# High throughput automated detection of axial malformations in Medaka embryo

Diane Genest[a, b,*, 1], Elodie Puybareau[a, c, 1], Marc Leonard[b], Jean Cousty[a], Noémie De Crozé[b], Hugues Talbot[a]

[a] *Université Paris-Est, LIGM (UMR 8049), CNRS, ENPC, ESIEE, UPEM, 2 Boulevard Blaise Pascal, 93162 Noisy-le Grand, France*

[b] *L'OREAL Research & Innovation, 1 avenue Eugène Schueller, 93600 Aulnay sous Bois, France*

[c] *EPITA Research and Development Laboratory (LRDE), 14-16 rue Voltaire, 94270 Le Kremlin-Bicêtre*

## Abstract

Fish embryo models are widely used as screening tools to assess the efficacy and /or toxicity of chemicals. This assessment involves the analysis of embryo morphological abnormalities. In this article, we propose a multi-scale pipeline to allow automated classification of fish embryos (Medaka: *Oryzias latipes*) based on the presence or absence of spine malformations. The proposed pipeline relies on the acquisition of fish embryo 2D images, on feature extraction based on mathematical morphology operators and on machine learning classification. After image acquisition, segmentation tools are used to detect the embryo before analysing several morphological features. An approach based on machine learning is then applied to these features to automatically classify embryos according to the presence of axial malformations. We built and validated our learning model on 1,459 images with a 10-fold cross-validation by comparison with the gold standard of 3D observations performed under a microscope by a trained operator. Our pipeline results in correct classification in 85% of the cases included in the database. This percentage is similar to the percentage of success of a trained human operator working on 2D images. The key benefit of our approach is the low computational cost of our image analysis pipeline, which guarantees optimal throughput analysis.

*Keywords:* Toxicology screening, discrete mathematical morphology, segmentation, features screening, machine learning, random forest.

[*] Corresponding author.

*Email address:* diane.genest@esiee.fr (Diane Genest),

elodie.puybareau@lrde.epita.fr (Elodie Puybareau)

[1] These authors should be considered as equal first authors.

## Introduction

Toxicological screening of chemicals is based on the analysis of reliable biological descriptors of model organisms. To assess the effect of compounds, several endpoints are analysed per individual, generating a large amount of data. For processing the data, automation appears to be necessary.

In compliance with international regulations relative to the welfare of animals used for scientific purposes [1][2], fish embryos provide ethically acceptable models for the development of screening methods to assess human and environmental toxicity of chemicals [3][4]. Moreover, early developmental stages of certain species such as Zebrafish (*Danio rerio*) and Medaka (*Orizia latipes*) are transparent, which facilitates observation of their organogenesis. Finally, fishes are vertebrates and key mechanisms of embryonic development are conserved throughout evolution from fishes to human. Fish embryos are thus considered to be a relevant model for studying the impact of chemicals on human embryonic development [5][6] and are commonly used in pharmacology and toxicology studies [7][8]. In this study, Medaka embryos are used. We focus on the eleutheroembryo stage that follows hatching and that is characterized by the presence of the yolk sac providing the energy supply necessary to organism development [9][10]. In the further article, eleutheroembryos are referred to as alevins.

### *Objectives and constraints*

Developmental toxicology assessment consists of classifying alevins according to the presence or the absence of malformations and is performed manually most of the time [11][12]. This assessment uses visual analyses that depend on both the operator and the observation conditions. This means that an operator can have a different analysis of the same data set depending on the observation conditions. To improve this process, which is time-consuming and subjective, some form of automation is required. Image processing tools and pattern recognition have been widely used in alevins studies and high-throughput screening [13][14][15]. In particular, several articles have shown the efficiency of supervised learning techniques in the scope of alevins phenotypes classification [16]. Nevertheless, most of the proposed methods are limited to the analysis of the alevin seen from a precise orientation, implying to manually position each alevin in this specific orientation before starting the image acquisition [17]. In some of these studies, image-based observations are considered as ground truth [16]. Because every malformation is not always visible from every point of view, taking image-based observations as a reference can occult some of these malformations. Here, we propose an experimental protocol that does not require manual positioning of the alevin in a given orientation and that considers as ground truths the observation of the alevin under a microscope by a trained user who has the possibility to analyse the given alevin from any possible orientations. Such conditions correspond to the use case of the software in a real assay.

Our objective is to propose an automated method for classification of alevins with or without a spine malformation, one of the most common developmental abnormalities observed [17][18]. This classification is based on the analysis of 2D images acquired according to the protocol described in [19]. In the acquired images, the alevins can appear in any orientation from the lateral view to the dorsal view (Figure 1a to c). Moreover, spine malformations cover an important variety of phenotypes, from the most obvious malformation to slightest defects of the spine curvatures (Figure 1d and e). Some specific cases of strongly bent alevins are refered as hook-shaped (Figure 1f). This huge variety in alevins phenotypes makes spine malformation complicated to characterize.  Mathematical morphology operators can provide an accurate description serving as input to a Machine Learning classifier. Working with 2D images implies loss of information compared to 3D observations made under a microscope. To validate the proposed set up, we challenge ground truth reliability by quantifying the gap between observations under a microscope and on 2D images. In addition, in order to quantify human subjectivity, we provide an estimation of the inter-observer subjectivity rate according to image-based observations made by three different observers.

