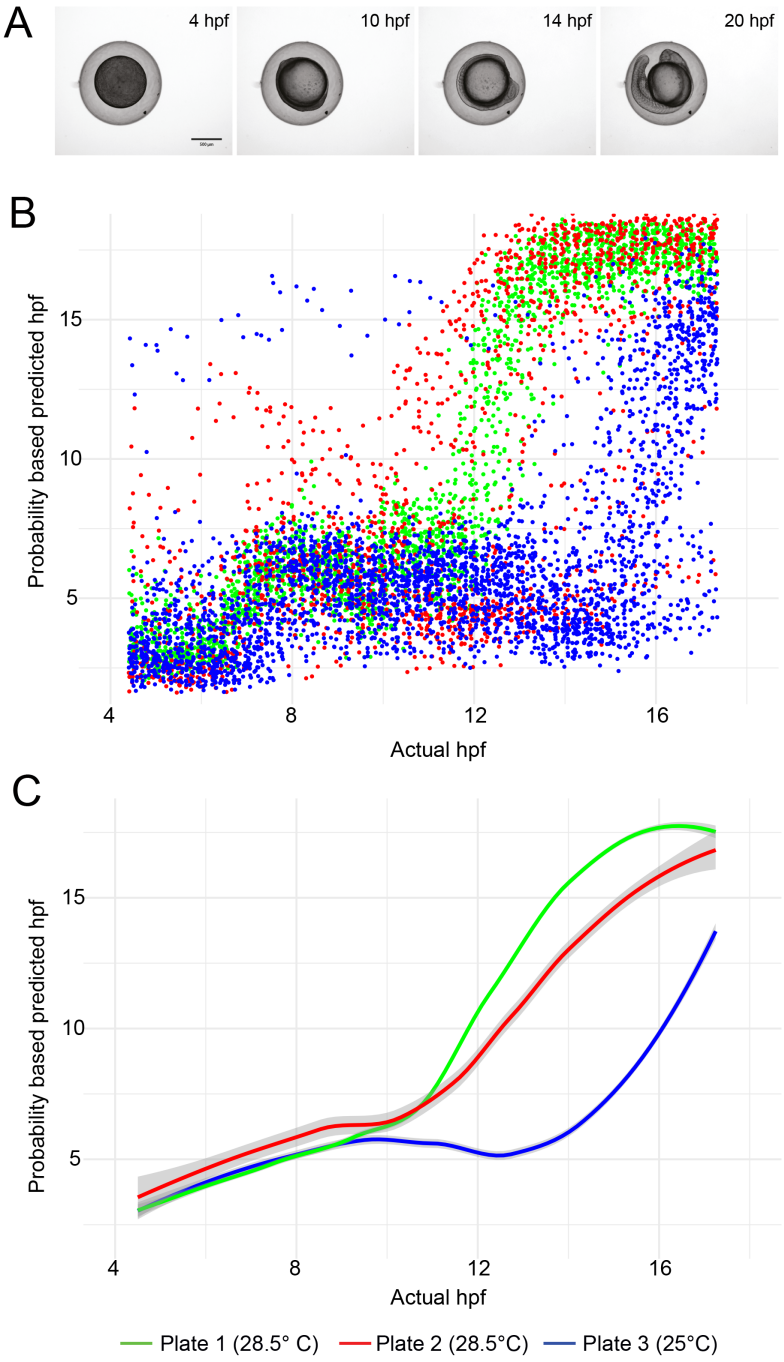

Figure 3. Temporal development profile of zebrafish embryos. (**A**) Examples of still images from time-lapse movies used to train the object classification algorithm. (**B**) Scatter plot showing the hpf predicted by the ilastik object classifier versus the actual hpf for each embryo image. Each dataset contains approximately 5,000 data points (96 wells per experiment, imaged every 15 minutes for 13 hours). (**C**) Line fit of the data in (**B**) using locally estimated scatterplot smoothing (LOESS). The grey region around each line shows the 95% confidence interval.

temporal development profile of embryos maintained at 28.5°C compared to 25°C. 95% confidence intervals (CI) calculated in R are displayed. All R scripts are available in the software availability section (Barry, 2022).

Results

Development of embryos is slower when they are maintained at 25°C than at 28.5°C. As proof of principle, we challenged our trained machine learning model with previously unseen test

data, consisting of embryos incubated at different temperatures (Table 1). The classifier was able to clearly differentiate between embryos maintained at 28.5°C and those maintained at 25°C (Figure 3). For the test data derived from plates incubated at 28.5°C, a greater spread in datapoints is evident for data derived from one plate versus the other (Figure 3C). The standard error of the mean predicted hpf, averaged over all timepoints, was 0.16 hpf for Plate 1 versus 0.60 hpf for Plate 2. This may be because slightly more training data was drawn from plate 1. But it should also be considered that the test data was not subjected to any quality control, so it is possible that more embryos on plate 2 died or drifted out of the field of view than on plate 1.

Having shown that our classifier can make meaningful relative comparisons between the developmental speed of embryos incubated at different temperatures, we next asked how accurate our classifier is at determining the actual developmental stage of specific embryos. More specifically, could our classifier identify the actual developmental stage, in hpf, of the embryos imaged? Importantly, our classifier was trained to give the probability that a given embryo belongs to one of two classes (4.5 hpf or 17.5 hpf) with the intention of detecting developmental delays. However, we were interested to ask how it compared with manual (human) staging. Crucially, images captured and assessed by the classifier are not controlled in relation to embryo orientation. In practice, this means that in some images, the embryonic stage can be clearly seen and identified (*e.g.* by counting the somites). In other images however, it is much more difficult, because the embryo is in an orientation in which key morphological features cannot be distinguished, or indeed the image itself is blurred. Therefore, unsurprisingly, considerable variation was observed in the manual staging between three individuals — for approximately 60% of timepoints, the maximum difference between any two human estimates was two hours or greater (Figure 4a). The random orientation of the embryos imaged in our system frequently did not permit the counting of somites, nor clear visualization of a specific developmental landmark such at the otic vesicle. Our data therefore demonstrate the importance of having multiple people stage the same samples to reach a consensus where the images are obtained in an automated fashion. When the same images were analyzed by our classifier, even given the training limitations described above, it was able to estimate the specific hpf of embryos with a similar success rate to manual (human) staging (Figure 4b). The errors produced by the classifier (0.0 ± 0.804; mean ± 95% confidence interval) are comparable to the errors made by humans (0.0 ± 0.239). But given the imbalance in the number of data points in each population (42 versus 126), making any kind of rigorous statistical analysis is difficult. What these data do show is that despite the classifier not having been trained to identify discrete developmental timepoints, it still fares well compared to humans, and is capable of analyzing images far more rapidly.

Discussion

Machine learning approaches in developmental biology are not new and have become increasingly popular as our ability to generate large amounts of data has evolved (Jones, 2019; Tarca *et al.,* 2007). The generation of 'big data', particularly from 'omics' technologies, has necessitated ever more sophisticated analysis tools, and the collection of live-imaging data is no different. Our ability to obtain thousands of images of hundreds of live biological samples means there is an increasing need for more automated methods of analysis. Moreover, automated data analysis helps to minimize the proclivity for human error and unconscious bias, a particular problem in our perception of images (Jost & Waters, 2019).

In this work, we have developed a new machine learning classifier for quantification of temporal development of the zebrafish, a commonly used model organism, particularly in the field of developmental biology. Until now, identification of developmental delay in mutant or treated zebrafish lines has only been possible by human observation and manual staging; a methodology inherently restricted in terms of numbers of embryos that can be observed over a given time-course. Moreover, as our data have shown, there is an intrinsic subjectivity in manual staging that may render results hard to reproduce, for example, over half of the images assessed by humans in our study showed at least a 2 hpf variability between individuals, and in some cases, considerably more. Our classifier at present uses relatively simple brightfield images, and therefore accuracy could be improved by incorporating gene expression data using fluorescent transgenic reporter lines. The expression profiles of numerous key genes during zebrafish development are clearly defined both spatially and temporally, so it follows that we could improve the accuracy of our classifier by the addition of gene expression data of selected genes. These could include for example, *tbxta* (*brachyury, T, no tail*) (germ-ring from ~5 hpf, notochord from ~10 hpf) (Schulte-Merker *et al.,* 1992; Schulte-Merker *et al.,* 1994), *sox10* (neural crest from ~10 hpf) (Dutton *et al.,* 2001) and *myod* (presumptive mesoderm from ~5hpf, somites from ~10hpf) (Weinberg *et al.,* 1996). In a similar way, Pond *et al.* (2021) incorporated gene expression data from confocal microscopy images to enhance their algorithm training, using fluorescent *in situ* hybridisation techniques to profile gene expression. Although it precludes the use of fluorescent *in situ* hybridisation techniques to profile gene expression, a key advantage of our classifier system is the use of live imaging, whereby the course of developmental progression is captured, as opposed to a series of fixed images of different embryos. Moreover, another key strength of our classifier is its ability to accurately quantify temporal development from images of embryos in random orientations with absolutely no image quality control. We envisage that our classifier will be a particularly useful tool in studies where accurate quantification of developmental delay is imperative, such as for developmental toxicity testing of drugs and toxicants (Dasgupta *et al.*, 2020; Nishimura *et al.*, 2016; Song *et al.*, 2021).

Limitations of this study include testing one 96-well plate at a given time meaning the 28.5°C and 25°C experiments were conducted on different days, and the lack of testing using a genetically perturbed/drug-treated zebrafish line. Additionally, in cases where only a small sample of embryos is to be

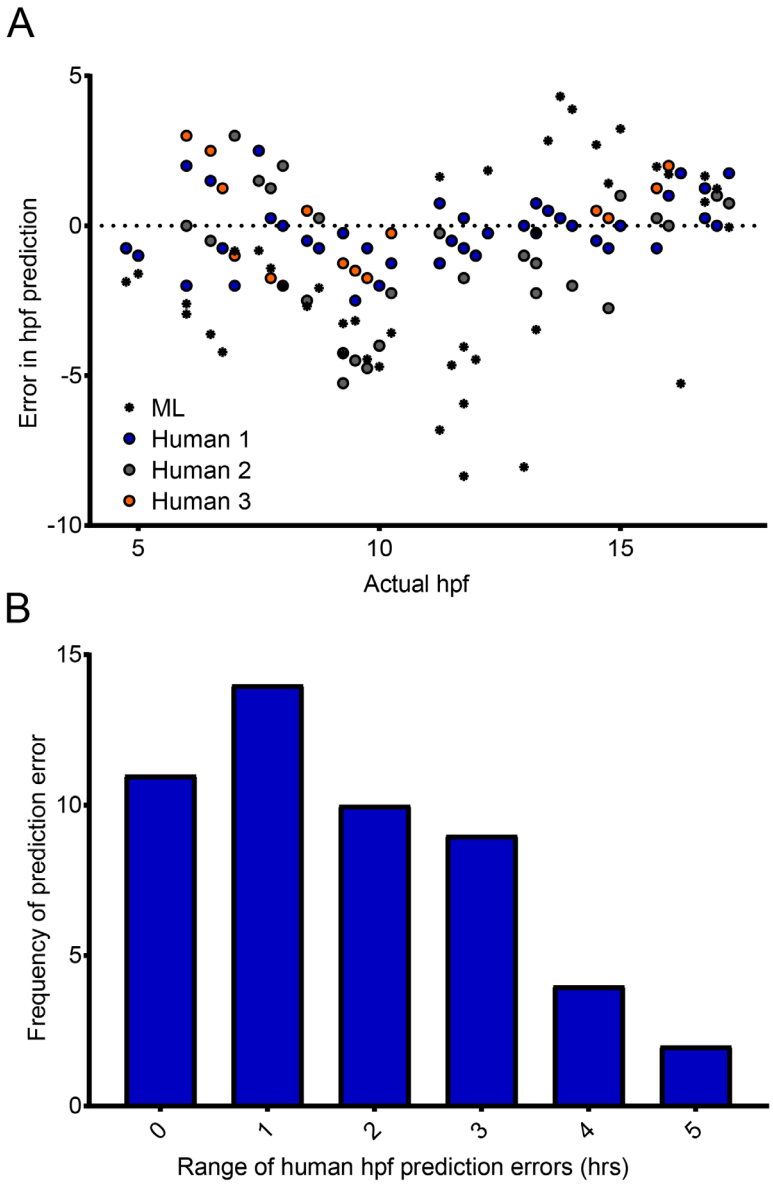

Figure 4. Comparison of manual and automated predictions of developmental stage. (**A**) Machine Learning (ML) classifier- and human-predicted hpf for 42 images of zebrafish embryos ranging from 4 to 17.5 hpf – each dot represents a single prediction for a single image. (**B**) Distribution of the range of human-predicted hpfs in (**A**), where the range represents the difference between the maximum and minimum error at each timepoint in (**A**).

tested (e.g. <50), it may be less labor-intensive to monitor development manually, albeit with appropriate controls to reduce subjectivity.

Other studies have used 3D imaging and OPT to enhance the ability of machine learning approaches to accurately stage and identify morphological features (Guglielmi *et al.*, 2021; Pond *et al.*, 2021). Our classifier at present uses relatively simple 2D images, taken using a standard wide-field microscope, and its simplicity in both image acquisition and analysis makes it accessible to a wide audience. Similarly, although our classifier has been trained using WT embryos, the same pipeline could be used to analyze zebrafish embryos with aberrant morphologies, e.g. the *no tail* (*Brachyury*) mutant (Halpern *et al.*, 1993; Schulte-Merker *et al.*, 1994), providing the algorithm is retrained on a subset of the given mutant embryos.

Finally, while we implemented our classifier using "conventional" image analysis tools such as ilastik and FIJI, the use of deep learning in biological research is becoming ever more popular (Hallou *et al.*, 2021). However, the application of deep learning for staging zebrafish embryos would require

optimization of neural network architecture, along with a substantially larger volume of training data — this requires considerable computational time and resources.

Conclusion

The developing zebrafish embryo is used in many different types of studies and accurate staging is essential. When comparing an experimental group of embryos with a control group, ensuring the embryos have reached the same developmental stage allows for meaningful comparisons to be made. Moreover, identification of a developmental delay in an experimental group is itself an important phenotypic observation. Our machine learning based classifier enables the unbiased assessment of thousands of images, across hundreds of embryos, with minimal time commitment. We anticipate that our classifier will be a useful tool for the zebrafish community to determine whether experimental animals (mutants, morphants, drug treated embryos) develop at the same rate as WT counterparts.

Data availability

Underlying data

All image data generated in this study is available to download from the BioImage Archive (accession number S-BIAD531) https://www.ebi.ac.uk/biostudies/bioimages/studies/S-BIAD531

Reporting guidelines

Zenodo: ARRIVE 2.0 checklist for "Automated staging of zebrafish embryos using machine learning" https://doi.org/10.5281/zenodo.7198533 (Barry, 2022a)

Software availability

Source code available from: https://github.com/djpbarry/fish-quant

Archived source code available from: https://doi.org/10.5281/zenodo.7189408 (Barry, 2022).

License: GNU v3.0

Acknowledgements

We thank Mollie Millington, Sarah Wheatley and all of the Francis Crick Institute Aquatics team for their invaluable help. We thank the Francis Crick Advanced Light Microscopy Science Technology Platform (STP) as well as the Scientific Computing STP and the Crick Research Illustration and Graphics Team. We also thank members of the Smith lab for helpful discussion and manuscript feedback. We thank Marvin Cortez of Princeton University for blind staging of zebrafish images. For the purpose of Open Access, the senior author has applied a CC BY public copyright license to any Author Accepted Manuscript version arising from this submission.

References

Aksakal FI, Sisman T: **Developmental toxicity induced by $Cu(OH)_2$ nanopesticide in zebrafish embryos.** *Environ Toxicol.* 2020; **35**(12): 1289–1298.
PubMed Abstract | Publisher Full Text

Akthar IST, Pichiah PBT, Arunachalam S, *et al.*: **Adriamycin inhibits embryonic development in zebrafish through downregulation of *Kruppel-like factor4*.** *J Biochem Mol Toxicol.* 2019; **33**: e22235.
PubMed Abstract | Publisher Full Text

Barry D: **ARRIVE 2.0 checklist for "Automated staging of zebrafish embryos using machine learning".** *[Reporting guidelines] Zenodo.* 2022a.
http://www.doi.org/10.5281/zenodo.7198533

Barry D: **djpbarry/fish-quant: Published Archive (v1.0.0).** *Zenodo.* [Code]. 2022.
http://www.doi.org/10.5281/zenodo.7189408

Berg S, Kutra D, Kroeger T, *et al.*: **ilastik: interactive machine learning for (bio)image analysis.** *Nat Methods.* 2019; **16**(12): 1226–1232.
PubMed Abstract | Publisher Full Text

Byrnes J, Ganetzky R, Lightfoot R, *et al.*: **Pharmacologic modeling of primary mitochondrial respiratory chain dysfunction in zebrafish.** *Neurochem Int.* 2018; **117**: 23–34.
PubMed Abstract | Publisher Full Text | Free Full Text

Choi TY, Choi TI, Lee YR, *et al.*: **Zebrafish as an animal model for biomedical research.** *Exp Mol Med.* 2021; **53**(3): 310–317.
PubMed Abstract | Publisher Full Text | Free Full Text

Dasgupta S, Reddam A, Liu Z, *et al.*: **High-content screening in zebrafish identifies perfluorooctanesulfonamide as a potent developmental toxicant.** *Environ Pollut.* 2020; **256**: 113550.
PubMed Abstract | Publisher Full Text | Free Full Text

Dutton KA, Pauliny A, Lopes SS, *et al.*: **Zebrafish *colourless* encodes *sox10* and specifies non-ectomesenchymal neural crest fates.** *Development.* 2001; **128**(21): 4113–4125.
PubMed Abstract | Publisher Full Text

Edelstein AD, Tsuchida MA, Amodaj N, *et al.*: **Advanced methods of microscope control using μManager software.** *J Biol Methods.* 2014; **1**(2): e10.
PubMed Abstract | Publisher Full Text | Free Full Text

Eisen JS: **Zebrafish Make a Big Splash.** *Cell.* 1996; **87**(6): 969–977.
PubMed Abstract | Publisher Full Text

Elabd S, Jabeen NA, Gerber V, *et al.*: **Delay in development and behavioural abnormalities in the absence of p53 in zebrafish.** *PLoS One.* 2019; **14**(7): e0220069.
PubMed Abstract | Publisher Full Text | Free Full Text