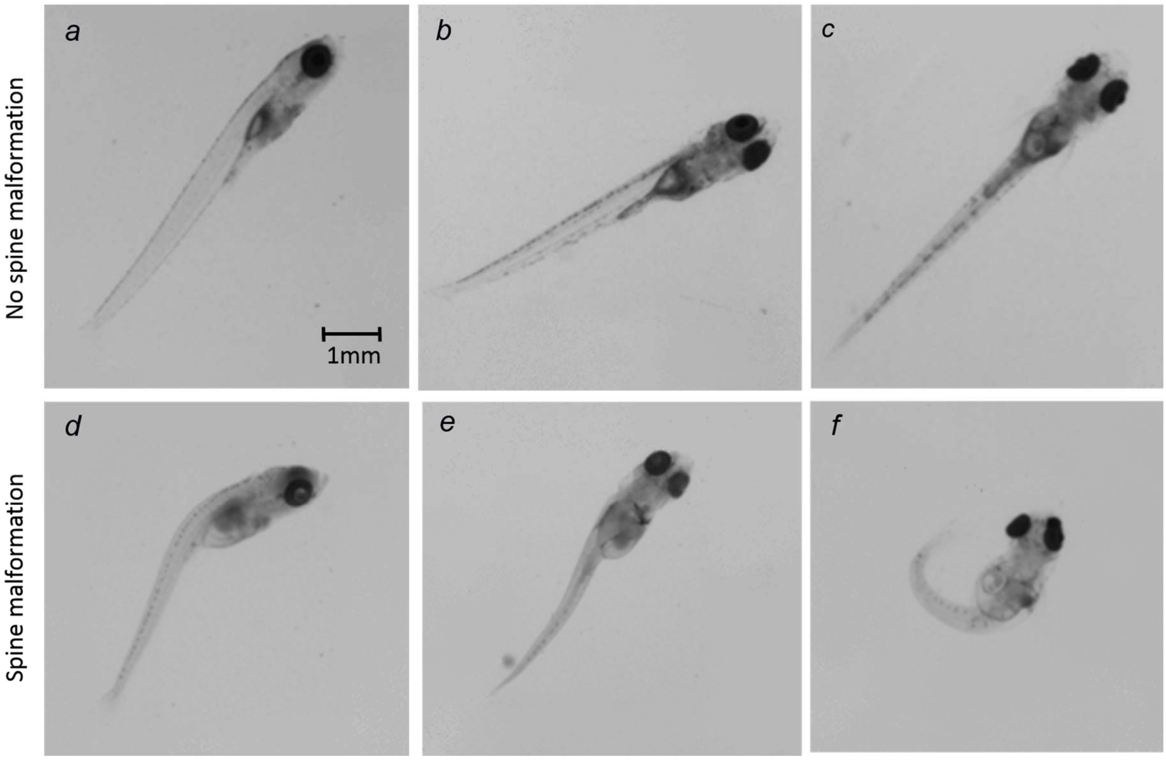


*Figure 1. Images of 9 dpf Medaka alevins as acquired by our set-up. a to c: healthy alevins shown in lateral view in a, three-quarters view in b and dorsal view in c. d to f: alevins showing different types of spine malformations, d being a major spine malformation (lateral view), e a slight "S-shaped" malformation (three quarter view) and f a hook-shaped alevin (dorsal view).*

Assessing the efficacy of our automated classifier implies to pay attention to both the sensitivity and the specificity of the classification. The sensitivity (i.e. the capacity of a test to indicate a correct positive result) corresponds to the proportion of malformed alevins correctly detected. Specificity refers to the ability of the test to correctly indicate a negative result, i.e. the ratio of healthy alevins correctly detected. The overall accuracy is the average of both numbers weighted by their population. The chemicals safety assessment involves reducing the number of false negatives, i.e. high sensitivity. On the other hand, in particular in an industrial context, specificity also needs to be high, because false detection of abnormalities could penalize production. Consequently, both specificity and sensitivity tests must be optimized, which corresponds to the conventional choice of optimizing the overall accuracy.

### *Proposed method*

In this article, we describe a new automated method to detect alevin spine malformations. Most of these malformations are characterized by abnormal spine curvature. Some alevins also exhibit shortened spines or humps. Inter-individual variability and the single orientation acquired in 2D images complicates the detection of axial malformations, due to the variability in alevin orientation from one image to the other. The first difficulty is thus to identify relevant parameters in order to characterize such a panel of malformations. Our method is based on binary spine modelling in order to extract numerical values relative to spine characterization. To this end, we consider an approach based on the morphological skeleton [20][21]. Features such as dimensions, curvature, angles are then deduced from this skeleton and gathered in a features vector in order to feed a random forest classifier [22]. The flowchart of our methodology is summarized in Figure 2.

The proposed method comprises two phases. The learning phase builds the classification model, which is then used to classify data during the testing phase. Learning is based on a set of labelled data. It begins with a pre-processing step (described in detail in the Appendix) that reduces the acquired data to the region of interest [19]. In the feature extraction step, the alevin spine is segmented using mathematical morphology operators [23]. Following segmentation, morphological parameters are measured on the spine and the alevin mask. A random forest classifier is built and fitted to the set of labelled data. During the testing phase, features are also extracted from the testing dataset and images are classified according to the trained random forest model.

Our pipeline is made up of simple and fast operators, that are, for the most part, available in off-the-shelf image analysis software packages such as the PINK image processing library [24] and scikit-learn library [25].

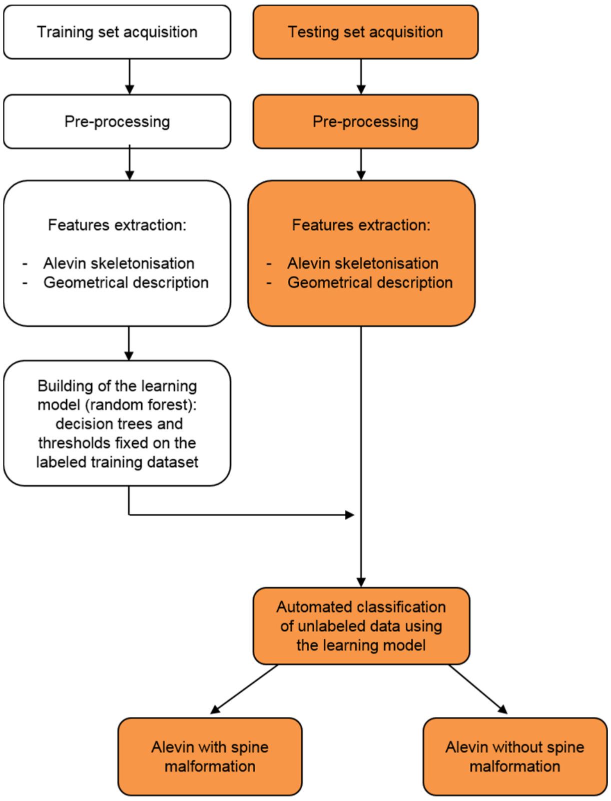

Training set acquisition
Testing set acquisition
Pre-processing
Pre-processing
Features extraction:
- Alevin skeletonisation
- Geometrical description
Features extraction:
- Alevin skeletonisation
- Geometrical description
Building of the learning model (random forest): decision trees and thresholds fixed on the labeled training dataset
Automated classification of unlabeled data using the learning model
Alevin with spine malformation
Alevin without spine malformation

*Figure 2. Flowchart of the alevin morphological abnormalities detection assay based on image processing. This detection method is assessed by cross-validation in the presented study.*

*Contributions and outlines*

The main contributions to this article are the following:

- A dataset of 1459 alevins associated with ground-truth is built. Each alevin is screened under a microscope and through observation of acquired 2D images. For each alevin, the presence or absence of malformations is established by a trained operator during the microscope observations, such labelling being considered as ground-truth. Furthermore, a second labelling is produced by independently reading the 2D images, allowing one to assess the loss of information due to the 2D acquisition of 3D alevins. To challenge ground truth reliability, additional observations and labelling are performed by three experts on a subset of 200 images, making assessment of inter-expert subjectivity possible;
- The efficacy of mathematical morphology operators is shown for characterizing alevin malformations and for feeding an automated classifier;

- The 2D images are used to show that alevins can automatically be classified with an accuracy similar to image-based human classification and with time efficiency (a few seconds for each image) that is compatible with its use in a high throughput industrial context.