Farooq M, Sharma A, Almarhoon Z, *et al.*: **Design and synthesis of mono-and di-pyrazolyl-*s*-triazine derivatives, their anticancer profile in human cancer cell lines, and *in vivo* toxicity in zebrafish embryos.** *Bioorg Chem.* 2019; **87**: 457–464.
PubMed Abstract | Publisher Full Text

Flinn L, Bretaud S, Lo C, *et al.*: **Zebrafish as a new animal model for movement disorders.** *J Neurochem.* 2008; **106**(5): 1991–1997.
PubMed Abstract | Publisher Full Text

Ge S, Li J, Huang D, *et al.*: **Strong static magnetic field delayed the early development of zebrafish.** *Open Biol.* 2019; **9**(10): 190137.
PubMed Abstract | Publisher Full Text | Free Full Text

Giraldez AJ, Cinalli RM, Glasner ME, *et al.*: **MicroRNAs Regulate Brain Morphogenesis in Zebrafish.** *Science.* 2005; **308**(5723): 833–8.
PubMed Abstract | Publisher Full Text

Gomes MC, Mostowy S: **The Case for Modeling Human Infection in Zebrafish.** *Trends Microbiol.* 2020; **28**(1): 10–18.
PubMed Abstract | Publisher Full Text

Greener JG, Kandathil SM, Moffat L, *et al.*: **A guide to machine learning for biologists.** *Nat Rev Mol Cell Biol.* 2022; **23**(1): 40–55.
PubMed Abstract | Publisher Full Text

Guglielmi L, Heliot C, Kumar S, *et al.*: **Smad4 controls signaling robustness and morphogenesis by differentially contributing to the Nodal and BMP pathways.** *Nat Commun.* 2021; **12**(1): 6374.
PubMed Abstract | Publisher Full Text | Free Full Text

Hallou A, Yevick HG, Dumitrascu B, *et al.*: **Deep learning for bioimage analysis in developmental biology.** *Development.* 2021; **148**(18): dev199616.
PubMed Abstract | Publisher Full Text | Free Full Text

Halpern ME, Ho RK, Walker C, *et al.*: **Induction of muscle pioneers and floor plate is distinguished by the zebrafish *no tail* mutation.** *Cell.* 1993; **75**(1): 99–111.
PubMed Abstract | Publisher Full Text

Hung FC, Cheng YC, Sun NK, *et al.*: **Identification and functional characterization of zebrafish Gas7 gene in early development.** *J Neurosci Res.* 2013; **91**(1): 51–61.
PubMed Abstract | Publisher Full Text

Ishaq O, Sadanandan SK, Wählby C: **Deep Fish.** *SLAS Discov.* 2017; **22**(1): 102–107.
PubMed Abstract | Publisher Full Text

Jeanray N, Marée R, Pruvot B, *et al.*: **Phenotype Classification of Zebrafish Embryos by Supervised Learning.** *PLoS One.* 2015; **10**(1): e0116989.
PubMed Abstract | Publisher Full Text | Free Full Text

Jia S, Wu X, Wu Y, *et al.*: **Multiple Developmental Defects in *sox11a* Mutant Zebrafish with Features of Coffin-Siris Syndrome.** *Int J Biol Sci.* 2020; **16**(15): 3039–3049.
PubMed Abstract | Publisher Full Text | Free Full Text

Jones DT: **Setting the standards for machine learning in biology.** *Nat Rev Mol Cell Biol.* 2019; **20**(11): 659–660.
PubMed Abstract | Publisher Full Text

Jost APT, Waters JC: **Designing a rigorous microscopy experiment: Validating methods and avoiding bias.** *J Cell Biol.* 2019; **218**(5): 1452–1466.
PubMed Abstract | Publisher Full Text | Free Full Text

Kamei H, Yoneyama Y, Hakuno F, *et al.*: **Catch-Up Growth in Zebrafish Embryo Requires Neural Crest Cells Sustained by Irs1 Signaling.** *Endocrinology.* 2018; **159**(4): 1547–1560.
PubMed Abstract | Publisher Full Text

Kemmler CL, Riemslagh FW, Moran HR, *et al.*: **From Stripes to a Beating Heart: Early Cardiac Development in Zebrafish.** *J Cardiovasc Dev Dis.* 2021; **8**(12): 17.
PubMed Abstract | Publisher Full Text | Free Full Text

Kimmel CB, Ballard WW, Kimmel SR, *et al.*: **Stages of embryonic development of the zebrafish.** *Dev Dyn.* 1995; **203**(3): 253–310.
PubMed Abstract | Publisher Full Text

Li SZ, Liu W, Li Z, *et al.*: ***greb1* regulates convergent extension movement and pituitary development in zebrafish.** *Gene.* 2017; **627**: 176–187.
PubMed Abstract | Publisher Full Text

Li W, Yuan M, Wu Y, *et al.*: **Bixafen exposure induces developmental toxicity in zebrafish (*Danio rerio*) embryos.** *Environ Res.* 2020; **189**: 109923.
PubMed Abstract | Publisher Full Text

Liu K, Petree C, Requena T, *et al.*: **Expanding the CRISPR Toolbox in Zebrafish for Studying Development and Disease.** *Front Cell Dev Biol.* 2019; **7**: 13.
PubMed Abstract | Publisher Full Text | Free Full Text

Mannucci I, Dang NDP, Huber H, *et al.*: **Genotype-phenotype correlations and novel molecular insights into the *DHX30*-associated neurodevelopmental disorders.** *Genome Med.* 2021; **13**(1): 90.
PubMed Abstract | Publisher Full Text | Free Full Text

Martinez CS, Feas DA, Siri M, *et al.*: ***In vivo* study of teratogenic and anticonvulsant effects of antiepileptics drugs in zebrafish embryo and larvae.** *Neurotoxicol Teratol.* 2018; **66**: 17–24.
PubMed Abstract | Publisher Full Text

Meijering E, Carpenter AE, Peng H, *et al.*: **Imagining the future of bioimage analysis.** *Nat Biotechnol.* 2016; **34**(12): 1250–1255.
PubMed Abstract | Publisher Full Text

Mesquita B, Lopes I, Silva S, *et al.*: **Gold nanorods induce early embryonic developmental delay and lethality in zebrafish (*Danio rerio*).** *J Toxicol Environ Health A.* 2017; **80**(13–15): 672–687.
PubMed Abstract | Publisher Full Text

Nishimura Y, Inoue A, Sasagawa S, *et al.*: **Using zebrafish in systems toxicology for developmental toxicity testing.** *Congenit Anom (Kyoto).* 2016; **56**(1): 18–27.
PubMed Abstract | Publisher Full Text

Nüsslein-Volhard C: **The zebrafish issue of *Development*.** *Development.* 2012; **139**(22): 4099–103.
PubMed Abstract | Publisher Full Text

Parichy DM, Elizondo MR, Mills MG, *et al.*: **Normal table of postembryonic zebrafish development: Staging by externally visible anatomy of the living fish.** *Dev Dyn.* 2009; **238**(12): 2975–3015.
PubMed Abstract | Publisher Full Text | Free Full Text

Pond AJR, Hwang S, Verd B, *et al.*: **A deep learning approach for staging embryonic tissue isolates with small data.** *PLoS One.* 2021; **16**(1): e0244151.
PubMed Abstract | Publisher Full Text | Free Full Text

Schindelin J, Arganda-carreras I, Frise E, *et al.*: **Fiji: an open-source platform for biological-image analysis.** *Nat Methods.* 2012; **9**(7): 676–82.
PubMed Abstract | Publisher Full Text | Free Full Text

Schulte-Merker S, Ho RK, Herrmann BG, *et al.*: **The protein product of the zebrafish homologue of the mouse T gene is expressed in nuclei of the germ ring and the notochord of the early embryo.** *Development.* 1992; **116**(4): 1021–32.
PubMed Abstract | Publisher Full Text

Schulte-Merker S, Van Eeden FJ, Halpern ME, *et al.*: ***no tail* (*ntl*) is the zebrafish homologue of the mouse *T* (*Brachyury*) gene.** *Development.* 1994; **120**(4): 1009–15.
PubMed Abstract | Publisher Full Text

Shang S, Lin S, Cong F: **Zebrafish Larvae Phenotype Classification from Bright-field Microscopic Images Using a Two-Tier Deep-Learning Pipeline.** *Appl Sci.* 2020; **10**(4): 1247.
Publisher Full Text

Sidik A, Dixon G, Buckley DM, *et al.*: **Exposure to ethanol leads to midfacial hypoplasia in a zebrafish model of FASD via indirect interactions with the Shh pathway.** *BMC Biol.* 2021; **19**(1): 134.
PubMed Abstract | Publisher Full Text | Free Full Text

Singleman C, Holtzman NG: **Growth and maturation in the zebrafish, *Danio rerio*: a staging tool for teaching and research.** *Zebrafish.* 2014; **11**(4): 396–406.
PubMed Abstract | Publisher Full Text | Free Full Text

Song YS, Dai MZ, Zhu CX, *et al.*: **Validation, Optimization, and Application of the Zebrafish Developmental Toxicity Assay for Pharmaceuticals Under the ICH S5(R3) Guideline.** *Front Cell Dev Biol.* 2021; **9**: 721130.
PubMed Abstract | Publisher Full Text | Free Full Text

Suzuki H, Ishizaka T, Yanagi K, *et al.*: **Characterization of *biklf/klf17*-deficient zebrafish in posterior lateral line neuromast and hatching gland development.** *Sci Rep.* 2019; **9**(1): 13680.
PubMed Abstract | Publisher Full Text | Free Full Text

Tarca AL, Carey VJ, Chen XW, *et al.*: **Machine Learning and Its Applications to Biology.** *PLoS Comput Biol.* 2007; **3**(6): e116.
PubMed Abstract | Publisher Full Text | Free Full Text

Teixidó E, Kießling TR, Krupp E, *et al.*: **Automated Morphological Feature Assessment for Zebrafish Embryo Developmental Toxicity Screens.** *Toxicol Sci.* 2019; **167**(2): 438–449.
PubMed Abstract | Publisher Full Text | Free Full Text

Trikić MZ, Monk P, Roehl H, *et al.*: **Regulation of Zebrafish Hatching by Tetraspanin *cd63*.** *PLoS One.* 2011; **6**(5): e19683.
PubMed Abstract | Publisher Full Text | Free Full Text

Tshering G, Plengsuriyakarn T, Na-Bangchang K, *et al.*: **Embryotoxicity evaluation of atractylodin and β-eudesmol using the zebrafish model.** *Comp Biochem Physiol C Toxicol Pharmacol.* 2021; **239**: 108869.
PubMed Abstract | Publisher Full Text

Walpita CN, Crawford AD, Darras VM: **Combined antisense knockdown of type 1 and type 2 iodothyronine deiodinases disrupts embryonic development in zebrafish (*Danio rerio*).** *Gen Comp Endocrinol.* 2010; **166**(1): 134–141.
PubMed Abstract | Publisher Full Text

Weinberg ES, Allende ML, Kelly CS, *et al.*: **Developmental regulation of zebrafish MyoD in wild-type, no tail and spadetail embryos.** *Development.* 1996; **122**(1): 271–280.
PubMed Abstract | Publisher Full Text

Westerfield M: **The Zebrafish Book: A Guide for the Laboratory Use of Zebrafish (*Danio Rerio*).** University of Oregon Press, 2000.
Reference Source

Zanandrea R, Bonan CD, Campos MM: **Zebrafish as a model for inflammation and drug discovery.** *Drug Discov Today.* 2020; **25**(12): 2201–2211.
PubMed Abstract | Publisher Full Text

Zhang Q, Cheng J, Xin Q: **Effects of tetracycline on developmental toxicity and molecular responses in zebrafish (*Danio rerio*) embryos.** *Ecotoxicology.* 2015; **24**(4): 707–719.
PubMed Abstract | Publisher Full Text

Open Peer Review

Current Peer Review Status: ✔ ? ?

Version 3

Reviewer Report 21 August 2023

https://doi.org/10.21956/wellcomeopenres.21485.r64540

?

Christian Tischer
Centre for BioImage Analysis, European Molecular Biology Laboratory, Heidelberg, Germany
Arif Khan
EMBL Heidelberg, Heidelberg, Germany
Sebastian Gonzalez-Tirado
EMBL Heidelberg, Heidelberg, Germany

 # Overall assessment

In this manuscript the authors describe a computational method for staging zebrafish embryos.

Firstly, we would like to congratulate the authors for publishing all image data on the Bioimage Archive and putting all code onto a publicly accessible github repository.

However, in our opinion, the overall rationale does not scientifically adhere to the standards in the field. The authors use a linear regression to interpolate random forest probabilities, which are not meant to provide a linear response to the input data. It is not obvious whether the object features that are used as an input to the random forest should change linearly during zebrafish development. For example, maybe the “variance in intensity” (https://www.ilastik.org/documentation/objects/objectfeatures) would change only in later stages while it stays constant for some duration during early development, potentially resulting in a non-linear response. The random forest classifier is adding even more nonlinearities as each decision tree is composed of several thresholds on the image features and the vote of each tree is a majority vote of all its leaves. The ilastik probabilities summarize the votes of all trees. Thus, overall, the rationale of a linear interpolation of those, by design, highly non-linear probabilities is in our view scientifically questionable. We feel that this notion is supported by the data. In our view, the data points for plate 1 in Figure 3A could be well fit by a step function with three steps where the random forest output probabilities stay relatively constant from 4 to 6 hpf and from 6 to 13 hpf and from 13 to 20 hpf, consistent with the non-linear design of the random forest that is meant to create decisions rather than a linear response.

Along those lines, we were also wondering that why the authors chose to paint a schematic non-linear response in Figure 1E. Could you please comment?

We would like to offer two suggestions for how to modify the approach of the authors while still keeping a large part of the overall pipeline intact:

1. As far as we know, the measured features can be extracted from ilastik and it would be very interesting to plot all of them against hpf and see whether some of them change linearly. If yes, those linearly responding features could be used to train a linear regression model to achieve the aim of the publication.

2. As an alternative, the authors could decide to split the developmental time, e.g. into bins of 2 hours and then train ilastik to predict into which bin the embryo belongs. That is, for 20 hours one would need to train 10 classes. We hope that this should be an acceptable effort in terms of annotations. The output of the overall pipeline would then simply be the class that is predicted by ilastik with the highest probability.

Another general comment is that for all approaches it is critical to keep the imaging conditions identical in terms of illumination intensity and exposure time, to ensure that the algorithm trained on one data set can be applied to another one. It would be important to mention this in the manuscript. In addition, it should be considered whether the images could be normalized, e.g. using intensity percentiles, before being input to the algorithms. In fact, we think it would be essential to test whether such a normalization could improve the robustness of the results, especially when staging embryos that are imaged on different days and that were not part of the training data.

 # Manuscript

Page3- Text: *"The staging of ..... well-suited"*. A reference (citation) would be helpful.

Page3- Text: *"Jeanray et al. ..... Shang et al. (2022)"*. Hard to comprehend. Please rephrase.

Page3- Text: *" ... semi-automated segmentation and quantitation .... "*. Use quantification ?!

Page4- Text: *" ... using simple grey level threshold .... "*. What is meant by 'simple' here? What was the criterion for threshold selection? We think it is generally good to avoid qualifications such as "simple".

Page5- Table1: Could you please elaborate in the text why no data from plate 3 was used for training?!

Figure 2, A: How were these features for pixel classification selected? Can you please explain the rationale why these features were used and which filter sizes were selected? In case, the authors just used all features that are available in ilastik, we think this should be mentioned like this instead of listing them individually.

Figure 2, B: Can you please label the individual morphological operation steps?

Figure 2,- Caption B: What kind of thresholding and what morphological operations were used?

Figure 2, C: How were these features selected?

Page8- Text: *"..... image itself is blurred"*. Why are the images blurred? Could they be removed from the data set?

Page8- Text: *"..... Test data was not subjected to any quality control, ..."*. Why not?

Page8- Text: *"But given the imbalance ..... analysis is difficult"*. Why was this imbalance created in the first place?

Page8- Text: *" ..... analyzing images far more rapidly"*. Do you have any estimation on how fast in minutes/hours/....?

Page10- Text: *" When comparing ..... To be made"*. Not comprehensible. Can you please rephrase?

Page10- Text: *" ..... unbiased assessment ....."*. Replace with automated?

 # Software

In general, we feel that the software is missing documentation, especially to make the usage of it clear to non-computational users that would like to use it. For example, there is no information on how a user could start using the software. Information such as *"clone the repository, you will find the trained models inside this directory, etc"* is missing. Other basic information is also missing on describing how each command should be executed. For example, part of the first step of the pipeline is written as:

 /run_ilastik.sh ... --output_filename_format="/{nickname}_{result_type}.tiff" "input.tiff"

But there are no comments indicating that the user should replace the parts between angle brackets (< >) with their own input. Also, please comment about the importance of keeping the other brackets (such as in **{nickname}_{result_type}.tiff**) for the software to work properly.

The Fiji scripts are devoid of comments in the code describing the different steps, functions, etc.

We think that the repository should have a minimal example pointing to a dataset that could be used to run the software with the exact instructions to showcase how to run it!

Regarding other technical issues, we found that the data was missing some information in their metadata. For example, when opening the image properties on Fiji, we could not find the value of the frame intervals (it was zero and it would be great if it was giving the hpf).

Also, there should be information regarding the software versions used to run the analysis. We encourage this last point since we were not able to reproduce the workflow using the following time-series from the test-set:

FishDev_WT_01_1_MMStack_A8-Site_0.ome.tiff

First, we noticed that the first script of the tool only outputs the pixel predictions for two time-points, which was not clear if this is how the tool is supposed to work or whether this is a bug and one should expect the pixel predictions for all time points of the dataset. Also, we could not reproduce the whole workflow since we had an error after attempting to run the object classification part of the workflow (step 3, error input/output marked below).