Section 1 introduces the classifier used in the proposed approach, presenting the functions and mechanisms related to the random forest estimator. The features extraction process is presented in Section 2, including the description of our automated method for alevin spine segmentation and for spine geometrical description. Section 3 explains how the learning model is established. The experimental setup is described in Section 4 and the assessment results are provided in Section 5.

## 1. Background notions for random forest classification

Decision trees are often used as predictive models for classification purposes in supervised learning. In this section, we quickly recall the basic concepts behind decision trees and random forest classifiers.

A decision tree is a directed binary tree where non-leaf nodes carry decision rules and where leaves are labelled. The decision rules associated with each node take the form of Boolean test functions pointed toward their respective children. The label associated to a leaf corresponds to a final class. More formally, a decision tree is a 4-tuple $(N, P, F, L)$ defined by the ensemble of nodes $N$, the ensemble of parent relations between them $P$, the mapping $F$ which associates a Boolean test function to each non-leaf node and a mapping $L$ that provides a label to each leaf node.

A decision tree-based algorithm classifies data based on a set of features (a.k.a. descriptors). At each non-leaf node, an associated test function takes a single feature as argument and compares it to a fixed threshold. Depending on the result of the comparison, either the right or the left child node is chosen. Thus, starting from the root of the tree and given a feature vector, a path is created from the root through the nodes until it reaches a leaf. The algorithm returns as output, the label of this leaf. The definition of a test function ensemble $F$ is given in Section 3.

The accuracy of a decision tree-based algorithm can be assessed on a data sample by comparing the predicted values on this sample with a corresponding set of correctly labelled data. On a sample of size

$n_{\text{sample}}$, we respectively call $y$ and $\hat{y}$ the series of labelled and predicted values. If $y_i$ is the label of the i$^{\text{th}}$ data and $\hat{y}_i$ is the corresponding predicted value, then we calculate the accuracy rate of the algorithm on this sample as the fraction of correct predictions over the total number of data in this sample. More precisely, the accuracy of the sample is given by:

$$(1)\ accuracy(y, \hat{y}) = \frac{1}{n_{\text{sample}}} \sum_{i=1}^{n_{\text{sample}}} 1(y_i, \hat{y}_i),$$

where $1(y_i, \hat{y}_i)$ is equal to 1 if $y_i$ is equal to $\hat{y}_i$ and 0 otherwise.

Fitting a Boolean test function to a training set of labelled data consists of finding the most relevant feature and its associated optimal threshold, according to certain criteria, like optimal accuracy on a training set. Then, the training set is split into two parts according to this Boolean test function and the process is carried out recursively on the two child nodes, until another criterion is fulfilled, such as desired accuracy or maximum branch depth. A limitation of decision trees is their tendency to overfit the data. Overfitting is defined as the tendency of a classifier to correspond too closely to a particular set of training data, jeopardizing its ability to correctly classify future observations. For this reason, it is recommended to not train decision trees on the entire available dataset but to train and test respectively on a collection of subsets and their complement in multiple ways. This process is called cross-validation.

Overfitting can also be reduced significantly by training multiple decision trees, using multiple subsets of features and submitting the results of these trees to a voting process. This process is what forms the basis to Random Forests (RF). RF are defined as an ensemble of decision trees that outputs a final prediction class corresponding to a function of every tree output classes. This principle is based on the idea that, as a single entity, a decision tree is not effective for high dimensional data. However the combination of many weak decision trees can produce a stronger and more reliable classifier [22]. To this end, RF are fitted using the general technique of bootstrap aggregating, or bagging. Each decision tree is computed (node split functions are defined) on a random subset of the training dataset, using a randomly selected set of features [26]. This technique is currently used to reduce misclassification error due to single application of the partitioning clustering procedure [27][28].

Feature characterization is a key requirement for decision tree building. The aim of this process is to obtain various objective descriptions of the data that needs to be classified. Such descriptions are then used as arguments to the decision functions. In our method, the classifier is designed to classify images of alevins depending on the presence or the absence of axial malformation. The following section describes features that enable characterization of such malformations.

## 2. Feature extraction for alevin spine characterization

We describe in this section a method for obtaining a geometric description of alevins from 2D images. Image analysis, including mathematical morphology, is used to characterize the spinal shape of alevins from grey-scale images [23][29]. Section 2.1 proposes a procedure to approximate the alevin's spine. Feature characterization is presented in Section 2.2.

### *2.1. Alevin axial segmentation method*

In this section, we start from a first segmentation of the whole alevin contour obtained during a pre-processing step summarized in the appendix. We denote by $\mathcal{M}$ the resulting segmentation (Figure 4a). Our aim is then to obtain, from $\mathcal{M}$, a segmentation which approximates the curve of the alevin's spine. After smoothing the contour of the alevin, this methodology implements morphological skeletonisation. More precisely, the spine approximation method uses the curvilinear skeleton principle described in [20]. An overview of the spine segmentation from the alevin mask is given in Figure 3.

Firstly, in order to filter out any artefact ramification, we begin by filling the convex areas on the alevin contour $\mathcal{M}$ with a morphological closing $\varphi_{\Gamma_{r_1}}$ by a disk-shaped structuring element $\Gamma_{r_1}$ of size $r_1$ [23]. In the following, we denote by $\mathcal{M}'$, the result of this process applied to $\mathcal{M}$:

(1) $\mathcal{M}' = \varphi_{\Gamma_{r_1}}(\mathcal{M})$.

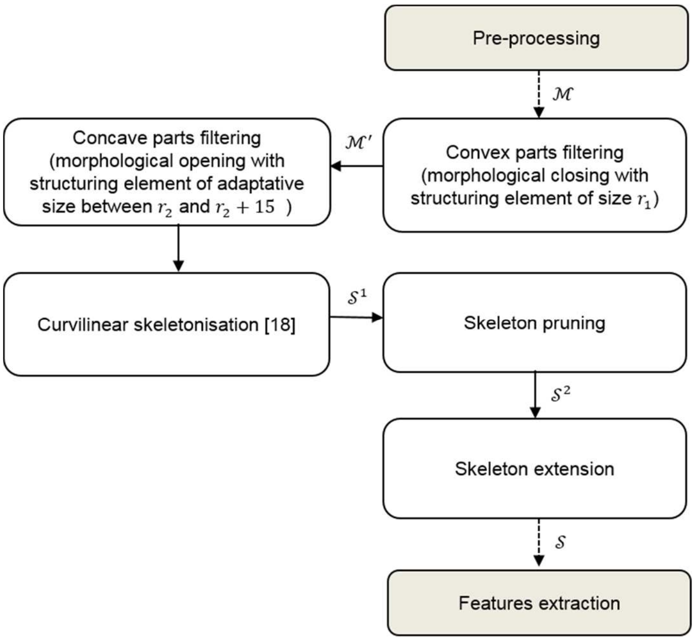

Pre-processing
$\mathcal{M}$
Convex parts filtering (morphological closing with structuring element of size $r_1$)
$\mathcal{M}'$
Concave parts filtering (morphological opening with structuring element of adaptative size between $r_2$ and $r_2 + 15$ )
Curvilinear skeletonisation [18]
$\mathcal{S}^1$
Skeleton pruning
$\mathcal{S}^2$
Skeleton extension
$\mathcal{S}$
Features extraction

*Figure 3. Flowchart of alevin spine approximation.*

On the other hand, concave areas due to alevin abnormalities such as significant oedemas or poor initial segmentation are more problematic because they may cause important ramifications in the subsequent skeleton application step. To filter out these concave areas, which can be more or less significant in size, we consider an iterative process which determines the smallest amount of filtering used to obtain a skeleton without any ramification. In our methodology, such filtering is performed with morphological openings by disk-shaped structuring elements. More precisely, we consider the curvilinear skeleton $S_i(X)$ of the largest connected component of the opening of $X$ by a disk-shaped structuring element of radius $i$. Hence, if we denote by $r_2$ the minimal radius considered in the proposed setting, we consider the resulting skeleton $\mathcal{S}^1$ defined by:

(2) $\mathcal{S}^1 = S_{r_2+3.\min(5,\lambda)}(\mathcal{M}')$,

where $\lambda = \min\{i \in \mathbb{N}$ such as $S_{r_2+3i}(\mathcal{M}')$ has two extremities$\}$. A further pruning step removes potential residual ramifications in $\mathcal{S}^1$, by filtering out the skeleton branches with a length less than $\alpha$ pixels. We write:

(3) $\mathcal{S}^2 = pruning_\alpha(\mathcal{S}^1)$,

where $pruning_\alpha$ denotes the skeleton pruning strategy of parameter $\alpha$.

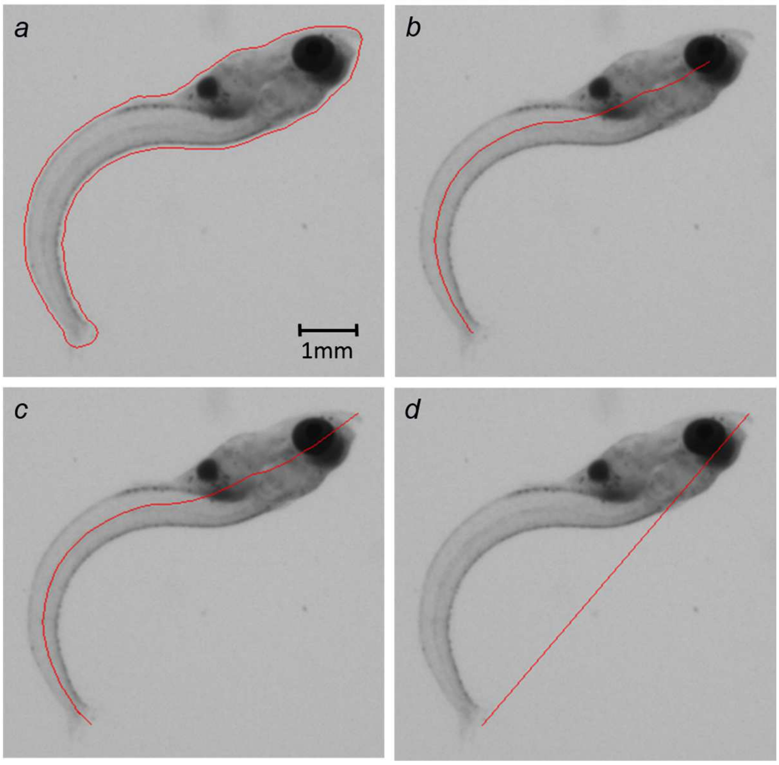


*Figure 4. Spine approximation steps on the cropped image of an alevin. The red line represents the contour of the initial mask $\mathcal{M}$ in a, the initial curvilinear skeleton $\mathcal{S}^2$ in b, the extended curvilinear skeleton $\mathcal{S}$ in c and the straight line $\mathcal{L}$ linking both ends.*

From its definition, the curvilinear skeleton $\mathcal{S}^2$ (Figure 4b) does not reach the borders of the alevin's shape $\mathcal{M}$ (Figure 4a). In order to more effectively approximate the alevin's actual spine, both extremities of the skeleton $\mathcal{S}^2$ are detected and extended up to the mask boundaries. To achieve this, for each skeleton extremity $p^i$, we draw the straight line linking $p^i$ to the point located five points behind the skeleton curve. This segment extends past $p^i$all the way to the border of $\mathcal{M}$. The resulting skeleton is denoted by $\mathcal{S}$ in the following (Figure 4c). This spine segmentation is accurate in cases of alevins seen in dorsal view because such alevins appear symmetric. However, in lateral view, the spine segmentation is systematically deviated near the yolk sac, instead of following the dorsal line. Nevertheless, it is not a problem for our purpose. Indeed, exact spine segmentation is not a goal per-se. It is a way to measure features for classification (see Section 2.2), and the observed deviation does not highly impact the features measurement further described. Finally, both skeleton extremities are then linked via a line segment $\mathcal{L}$, as shown in (Figure 4d). Because a healthy alevin is expected to present a straight spine when it is anaesthetized, this segment is used in the following section as a reference to compare the actual alevin's spine to a healthy spine.

### *2.2. Geometrical features description*

Classifying alevin malformations from images by using a learning-based approach requires an accurate description of the malformation that we want to detect. Hence, from the segmentations obtained as described in Section 2.1, we select relevant and discriminative features to reliably distinguish between alevins with and without a spine abnormality. Features are measured through the assessments of (i) the alevin dimensions (Section 2.2.1), (ii) the curvature (Section 2.2.2), (iii) the regularity (Section 2.2.3) and (iv) the discontinuities of the alevin's shape (Section 2.2.4).