Error output from step 3:

 /Applications/ilastik-1.4.0b27-OSX.app/Contents/ilastik-release/run_ilastik.sh --headless --project="./ObjectClassifier.ilp" --export_source="Object Probabilities" --output_format="multipage tiff" --output_filename_format="./{nickname}_{result_type}.tiff" --raw_data="/Users/segonzal/Downloads/Zebrafish_ML_Archive/test_data/FishDev_WT_01_1/FishDev_WT_01_1_M Site_0.ome.tif" --segmentation_image="/Users/segonzal/Downloads/Zebrafish_ML_Archive/test_data/FishDev_WT_01_1/FishDev Site_0.ome_Probabilities.tiff_Binary.tiff" --export_dtype="float32" --readonly="true" > error.txt

"/Applications/ilastik-1.4.0b27-OSX.app/Contents/ilastik-release/lib/python3.7/site-packages/lazyflow/operators/ioOperators/opTiffReader.py", line 58, in setupOutputs
   assert axes[last_C_pos] == "C"
AssertionError

Is the rationale for developing the new software tool clearly explained?
Yes

Is the description of the software tool technically sound?
No

Are sufficient details of the code, methods and analysis (if applicable) provided to allow replication of the software development and its use by others?
Partly

Is sufficient information provided to allow interpretation of the expected output datasets and any results generated using the tool?
Yes

Are the conclusions about the tool and its performance adequately supported by the findings presented in the article?
No

Competing Interests: No competing interests were disclosed.

Reviewer Expertise: Bioimage analysis

We confirm that we have read this submission and believe that we have an appropriate level of expertise to confirm that it is of an acceptable scientific standard, however we have significant reservations, as outlined above.

Version 2

Reviewer Report 31 March 2023

https://doi.org/10.21956/wellcomeopenres.21251.r55612

?

Amin Allalou
[1] Division of Visual Information and Interaction, Department of Information Technology, Image Analysis and Human-Computer Interaction, Uppsala University, Uppsala, Sweden
[2] Science for Life Laboratory BioImage Informatics Facility, Uppsala, Sweden

Responses (1)

AUTHOR RESPONSE16 MARCH 2023
Rebecca Jones
The Francis Crick Institute, 1 Midland Road, London, UK
Responses to Reviewer 2 (Amin Allalou) In this manuscript the author is describing a method for automatic detection of different development stages for the zebrafish. A robust method that can accurately classify a zebrafish into different development stages is highly relevant for many researchers working with zebrafish and hence the importance in developing methods like this is high.
We thank the reviewer for their positive comments.

The dataset used for the method is quite small and limited. It is good for proving a proof of the concept of the method but for a more general use of the method it needs to be further tested with embryos with different genetic backgrounds or treatment (as mentioned by the author in the discussion).
We agree with the reviewer that the data used in this study is limited, in the sense that the images used are all of wild type zebrafish embryos and all images were acquired on the same microscopy system. However, the aim of the manuscript was indeed to demonstrate a proof of concept (that developmental delay could be quantified in an automated manner). We did explore the possibility of testing our method on images of zebrafish whose development was perturbed either by some pharmacological treatment or genetic mutation, but, as far as we are aware, no such treatments or mutants are readily available. Therefore, including such additional data would constitute a substantial body of additional work and experimentation.
- I see your point even though more data would make the manuscript more appealing.

The author should improve the overall description of the method and the data used for training. It is sometimes difficult to understand what dataset is used for training and for evaluation. Same data cannot be used in training as in the evaluation of the method, it needs to be made clear to the reader that this is not the case.
We thank the reviewer for drawing our attention to this and agree that the wording of the manuscript could be improved in places. We have made clarifications regarding the nature of the

training and test data, detailed in the responses below.
- Thank you for clarifying this.

Page 4 Zebrafish husbandry and Live imaging Later in the paper the author is discussing 2 different datasets for 28.5 degree data (rep1 and rep2). Could the author in this section describe the difference between those to datasets.
We thank the reviewer for bringing this to our attention. We have revised the methods section and added a table (Table 1) to clarify exactly what data was used for training and what was used for testing. In total, we imaged three 96-well plates of embryos, two at 28.5C and one at 25.0C. A random selection of wells from the two 28.5C plates were used for training purposes, while the remainder of wells in those plates were reserved for testing. None of the training data was subsequently used for testing. The plate incubated at 25.0C was used for testing only. To be certain that none of the training data "leaked" into the test datasets, we modified the R code used to generate the plots in Figure 3 to explicitly rule out (by name) the inclusion of the wells used for training. New plots for Figure 3 have been generated with this revised R code which, as far as we can tell, are identical to those included with the original submission.
- Table 1 is a good addition that makes the data selection much more clear.

Page 4 “A total of 20 embryos were used for training, selected at random from the two”... Why only 20 embryos?
This was in fact a typo (the correct number, as per Table 1, is 24). But the reason for the small number is the requirement for manual annotation, which is time consuming. However, we view this as a positive - with relatively little training data, the classifier was still successful in distinguishing between two different populations of embryos.
- Thank you. I agree a small number of training samples that can generate a good method is positive.

How are these samples used in the training. Is the set divided into training, validation and test set? This is usually the procedure.
The manner in which the data is divided into training and test sets is outlined in Table 1.
- Thank you

“The resultant masks were then combined with the raw pixel data in an ilastik object classification pipeline (ObjectClassifier.ilp; Barry, 2022), whereby the embryos were manually classified as being either 4.5 or 17.5 hpf (Figure 2C)."Some more detail on the classification method would be useful for the reader.
We have now added some further details on the object classification workflow and included a link to the relevant section of the ilastik documentation (https://www.ilastik.org/documentation/objects/objects), should the reader require further information.
-Thank you

Page 5 “...for the same 50 images, and plotted accordingly”. Where are they plotted?
This corresponds to the data shown in Figure 4

Section “Results An accuracy measure of the correct classification would be good to have. How many samples are correctly classified into the correct class.
We agree with the reviewer that such a metric would be useful. However, while our machine

learning model is trained on the basis of placing objects into one of two different classes (4.5 or 17.5 hpf embryos), the result is essentially a regression model, as we are testing this classifier on embryos at various stages of development between 4.5 and 17.5 hpf (inclusive). So, while we could certainly include a measure of how successful our model would be in discriminating between 4.5 and 17.5 hpf embryos, this would tell us little about how successful the model is at discriminating between different rates of development, which was our primary goal.
- Ok, thank you for the clarification.

Development of embryos is slower when they are maintained at 25°C than at 28.5°C. As proof of principle, following training on 14 embryos and subsequent testing on two 96-well plates of WT embryos at 28.5°C as ...” What was trained here? The author should be more clear on what was trained here. Is anything from previous training used here (where 20 embryos were used) or is it a completely new training? Was the same classes used in the training as previous? If so why was it retrained?
We agree with the reviewer that this section was poorly worded and likely to cause confusion. We have therefore revised this text to make clear that no new training data has been introduced at this point - the results section refers to the same datasets referred to in the methods (and now listed in Table 1).
- Great!

If all training was done on one batch from 28.5°C it should not be used as a comparison in the evaluation where the same batch is compared to data from a new batch. This is not a valid comparison.
We agree with the reviewer on this point and have made changes to the text (and added Table 1) to clarify what was used as training data and what was used for testing - none of the data shown in Figure 3 was used in training.
-OK!

“our classifier was trained to give the probability that a given embryo belongs to one of two classes (4.5 hpf or 17.5 hpf) with the intention of detecting developmental delays – it was not trained to predict the actual hpf of a given embryo.” It is not really clear to the reader how the author is estimating the predicted hpf. The classifier seems to be trained on only two classes 4.5hpf and 17.5hpf. The output from the classification is a probability of belonging to one of each class. From these values how is the probability based predicted hpf in Figure 3 calculated? This should be clarified.
We thank the reviewer for drawing our attention to this omission. We have now added text to the relevant methods section explaining how the probability values output by the object classifier are converted into predicted hpf values.
-Ok, thanks for the clarification. However, it is a little bit confusing when you state “it was not trained to predict the actual hpf of a given embryo”, but you provide a regression where you try to do exactly this and also plot results for this. And at the same time you are not providing any classification accuracy. You are merely using a 2 class classifier and depending on how similar they are to the different classes you estimate the hpf. Figure 3B should be a straight line if the hpf prediction was good. Now it looks more like they are closer to one class and around 12 hpf they just shift from one class to the other. And for the 25 degree this shift happens around 16hpf. This to me does not seem like a good prediction of the hpf. However, this still shows the delay in development that the paper is aiming for but the importance and need of the regression part is not clear to me.

“hpf of embryos with a similar success rate to manual (human) staging (Figure 4b). The errors produced by the classifier (0.0 +-} 0.804; mean +-} 95% confidence interval) are comparable to the errors made by humans (0.0 +-} 0.239).” Is both rep 1 and rep 2 data used in this calculation? Are any of these batches used in the training? Please, clarify in the text.
A random selection of images from both plate 1 and plate 2 were used to generate this data. We have now noticed that, although none of the images used in training were included in this data, some images from the same wells as those used for training were included – these have now been removed and the figure revised. While there are now slightly fewer data points (44), the overall conclusion remains unchanged.
-OK!

“For example, at an actual hpf of 16.0, our classifier gave predicted hpfs of 16.... embryos and embryos maintained at 25°C, respectively.” This only provides info on that the development on the 25°C is slower, as it should be. But how accurate is the 10.21 ± 0.44 really? How much slower should an embryo at 25°C be? Did the author do some manual staging to compare this number with manual estimation?
We agree with the reviewer that this statement of accuracy was of limited value and have now removed it from the text. Estimations of accuracy are now limited to the data comparing our machine learning classifier with manual staging as shown in Figure 4.
-OK

Figure 1E What data is used here? It looks much cleaner than the ones shown in Figure 3? If this is just an artificial plot than are all three colors needed?
We have now updated the panel in Figure 1E to make it clear that this plot is not real data and is intended for illustrative purposes only.
- Ok, great!