### *2.2.1. Dimension measurement on the alevin masks*

A first set of parameters, namely $a_{\text{alevin}}$, $l_{\text{alevin}}$, $w_{\max}$, $w_{\text{mean}}$, $r_{\text{image}}^1$ and $r_{\text{image}}^2$ described below are related to the dimensions of the alevin. The alevin area $a_{alevin}$ is measured on mask $\mathcal{M}$ in number of pixels. The parameter $l_{\text{alevin}}$ refers to the alevin's length, measured as the Euclidean length of the skeleton $\mathcal{S}$. Maximum and average widths are calculated using the maximal balls principle. For that, the Euclidean distance map is computed to the exterior of the alevin mask $\mathcal{M}$ [30][31] and restricted to the skeleton $\mathcal{S}$. Thus, each point of the skeleton is associated with its distance to the external part of the alevin mask[1]. The largest and the average values are extracted and multiplied by two to obtain the maximal and average widths denoted by $w_{\max}$ and $w_{\text{mean}}$, respectively. We compute the ratios $r_{\text{image}}^1$ and $r_{\text{image}}^2$ between the alevin's length and width as follow:

$$(4)\ r_{\text{image}}^1 = \frac{w_{\text{mean}}}{l_{\text{alevin}}}\ ;\text{ and } r_{\text{image}}^2 = \frac{w_{\max}}{l_{\text{alevin}}}.$$

### *2.2.2. Curvature assessment from the graphical representation of the alevin's spine*

The aim of this part is to extract features related to spine deviation from the straight line joining its two extremities. The relevant parameters are denoted by $AUC$, $d_{\max}$, $d_{\text{mean}}$, $r_{\text{graph}}^1$, $r_{\text{graph}}^2$, and $r_{\text{graph}}^3$. We build an image representation of the alevin's spine in order to simplify its analysis in a direct

[1] This weighted skeleton is called the extinction function [23]

orthonormal frame. We aim to lay both the spine extremities on the abscissa axis. To this end, we search for the composition of the translation $\vec{\mathrm{T}}$ and the rotation R that register the line segment joining the extremities of the spine curve to the segment $[(0,0),(l,0)]$ where $l$ is the distance between the two extremities. The result is shown on Figure 5b.

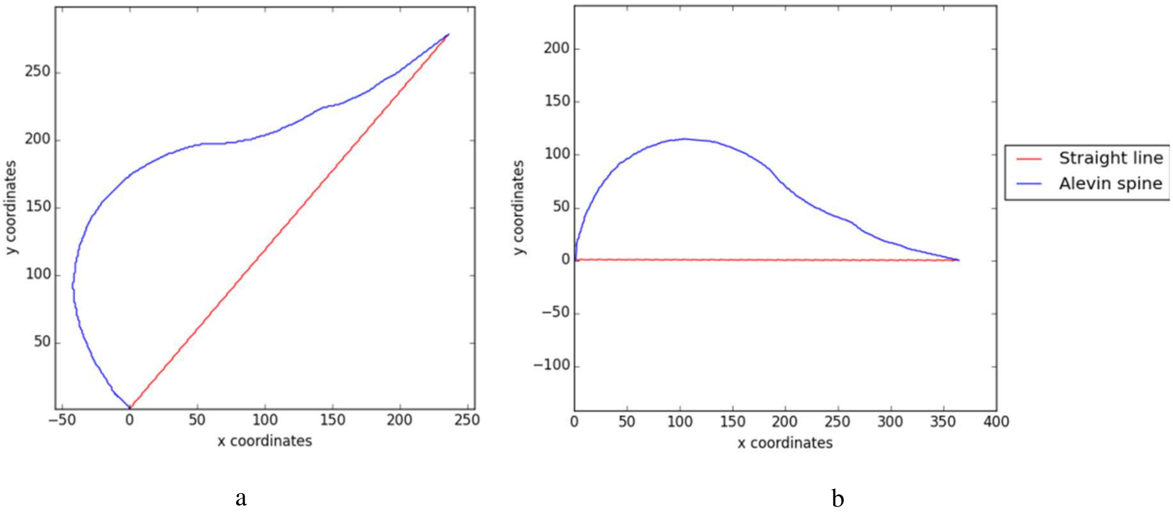


*Figure 5. Graphical representation of the curvilinear skeleton $S$ in a direct orthonormal frame. The spine curve is represented after translation $\vec{T}$ (a) and after translation $\vec{T}$ and rotation R (b).*

Depending on the curve shape, it is not always possible to represent the detected alevin spine as an explicit function. In particular, when multiple points of the curve, representing the alevin spine in the presenting orthonormal frame, have the same abscissa, the spine is considered to have a hook. This case is described in Section 4.2 and Figure *1*f. In the normal case, we consider the spine curve as the graphic representation of a function $f$ in an orthonormal frame. We write $(x_i, f(x_i))$ the coordinates of the i$^{th}$ point of the curve. The total number of points on the curve is $n$. This representation is used to measure several numerical parameters, which are chosen for their ability to characterize the spine shape. In particular, the abscissas axis is taken as reference and spine deviation is estimated with the following features.

The area under the curve (AUC) of the function $|f|$ is computed using the trapezoidal rule [32], where $|f|$ is the absolute value of $f(x)$ for every points $x$ of the domain:

(5) $AUC = \sum_{i=1}^{n} \frac{(|f(x_{i-1})|+|f(x_i)|)}{2} \times (x_i - x_{i-1})$.

The use of the absolute value allows analysing every alevin equally, even those with S-shaped spinal cord i.e. those for which function $f$ is somewhere above and somewhere below the line segment joining the extremities of the alevin's spine. The maximal deviation $d_{max}$ and the average deviation $d_{mean}$ are calculated considering the maximal and average distances between the spine curve and the abscissas axis respectively, meaning the maximum and average values of the curve ordinates:

(6) $d_{\text{max}} = \max(f(x_i))\, for\ i\ \in [0, n]$ ; and

(7) $d_{\text{mean}} = \frac{1}{n}\sum_{i=0}^{n} f(x_i)$.

From these parameters, three ratios $r_{\text{graph}}^1$, $r_{\text{graph}}^2$, and $r_{\text{graph}}^3$ are considered to characterize the flatness of the spine:

(8) $r_{\text{graph}}^1 = \frac{d_{\text{max}}}{l_{\text{alevin}}}$ ; $r_{\text{graph}}^2 = \frac{d_{\text{max}}}{d_{\text{mean}}}$ ; and $r_{\text{graph}}^3 = \frac{\text{AUC}}{l_{\text{alevin}}}$.

### *2.2.3. Curve regularity assessment*

The spine shape can also be discriminant even if no important deviation is detectable. Even a slight curve in the alevin's spine can be representative of an anomaly depending on the regularity of the curve. Indeed, a recently anaesthetized alevin or immediately after hatching and still undergoing deployment could have such an appearance without this necessarily pointing to a malformation. We now describe parameters $r_p^2$ and $r_c^2$ that represent information about the regular appearance of the spine curve. For this purpose, we approximate the function $f$ (see Section 2.2.2) by a parabola. Hence, we define the parabolic function $f_p$ defined by:

(9) $f_p(x) = a_1x^2 + b_1x + c_1$,

where the triplet $(a_1, b_1, c_1)$ is chosen to most effectively approximate the initial function $f$ via least-squares. We then consider the determination coefficient $r_p^2$ as follows:

(10) $r_p^2 = 1 - \frac{\sum_{i=0}^{n}(f(x_i)-f_p(x_i))^2}{\sum_{i=0}^{n}(f(x_i)\ -m)^2}$,

where $m = \frac{1}{n}\sum_{i=0}^{n} f(x_i)$ is the average of the function ordinates. In a similar way, we compute the determination coefficient $r_c^2$ of the cubic function $f_c$ defined by the equation

$f_c(x)\ =\ a_2x^3\ +\ b_2x^2\ +\ c_2x\ +\ d_2$ and that most effectively approximates the initial function $f$:

$$(11) \qquad r_c^2 = 1 - \frac{\sum_{i=0}^{n}(f(x_i)-f_c(x_i))^2}{\sum_{i=0}^{n}(f(x_i)-m)^2}.$$

Both $r_p^2$ and $r_c^2$ coefficients are used as descriptors of spine curve regularity.