Figure 3 In Figure 3b the spread of the data points for red (WT rep 2) seems much greater than for green (WT rep 1) (this can also be seen in the confidence interval in Fig 3b). It is not clear, but if the data for “rep 1” is coming from the same dataset as was used for the training, then this pattern is an indication that the method does not generalise well to a new batch. Could the author clarify where the data is coming from and provide a plot with the data separated? In addition, some quantification of the spread of the data in rep1 vs rep 2 would be of interest to the reader.
All of the data presented in Figure 3 was generated on test data only, previously unseen by the classifier. With regard to the differences in data spread, we have now added some discussion of this in the main text. Slightly more training data was drawn from plate 1 than plate 2, which may explain the slightly greater spread of data points for plate 2. It should also be noted that no quality control was applied to the test data (as this is laborious and time-consuming), so images of dead embryos and/or embryos that have drifted to the edge (our outside) of the field of view will have been included - it's possible that plate 2 produced a greater quantity of these lower quality images than plate 1. But the standard error of the mean for both plates is still less than 1.0 hpf.
- One problem I see with the division of the data is that in your true test set (data coming from an isolated batch) you only have data that should show a difference from your data used in the training. No real isolated control data. To make the results more clear an isolated test set for 28 degrees and 25 degrees would be a much stronger evidence of the method performance. What if the method is doing some overfit to the two training batches, how do you know the performance on the test set is accurate when you don’t have a control

group (28 degree)?

(3b) From the graph there seems to be a cluster of blue (25 degree) with actual hpf 4-10 being classified as approx 15 hpf. A comment on this pattern would be useful to the reader since the error is quite large.
The majority of these datapoints correspond to the same two wells on that particular plate and it appears that segmentation error is to blame for the misclassification, as illustrated in the example below – in both cases, the embryos in the early stages of development were consistently over-segmented, which likely resulted in their misclassification.
-ok, If this is the case then an additional plot with those wells removed (or marked in some way in your current plot) would clearly show the reader that the errors are coming from those wells.

Figure 4. "Machine Learning (ML) classifier- and human-predicted hpf for 50 images of zebrafish"
Are the data taken from the same batch as was used for training? No, as explained above, we have now made absolutely sure that none of the random images used for the human-ML comparison were drawn from the training data.
- Ok!

Is the rationale for developing the new software tool clearly explained?
No

Is the description of the software tool technically sound?
No

Are sufficient details of the code, methods and analysis (if applicable) provided to allow replication of the software development and its use by others?
No

Is sufficient information provided to allow interpretation of the expected output datasets and any results generated using the tool?
No

Are the conclusions about the tool and its performance adequately supported by the findings presented in the article?
No

Competing Interests: No competing interests were disclosed.

Reviewer Expertise: Quantitative Microscopy, Image analysis, Machine learning, zebrafish

I confirm that I have read this submission and believe that I have an appropriate level of expertise to confirm that it is of an acceptable scientific standard, however I have significant reservations, as outlined above.

Author Response 07 Apr 2023

Rebecca Jones

Ok, thanks for the clarification.

However, it is a little bit confusing when you state "*it was not trained to predict the actual hpf of a given embryo*", but you provide a regression where you try to do exactly this and also plot results for this. And at the same time you are not providing any classification accuracy. You are merely using a 2 class classifier and depending on how similar they are to the different classes you estimate the hpf. Figure 3B should be a straight line if the hpf prediction was good. Now it looks more like they are closer to one class and around 12 hpf they just shift from one class to the other. And for the 25 degree this shift happens around 16hpf. This to me does not seem like a good prediction of the hpf. However, this still shows the delay in development that the paper is aiming for but the importance and need of the regression part is not clear to me.

We agree with the reviewer's point here that, ideally, a plot of predicted hpf versus actual hpf should result in a straight line. In our case, we have curves that could more accurately be described as s-curves, due to the nature of the model we trained, as outlined above by the reviewer. While such curves are certainly not ideal for accurately predicting the hpf of individual embryos, they are completely adequate for the identification of developmental delay, which was our primary goal.

It is certainly possible that we could improve our model by including additional training classes (perhaps resulting in a predicted versus actual hpf plot more closely resembling a straight line), but as previously stated, this would require a substantial volume of additional work to manually annotate new training data. We did actually consider presenting our data in a slightly different manner, which would have involved labelling the y axis in Figure 3B as "Probability of embryo being 17.5 hpf", but felt that this was perhaps a little abstract and may be difficult for biologists to interpret. But we also felt that "predicted hpf" wasn't quite accurate, for the reasons the reviewer has outlined, which is why we settled on (the slightly cumbersome!) "probability based predicted hpf".

We also agree that the statement "*it was not trained to predict the actual hpf of a given embryo* " is confusing and have removed this statement from the text.

One problem I see with the division of the data is that in your true test set (data coming from an isolated batch) you only have data that should show a difference from your data used in the training. No real isolated control data. To make the results more clear an isolated test set for 28 degrees and 25 degrees would be a much stronger evidence of the method performance. What if the method is doing some overfit to the two training batches, how do you know the performance on the test set is accurate when you don't have a control group (28 degree)? We take the reviewers point about the 25C test data being completely isolated from the training data, whereas the 28.5C test data is drawn from the same plates as training data. However, in our opinion, over-fitting to training data was more likely had we used only a single 28.5C plate for training. We chose to use training data from both 28.5C plates for this reason.

ok, If this is the case then an additional plot with those wells removed (or marked in some way in your current plot) would clearly show the reader that the errors are coming from those wells. We must respectfully disagree with the reviewer on this point. We believe that one of the major strengths of our approach is the lack of manual curation we have performed on the data - the method is 100% automated. While we accept that this will sometimes result in errors for individual images or wells, for a sufficiently large population of embryos (such as a 96 well plate), these errors are not significant, as is evidenced by our ability to detect developmental delay between two different populations.

Removing erroneous datapoints based on arbitrary criteria would not only be an extremely time-consuming endeavour (requiring the manual assessment of thousands of images), we believe it would also significantly weaken our argument that no manual intervention is required in our pipeline.

Competing Interests: No competing interests were disclosed.

Reviewer Report 30 March 2023

https://doi.org/10.21956/wellcomeopenres.21251.r55613

Steffen Scholpp
Living Systems Institute, School of Biosciences, College of Life and Environmental Sciences, University of Exeter, Exeter, UK

The amendments are sufficient, and therefore, I suggest the manuscript for indexing.

Is the rationale for developing the new software tool clearly explained?
No

Is the description of the software tool technically sound?
No

Are sufficient details of the code, methods and analysis (if applicable) provided to allow replication of the software development and its use by others?
No

Is sufficient information provided to allow interpretation of the expected output datasets and any results generated using the tool?
No

Are the conclusions about the tool and its performance adequately supported by the

findings presented in the article?
No

Competing Interests: No competing interests were disclosed.

Reviewer Expertise: Morphogen trafficking in zebrafish

I confirm that I have read this submission and believe that I have an appropriate level of expertise to confirm that it is of an acceptable scientific standard.

Version 1

Reviewer Report 26 January 2023

https://doi.org/10.21956/wellcomeopenres.20298.r53876

?

Amin Allalou
[1] Division of Visual Information and Interaction, Department of Information Technology, Image Analysis and Human-Computer Interaction, Uppsala University, Uppsala, Sweden
[2] Science for Life Laboratory BioImage Informatics Facility, Uppsala, Sweden

In this manuscript the author is describing a method for automatic detection of different development stages for the zebrafish. A robust method that can accurately classify a zebrafish into different development stages is highly relevant for many researchers working with zebrafish and hence the importance in developing methods like this is high.

The dataset used for the method is quite small and limited. It is good for proving a proof of the concept of the method but for a more general use of the method it needs to be further tested with embryos with different genetic backgrounds or treatment (as mentioned by the author in the discussion).

The author should improve the overall description of the method and the data used for training. It is sometimes difficult to understand what dataset is used for training and for evaluation. Same data cannot be used in training as in the evaluation of the method, it needs to be made clear to the reader that this is not the case.

Page 4 Zebrafish husbandry and Live imaging

- Later in the paper the author is discussing 2 different datasets for 28.5 degree data (rep1 and rep2). Could the author in this section describe the difference between those to datasets.

Page 4

- *"A total of 20 embryos were used for training, selected at random from the two"…*
  - Why only 20 embryos?
  - How are these samples used in the training. Is the set divided into training, validation and test set? This is usually the procedure.
- *"The resultant masks were then combined with the raw pixel data in an ilastik object classification pipeline (ObjectClassifier.ilp; Barry, 2022), whereby the embryos were manually classified as being either 4.5 or 17.5 hpf (Figure 2C)."*
  - Some more detail on the classification method would be useful for the reader.

Page 5

- *"…for the same 50 images, and plotted accordingly".*
  - Where are they plotted?