### *2.2.4. Curve discontinuities assessment*

Some alevins exhibit disruptions in their spine, that can be detected by the presence of large, abrupt angles. Such irregularities may not cause important deviations with respect to the straight line linking both extremities. As a result, they cannot be sufficiently characterized by the previously described features. To reveal such irregularities, an algorithm was developed in order to approximate the skeleton by a broken line and to assess the main angles in the alevin curve. It consists of searching for the significant extrema of the piecewise affine function that best represents the spine curve and of linking them by line segments.

We consider the skeleton curve as a 1D signal that is smoothed by a convolution with a Gaussian kernel of size $\sigma$. This step reduces the number of spurious angular variations that are mostly due to the discrete aspect of the pixel-supported signal. Reflective boundary conditions are used to limit border effects on the skeleton signal. We then search for local extrema. Their coordinates are gathered in a vector $V$. Both extremities are added at the beginning and at the end of $V$.

Because of the discrete domain representation, or due to some oscillations on the spine segmentation, some of these extrema are close to each other and do not represent significant angular changes. To filter out extrema that are not significant, we search for steady portions of the spine curve. We define as a steady portion a subsequence in vector $V$ that is as long as possible and whose successive points are close to each other. A vertical distance threshold $d_1$ is defined below which two successive points of $V$ are considered to be within a steady portion. From the vector $V$, all the extrema located between the two extremities of a steady portion are removed. A horizontal distance threshold $d_2$ is then defined, below which a steady portion is simplified by replacing its extremities with a unique centred point. The broken line that links the selected extrema is finally considered. An example of this process is presented in Figure

6. The number of angles $n_{angles}$ detected on the broken line created, the minimal angle $\theta_{min}$, and the maximal angle $\theta_{max}$ are saved as features.

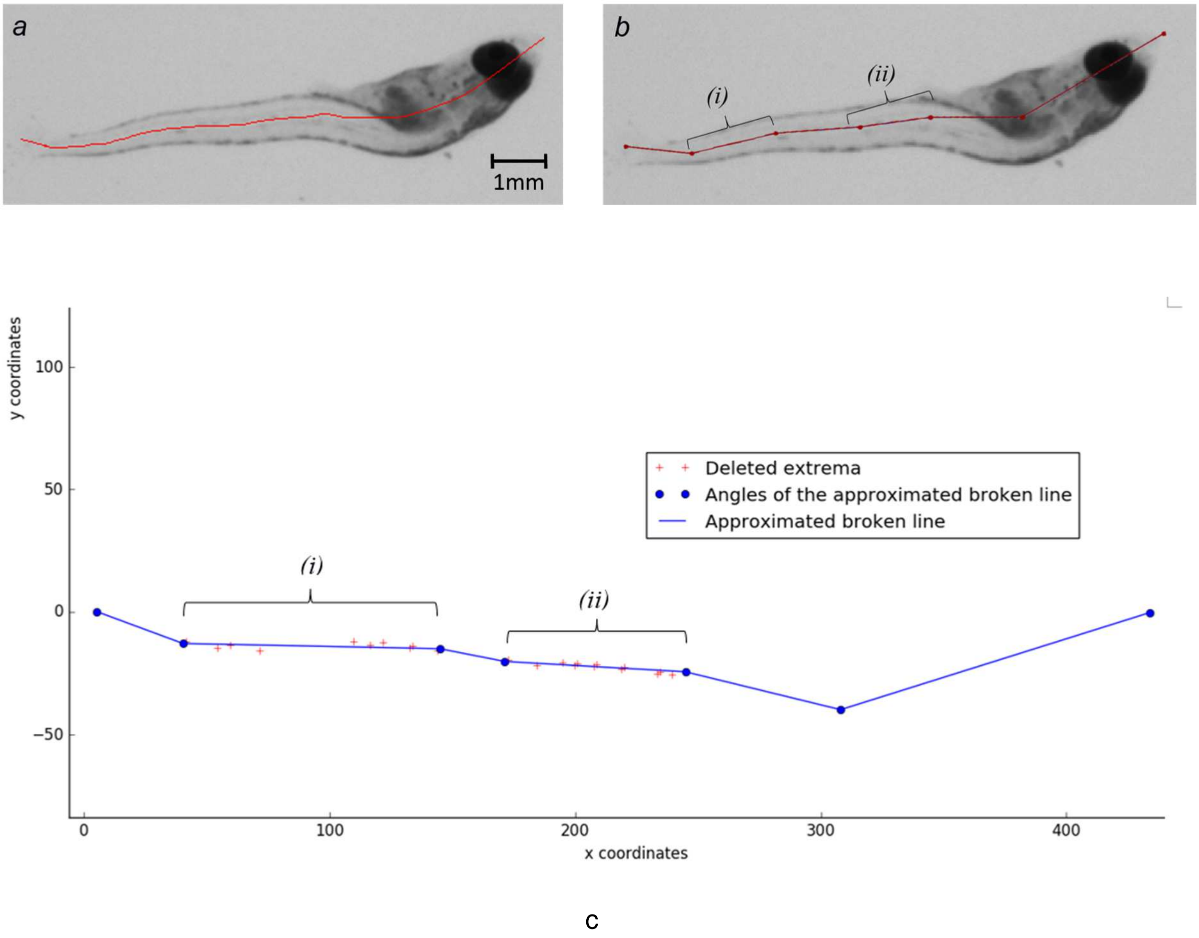


*Figure 6. Alevin spine approximation by a piecewise affine function. The red line shows the spine segmentation* $S$ *in a, the approximated spine in b, superimposed on the cropped image. The approximated spine is represented in a direct orthonormal frame in c. In b and c: the areas (i) and (ii) are detected as steady portions of the curve whose only extremities are maintained as the broken line angles. The red crosses represent the extrema deleted from the initial spine graphical representation. In fine, the retained angles and the delineation of the approximated broken line appear in blue. For this alevin, the following parameters are measured:* $n_{angles} = 5$, $\theta_{min} = 149°$, *and* $\theta_{max} = 172°$.

We summarize the parameters characterizing the alevin's spine and used during classification in Table 1.

| | |
|---|---|
| Alevin dimension descriptors | $a_{\text{alevin}}$ ; $l_{\text{alevin}}$ ; $w_{\text{mean}}$ ; $w_{\text{max}}$ ; $r^{1}_{\text{image}}$ , $r^{2}_{\text{image}}$ |
| Curvature descriptors | $AUC$ ; $d_{\text{max}}$, $d_{\text{mean}}$ ; $r^{1}_{\text{graph}}$, $r^{2}_{\text{graph}}$, $r^{3}_{\text{graph}}$ |
| Curve regularity descriptors | $r_p^2$ ; $r_c^2$ |
| Curve break descriptors | $n_{\text{angles}}$ ; $\theta_{\text{min}}$ ; $\theta_{\text{max}}$ |

*Table 1. List of features extracted from alevin segmentations and used during axial classification*

## 3. Learning model

Many parameters and rules are involved in our RF model and determine the capacity of the model to classify correctly. They are specified before the classifier training step and make it possible to adapt it to the data constraints. We present some of them in this section.

During learning, we search for the ensemble $N$ of nodes, the parent relations $P$ between them and the set $F$ of test functions associated with each node. For each tree, we firstly consider a single root node to which we associate all the labelled data from the training sample. Then, we recursively decide if the node needs to be split with the associated dataset. To decide if a node needs to be split or if the learning model needs to be stopped, we use the standard entropy criterion. Applied to a sample, entropy measures its level of impurity, in term of label distribution. A sample with an entropy of zero means this sample only contains elements with the same label. Conversely, entropy is maximal when uniform label distribution is observed in the sample. The entropy of a binary sample $S$ of labelled data is defined by:

$$(12) \qquad H(S) = -(p_{L_-}\log_2 p_{L_-}) - (p_{L_+}\log_2 p_{L_+}),$$

where $p_{L_+}$ and $p_{L_-}$ are respectively the relative frequencies of the positive label $L_+$ and the negative label $L_-$ in $S$. If the entropy of $S$ is higher than a given threshold, we divide the sample into two subsamples. In order to determine these two subsamples, we search for the related splitting function $s$ defined as follows. Given a feature function $\Phi$ and a threshold $\vartheta$, the splitting function $s$ associated with $\Phi$ and $\vartheta$ is the map $s_{\Phi,\vartheta}$ from the set of data into the set $\{\text{True, False}\}$ such as

$s_{\Phi,\vartheta}(x) =$ True whenever the feature $\Phi(\text{x})$ is higher than the value $\vartheta$ i.e $\Phi(\text{x}) > \vartheta$.

To any set $S$ of data and any splitting function $s_{\Phi,\vartheta}$, the $\text{Gain}(S, s_{\Phi,\vartheta})$ function is associated, defined as the difference between the entropy of $S$ and the weighted mean of the entropies of the subsets $S_{\text{True}}$ and $S_{\text{False}}$ made of the elements of $S$ for which the splitting function is True and for which the splitting function is False respectively:

$$(13) \qquad \text{Gain}(S, s_{\Phi,\vartheta}) = H(S) - [\tfrac{n_{\text{True}}}{n} \times H(S_{\text{True}}) + \tfrac{n_{\text{False}}}{n} \times H(S_{\text{False}})],$$

where $n$, $n_{\text{True}}$ and $n_{\text{False}}$ are the numbers of elements in $S$, in $S_{\text{True}}$, and in $S_{\text{False}}$, respectively. The gain can be interpreted as encoding the information that would be gained by branching the node on the attribute $\Phi$ with threshold $\vartheta$. At each node, all features $\Phi$ and thresholds $\vartheta$ are tested and we select the splitting function that maximizes the gain. This leads to a new partition, for which child nodes are then analysed recursively in the same way.

Some parameters control the size and the complexity of the trees. We can specify maximal tree depth, the minimum number of elements required to split an internal node and to be at a leaf node. Such parameters appear as stop criteria in the tree growing process described above.

A weighting system can be used in order to favour one of the two labels. Such weighting intervenes in the calculation of the labels relative frequencies $p_{L_-}$ and $p_{L_+}$. If we denote $w_{L_-}$ and $w_{L_+}$ the weights respectively associated with labels $L_-$ and $L_+$, then the final relative frequency of each label is given by:

$$(14) \qquad p_{L_-} = \frac{w_{L_-} \times n_{L_-}}{(w_{L_-} \times n_{L_-}) + (w_{L_+} \times n_{L_+})} \, ; \, p_{L_+} = \frac{w_{L_+} \times n_{L_+}}{(w_{L_-} \times n_{L_-}) + (w_{L_+} \times n_{L_+})}.$$

## 4. Experimental setup

In this section, we describe the experimental setup to which the proposed classification method is applied. This setup includes the acquisition protocol, the validation dataset and ground truth establishment, the relevant tested methods and the performance measures.

### *4.1. Experimental protocol and image acquisition*

On the first day of the experiment, the individual fish eggs are manually placed in a 24-well plate, 1 egg per well, each well containing 2 mL of incubation medium with or without the studied chemical [33]. Medium replacements are performed every two days. After nine-days' exposure, 1.5mL of the incubation medium is removed from each well and the embryos are anaesthetized with Tricaine (0,18 g/L final). For each well, we record one photograph at a resolution of $1776 \times 2360$ pixels. More details are provided in [19].

### *4.2. Database description*

The database described in this manuscript has not been gathered with the aim of a thorough toxicological test, but for developing and testing computer programs. It means that each image was selected according to the presence or the absence of spine malformation that is to be automatically detected. The pictured alevins have been exposed to a wide variety of chemicals, including none. The nature of the chemical used is not significant in this study.

As seen in Section 2.2, feature characterization of our abnormality detection test depends on the alevin's skeleton representation on an orthonormal coordinate system. Such a representation implies that each abscissa is linked to a single ordinate. However, some alevins are not compatible with this graphic representation process and so the geometric description cannot be obtained. It can apply to some alevins that are so tightly wound that their spine form a hook (Figure 1f). To deal with these cases, alevins identified as such are directly labelled as having a hook-shaped spinal malformation without undergoing the learning-based classification.

Thus, in our validation process, several subsets of our datasets need to be considered. From a total dataset of 1,471 images of alevins (called "Dataset 0"), 12 are identified before feature extraction as being hook-shaped by the early malformation detection step of our program. The remaining dataset of 1,459 usable images (called "Dataset 1") constitutes the database used for the machine learning validation step. The datasets establishment process is summarized in Figure 7.