Section "Results

- An accuracy measure of the correct classification would be good to have. How many samples are correctly classified into the correct class.
- *"Development of embryos is slower when they are maintained at 25°C than at 28.5°C. As proof of principle, following training on 14 embryos and subsequent testing on two 96-well plates of WT embryos at 28.5°C as …"*
  - What was trained here? The author should be more clear on what was trained here. Is anything from previous training used here (where 20 embryos were used) or is it a completely new training? Was the same classes used in the training as previous? If so why was it retrained?
  - If all training was done on one batch from 28.5°C it should not be used as a comparison in the evaluation where the same batch is compared to data from a new batch. This is not a valid comparison.
- *"our classifier was trained to give the probability that a given embryo belongs to one of two classes (4.5 hpf or 17.5 hpf) with the intention of detecting developmental delays – it was not trained to predict the actual hpf of a given embryo."*
  - It is not really clear to the reader how the author is estimating the predicted hpf. The classifier seems to be trained on only two classes 4.5hpf and 17.5hpf. The output from the classification is a probability of belonging to one of each class. From these values how is the probability based predicted hpf in Figure 3 calcualted? This should be clarified.
- *"hpf of embryos with a similar success rate to manual (human) staging (Figure 4b). The errors produced by the classifier (0.0 +-} 0.804; mean +-} 95% confidence interval) are comparable to the errors made by humans (0.0 +-} 0.239)."*
  - Is both rep 1 and rep 2 data used in this calculation? Are any of these batches used in

the training? Please, clarify in the text.

- "*For example, at an actual hpf of 16.0, our classifier gave predicted hpfs of 16.... embryos and embryos maintained at 25°C, respectively.*"

    - *This only provides info on that the development on the 25°C is slower, as it should be. But how accurate is the 10.21 ± 0.44 really? How much slower should an embryo at 25°C be? Did the author do some manual staging to compare this number with manual estimation?*

Figure 1E

- What data is used here? It looks much cleaner than the ones shown in Figure 3? If this is just an artificial plot than are all three colors needed?

Figure 3

- In Figure 3b the spread of the data points for red (WT rep 2) seems much greater than for green (WT rep 1) (this can also be seen in the confidence interval in Fig 3b). It is not clear, but if the data for *"rep 1"* is coming from the same dataset as was used for the training, then this pattern is an indication that the method does not generalise well to a new batch. Could the author clarify where the data is coming from and provide a plot with the data separated? In addition, some quantification of the spread of the data in rep1 vs rep 2 would be of interest to the reader.

- (3b) From the graph there seems to be a cluster of blue (25 degree) with actual hpf 4-10 being classified as approx 15 hpf. A comment on this pattern would be useful to the reader since the error is quite large.

Figure 4.

- *"Machine Learning (ML) classifier- and human-predicted hpf for 50 images of zebrafish"*

    - Are the data taken from the same batch as was used for training?

Is the rationale for developing the new software tool clearly explained?

Yes

Is the description of the software tool technically sound?

Yes

Are sufficient details of the code, methods and analysis (if applicable) provided to allow replication of the software development and its use by others?

Yes

Is sufficient information provided to allow interpretation of the expected output datasets and any results generated using the tool?

Partly

Are the conclusions about the tool and its performance adequately supported by the findings presented in the article?

Partly

Competing Interests: No competing interests were disclosed.

Reviewer Expertise: Quantitative Microscopy, Image analysis, Machine learning, zebrafish

I confirm that I have read this submission and believe that I have an appropriate level of expertise to confirm that it is of an acceptable scientific standard, however I have significant reservations, as outlined above.

Author Response 06 Mar 2023

Rebecca Jones

Responses to Reviewer 2 (Amin Allalou) *In this manuscript the author is describing a method for automatic detection of different development stages for the zebrafish. A robust method that can accurately classify a zebrafish into different development stages is highly relevant for many researchers working with zebrafish and hence the importance in developing methods like this is high.*
We thank the reviewer for their positive comments.

The dataset used for the method is quite small and limited. It is good for proving a proof of the concept of the method but for a more general use of the method it needs to be further tested with embryos with different genetic backgrounds or treatment (as mentioned by the author in the discussion).
We agree with the reviewer that the data used in this study is limited, in the sense that the images used are all of wild type zebrafish embryos and all images were acquired on the same microscopy system. However, the aim of the manuscript was indeed to demonstrate a proof of concept (that developmental delay could be quantified in an automated manner). We did explore the possibility of testing our method on images of zebrafish whose development was perturbed either by some pharmacological treatment or genetic mutation, but, as far as we are aware, no such treatments or mutants are readily available. Therefore, including such additional data would constitute a substantial body of additional work and experimentation.

The author should improve the overall description of the method and the data used for training. It is sometimes difficult to understand what dataset is used for training and for evaluation. Same data cannot be used in training as in the evaluation of the method, it needs to be made clear to the reader that this is not the case.
We thank the reviewer for drawing our attention to this and agree that the wording of the manuscript could be improved in places. We have made clarifications regarding the nature of the training and test data, detailed in the responses below.

Page 4 Zebrafish husbandry and Live imaging Later in the paper the author is discussing 2 different datasets for 28.5 degree data (rep1 and rep2). Could the author in this section describe the difference between those to datasets.
We thank the reviewer for bringing this to our attention. We have revised the methods section and added a table (Table 1) to clarify exactly what data was used for training and what was used for testing. In total, we imaged three 96-well plates of embryos, two at 28.5C and one at 25.0C. A random selection of wells from the two 28.5C plates were used for

training purposes, while the remainder of wells in those plates were reserved for testing. None of the training data was subsequently used for testing. The plate incubated at 25.0C was used for testing only. To be certain that none of the training data "leaked" into the test datasets, we modified the R code used to generate the plots in Figure 3 to explicitly rule out (by name) the inclusion of the wells used for training. New plots for Figure 3 have been generated with this revised R code which, as far as we can tell, are identical to those included with the original submission.

Page 4 "A total of 20 embryos were used for training, selected at random from the two"... Why only 20 embryos?
This was in fact a typo (the correct number, as per Table 1, is 24). But the reason for the small number is the requirement for manual annotation, which is time consuming. However, we view this as a positive - with relatively little training data, the classifier was still successful in distinguishing between two different populations of embryos.

How are these samples used in the training. Is the set divided into training, validation and test set? This is usually the procedure.
The manner in which the data is divided into training and test sets is outlined in Table 1.

"The resultant masks were then combined with the raw pixel data in an ilastik object classification pipeline (ObjectClassifier.ilp; Barry, 2022), whereby the embryos were manually classified as being either 4.5 or 17.5 hpf (Figure 2C)."Some more detail on the classification method would be useful for the reader.
We have now added some further details on the object classification workflow and included a link to the relevant section of the ilastik documentation (https://www.ilastik.org/documentation/objects/objects), should the reader require further information.

Page 5 "...for the same 50 images, and plotted accordingly". Where are they plotted?
This corresponds to the data shown in Figure 4

Section "Results An accuracy measure of the correct classification would be good to have. How many samples are correctly classified into the correct class.
We agree with the reviewer that such a metric would be useful. However, while our machine learning model is trained on the basis of placing objects into one of two different classes (4.5 or 17.5 hpf embryos), the result is essentially a regression model, as we are testing this classifier on embryos at various stages of development between 4.5 and 17.5 hpf (inclusive). So, while we could certainly include a measure of how successful our model would be in discriminating between 4.5 and 17.5 hpf embryos, this would tell us little about how successful the model is at discriminating between different rates of development, which was our primary goal.

Development of embryos is slower when they are maintained at 25°C than at 28.5°C. As proof of principle, following training on 14 embryos and subsequent testing on two 96-well plates of WT embryos at 28.5°C as ..." What was trained here? The author should be more clear on what was trained here. Is anything from previous training used here (where 20 embryos were used) or is it a completely new training? Was the same classes used in the training as previous? If so why was

it retrained?
We agree with the reviewer that this section was poorly worded and likely to cause confusion. We have therefore revised this text to make clear that no new training data has been introduced at this point - the results section refers to the same datasets referred to in the methods (and now listed in Table 1).

If all training was done on one batch from 28.5°C it should not be used as a comparison in the evaluation where the same batch is compared to data from a new batch. This is not a valid comparison.
We agree with the reviewer on this point and have made changes to the text (and added Table 1) to clarify what was used as training data and what was used for testing - none of the data shown in Figure 3 was used in training.

"*our classifier was trained to give the probability that a given embryo belongs to one of two classes (4.5 hpf or 17.5 hpf) with the intention of detecting developmental delays – it was not trained to predict the actual hpf of a given embryo." It is not really clear to the reader how the author is estimating the predicted hpf. The classifier seems to be trained on only two classes 4.5hpf and 17.5hpf. The output from the classification is a probability of belonging to one of each class. From these values how is the probability based predicted hpf in Figure 3 calculated? This should be clarified.*
We thank the reviewer for drawing our attention to this omission. We have now added text to the relevant methods section explaining how the probability values output by the object classifier are converted into predicted hpf values.

"hpf of embryos with a similar success rate to manual (human) staging (Figure 4b). The errors produced by the classifier (0.0 +-} 0.804; mean +-} 95% confidence interval) are comparable to the errors made by humans (0.0 +-} 0.239)." Is both rep 1 and rep 2 data used in this calculation? Are any of these batches used in the training? Please, clarify in the text.
A random selection of images from both plate 1 and plate 2 were used to generate this data. We have now noticed that, although none of the images used in training were included in this data, some images from the same wells as those used for training were included – these have now been removed and the figure revised. While there are now slightly fewer data points (44), the overall conclusion remains unchanged.

"For example, at an actual hpf of 16.0, our classifier gave predicted hpfs of 16.... embryos and embryos maintained at 25°C, respectively." This only provides info on that the development on the 25°C is slower, as it should be. But how accurate is the 10.21 ± 0.44 really? How much slower should an embryo at 25°C be? Did the author do some manual staging to compare this number with manual estimation?
We agree with the reviewer that this statement of accuracy was of limited value and have now removed it from the text. Estimations of accuracy are now limited to the data comparing our machine learning classifier with manual staging as shown in Figure 4.

Figure 1E What data is used here? It looks much cleaner than the ones shown in Figure 3? If this is just an artificial plot than are all three colors needed?
We have now updated the panel in Figure 1E to make it clear that this plot is not real data and is intended for illustrative purposes only.

Figure 3 In Figure 3b the spread of the data points for red (WT rep 2) seems much greater than for green (WT rep 1) (this can also be seen in the confidence interval in Fig 3b). It is not clear, but if the data for "rep 1" is coming from the same dataset as was used for the training, then this pattern is an indication that the method does not generalise well to a new batch. Could the author clarify where the data is coming from and provide a plot with the data separated? In addition, some quantification of the spread of the data in rep1 vs rep 2 would be of interest to the reader.
All of the data presented in Figure 3 was generated on test data only, previously unseen by the classifier. With regard to the differences in data spread, we have now added some discussion of this in the main text. Slightly more training data was drawn from plate 1 than plate 2, which may explain the slightly greater spread of data points for plate 2. It should also be noted that no quality control was applied to the test data (as this is laborious and time-consuming), so images of dead embryos and/or embryos that have drifted to the edge (our outside) of the field of view will have been included - it's possible that plate 2 produced a greater quantity of these lower quality images than plate 1. But the standard error of the mean for both plates is still less than 1.0 hpf.

(3b) From the graph there seems to be a cluster of blue (25 degree) with actual hpf 4-10 being classified as approx 15 hpf. A comment on this pattern would be useful to the reader since the error is quite large.
The majority of these datapoints correspond to the same two wells on that particular plate and it appears that segmentation error is to blame for the misclassification, as illustrated in the example below – in both cases, the embryos in the early stages of development were consistently over-segmented, which likely resulted in their misclassification.

Figure 4. "Machine Learning (ML) classifier- and human-predicted hpf for 50 images of zebrafish" Are the data taken from the same batch as was used for training? No, as explained above, we have now made absolutely sure that none of the random images used for the human-ML comparison were drawn from the training data.

Competing Interests: No competing interests were disclosed.

Reviewer Report 25 November 2022

https://doi.org/10.21956/wellcomeopenres.20298.r53245

?

Steffen Scholpp
Living Systems Institute, School of Biosciences, College of Life and Environmental Sciences, University of Exeter, Exeter, UK

The authors describe an automated method to stage zebrafish embryos. Using a machine learning approach, the authors developed a way to determine the stages of embryos without dechorionation from 4hpf – 18hpf automatically. Then, the authors use two different temperatures to evaluate the method's efficiency: staging embryos at 25°C and 28.5°C and comparing the results to manually staged embryos. The software seems to be less accurate than staging by researchers. However, the differences are minor and can be neglected. The obvious advantage is the possibility of staging a large number of embryos.

In my opinion, this is a valuable method to analyse many embryos in a short time. Moreover, the technique seems to cope reasonably well with specific challenges, such as the orientation of the individual embryos or the chorion.

Personally, I would like to see more ways in which the method is challenged. For example, it would be interesting to see how the algorithm copes with higher temperatures, i.e. 33°C or by mutations lacking fundamental body parts, such as in the headless mutant or the no tail mutant. How would the staging work in these embryos? These experiments are not strictly required for this manuscript but would provide additional insight into how the technology works.

The method was tested only for several stages (4 – 18hpf). Can that be extended to older stages, or does the twitching of the embryos after 20hpf affect the automated staging? Furthermore, is pigmentation an obstacle? Finally, does the method also work with fixed embryos? These aspects should be discussed.

Finally, the advantage of the machine learning approach was not immediately apparent. For example, does the algorithm improve after staging more embryos? The authors should clarify how many iterations have been performed and if the authors predict a further improvement in the future.

In general, this is a handy method for large chemical or genetic screens. However, with a smaller sample size, such a method seems too labour-intensive, and the specialised equipment, such as cameras and well plates, could be troublesome. Therefore, it would be interesting to read for which applications this technology would be beneficial. Similarly, it would help the readers to find a critical discussion on the limitations of this technology.

Minor comments:

The different colours in the plot in Fig. 1E are not explained and should refer to the explanation in Fig. 3.

Is the rationale for developing the new software tool clearly explained?

Yes

Is the description of the software tool technically sound?

Yes

Are sufficient details of the code, methods and analysis (if applicable) provided to allow replication of the software development and its use by others?

Partly

Is sufficient information provided to allow interpretation of the expected output datasets and any results generated using the tool?
Partly

Are the conclusions about the tool and its performance adequately supported by the findings presented in the article?
Yes

Competing Interests: No competing interests were disclosed.

Reviewer Expertise: Morphogen trafficking in zebrafish

I confirm that I have read this submission and believe that I have an appropriate level of expertise to confirm that it is of an acceptable scientific standard, however I have significant reservations, as outlined above.

Author Response 06 Mar 2023

Rebecca Jones

Responses to Reviewer 1 (Stefen Schlopp) *The authors describe an automated method to stage zebrafish embryos. Using a machine learning approach, the authors developed a way to determine the stages of embryos without dechorionation from 4hpf – 18hpf automatically. Then, the authors use two different temperatures to evaluate the method's efficiency: staging embryos at 25°C and 28.5°C and comparing the results to manually staged embryos. The software seems to be less accurate than staging by researchers. However, the differences are minor and can be neglected. The obvious advantage is the possibility of staging a large number of embryos. In my opinion, this is a valuable method to analyse many embryos in a short time. Moreover, the technique seems to cope reasonably well with specific challenges, such as the orientation of the individual embryos or the chorion.*
We thank the reviewer for taking the time to read and comment on our manuscript, and for the supportive and helpful comments.

Personally, I would like to see more ways in which the method is challenged. For example, it would be interesting to see how the algorithm copes with higher temperatures, i.e. 33°C or by mutations lacking fundamental body parts, such as in the headless mutant or the no tail mutant. How would the staging work in these embryos? These experiments are not strictly required for this manuscript but would provide additional insight into how the technology works.
We agree with the reviewer that it would be interesting to see how the algorithm copes with higher temperatures, and indeed we did try this at 32°C, however the embryos did not develop well and the majority died within a few hours of starting the experiment. In terms of testing the algorithm using mutants with a different morphology to WT (e.g. the headless mutant as suggested), our classifier was only designed to detect normal WT development, and whilst the same analysis pipeline could be used, the algorithm would need to be retrained on the 'new' morphology. We have added an additional comment in our

manuscript to reflect this, and we thank the reviewer for bringing it to our attention.

The method was tested only for several stages (4 – 18hpf). Can that be extended to older stages, or does the twitching of the embryos after 20hpf affect the automated staging? Furthermore, is pigmentation an obstacle? Finally, does the method also work with fixed embryos? These aspects should be discussed.
The reviewer asks whether the method could be extended to older stages, and this is indeed possible. The end point of the experiment (17.5 hpf) was sufficient for proof-of-concept, and there was therefore no reason for us to train the classifier beyond this stage. The end point could easily be extended, and the emerging pigmentation of the embryos as they develop would not pose a problem. Similarly, twitching of the embryos would not affect the outcome. This is because a unique advantage of our classifier is the complete absence of quality control; it is able to accurately quantify developmental temporal trajectory regardless of the orientation of the embryos. We have included a further sentence in our manuscript reflecting this, and we thank the reviewer for the helpful input. In respect of fixed embryos, this is not something we tried, and due to differences in opacity of live vs fixed embryos, we believe that the classifier may need retraining in the same way as described above.

Finally, the advantage of the machine learning approach was not immediately apparent. For example, does the algorithm improve after staging more embryos? The authors should clarify how many iterations have been performed and if the authors predict a further improvement in the future.
The reviewer has raised an interesting point regarding whether the algorithm improves after staging more embryos, and we have included some additional text in our manuscript to clarify the number of training iterations that were performed. Theoretically, the machine-learning based classifier could improve if we trained it on more WT embryos, however it is also possible to over-train the classifier, resulting in its abilities becoming too specific to the training sets and less generalizable over a wider range of samples. There is consequently a training 'sweet-spot', and less can be more. Deep-learning is a way to potentially address this, but discussion of this falls outside the scope of this current paper.

In general, this is a handy method for large chemical or genetic screens. However, with a smaller sample size, such a method seems too labour-intensive, and the specialised equipment, such as cameras and well plates, could be troublesome. Therefore, it would be interesting to read for which applications this technology would be beneficial. Similarly, it would help the readers to find a critical discussion on the limitations of this technology.
The reviewer comments that our method seems labor-intensive for smaller sample sizes. This is a valid point, and we have added text in the discussion section of the manuscript regarding the limitations of our classifier, which we trust addresses this issue.

The different colours in the plot in Fig. 1E are not explained and should refer to the explanation in Fig. 3.
Figure 1E has now been replaced with a simple schematic to show how the technology works; the figure legend has been updated accordingly.

Competing Interests: No competing interests were disclosed.