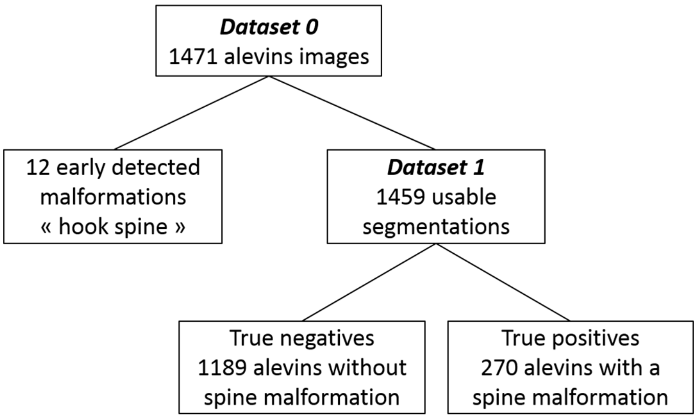

*Dataset 0*
1471 alevins images
12 early detected malformations « hook spine »
*Dataset 1*
1459 usable segmentations
True negatives
1189 alevins without spine malformation
True positives
270 alevins with a spine malformation

*Figure 7. Datasets description.*

### *4.3. Ground truth establishment*

On the day of image acquisition, each alevin is interactively observed under a microscope by an expert who manually and visually assesses the presence or the absence of any malformations. Interactive visual inspection using a microscope means that the alevin can be manipulated by the experts and thus observed from any relevant angle. Also, there is no discrete artefact due to image acquisition. This allows the operator to detect a malformation with high accuracy. For these reasons, this method is the most reliable way to assess whether an alevin has a morphological abnormality or not. It can be used to validate the automated classification method but also, more generally, to evaluate the quality of the complete alevin abnormalities detection assay, including plate preparation, data acquisition and data processing.

For our purpose, these microscope-based observations serve as ground truth. We focus on the expert observations that concern the presence or the absence of axial malformations. According to this ground truth and as it is shown in Figure 7, the dataset of 1,459 images contains 270 images of alevins with a spine malformation and 1,189 images of alevins without.

### *4.4. Tested classification methods*

This section introduces the details and the setup of the classification methods tested on the dataset and on the ground truth previously described in Sections 4.2 and 4.3. More precisely, we describe the setting of parameters presented in Section 2 as well as classification performed by an expert which is used for comparison purposes with the proposed automated method.

Since microscope-based observations are considered as ground truth for assessing axial deformations, it is necessary to point out that our proposed assay suffers from inherent limitations due to the 2D imaging acquisition system. Indeed, our data acquisition is restricted to a single 2D image, and so we observe one orientation only. Because some axial malformations are not visible from every point of view, it can happen that some abnormalities may not be detectable on the acquired images. As our automated classification (named $AC$) relies on image analysis, only considering the program misclassifications rate compared to ground truth does not paint the whole picture. To characterize the misclassification rate linked to data acquisition limitations, we compare our results with visual classification performed by an expert observing only 2D images. We term this "human classification" or $HC$. The following results of $AC$ and $HC$ are compared in Section 5.

The automated classifier parameters are set up as follow. All parameters described in Section 2 of this article are experimentally determined in order to optimize segmentation results. Segmentation and geometric parameters are listed in Table 2. To set up the classifier parameters described in Section 3 of this article, an implementation of the Iterative Grid Search algorithm is used that performs hyperparameter optimization by cross-validated grid-search over a specified parameters grid. We begin by defining a grid of parameters that will be searched during the process. Each grid parameter presents a range of test values. The algorithm exhaustively generates candidates from the specified parameters of this grid and fits the estimator on the whole dataset until finally retaining the best parameters combination. Manual specification of a limited set of hyperparameters reduces memory consumption during search. This method was used to set up the following parameters: the number of trees in the forest and the maximum depth of each tree are set to 30, the minimum number of samples required to split an internal node is set to 3 and the minimum number of samples required to be at a leaf node is set to 2. At each node, the quality of a split is measured with the entropy criterion presented in Section 3. In our program, we use the implemented algorithm GridSearch from the scikit-learn library [25].

| **Parameter name** | **Parameter description** | **Parameter value** |
|---|---|---|
| $r_1$ | Radius of $\Gamma_{r_1}$, the disk structuring element of the morphological closing $\varphi_{\Gamma_{r_1}}$ (Equation (2)) | 10 |
| $r_2$ | Minimal opening radius used for skeletonisation $S_{r_2+3.\mathrm{min}(5,\lambda)}$ (Equation (3)) | 14 |
| $\alpha$ | Minimal branch length used for skeleton pruning (Equation (4)) | 25 |
| $\sigma$ | Size of the convolution scaled window used for skeleton curve smoothing (Section 2.2.4) | 11 |
| $d_1$ | Minimal vertical distance that must separate two successive extrema to maintain them during spine approximation by a piecewise affine function (Section 2.2.4) | 4 |
| $d_2$ | Minimal horizontal distance required by a steady portion to be considered during spine approximation by a piecewise affine function (Section 2.2.4) | 10 |

*Table 2. Parameters determination for alevin spine segmentation and geometrical description of classification features.*

By testing different values for the weights $w_{L_-}$ and $w_{L_+}$ (see Equation (15)) associated with the negative positive dataset $L_-$ (non-malformed alevins) and to the true dataset $L_+$ (malformed alevins) respectively, we discovered that overall classification accuracy is stable. For 14 different weightings, overall accuracy varies by less than 1%. Since overall accuracy is essentially constant, given the screening nature of the assay, priority is given to specificity. In terms of methodology, that means minimizing the number of errors within the dataset $L_-$. It is equivalent with associating with the dataset $L_-$ the highest relative frequency $p_{L_-}$, which depends on both its number of data $n_{L_-}$ and the weight of each data $w_{L_-}$ as it is described in Equation 15, Section 3. According to the ground truth described in Section 4.3, the total database of 1,459 images contains 270 alevins with a spine malformation (positive dataset $L_+$) and 1,189 alevins without (negative dataset $L_-$). The relative frequencies are initially 80% for $L_-$ and 20% for $L_+$. In order to partially balance them, a higher weight value is given to the data of the sparsest sample $L_+$ than to the largest one $L_-$. Nevertheless, weighting remains in favour of dataset $L_-$ that is prioritized. The following weighting is chosen: 1 for the negative dataset $L_-$ and 2 for the positive dataset $L_+$. The following final relative frequencies are reached: 69% of negative data and 31% of positive data according to Equation 15.

Once all the model parameters are set up, the model can be trained. All features are gathered in a matrix and corresponding ground truths constitute a binary data vector used as true labelled data. Both are used as input for the training algorithm and the model is fitted as explained in Section 3.

### *4.5. Performance measurement*

In machine learning-based approaches, constructing a classifier involves optimizing its parameters on a predetermined training data sample with their associated labels. The classifier is then run on a test sample. In order to optimally use available data and minimize adverse training effects, we apply a cross-validation splitting strategy for our study. The basic k-fold approach is chosen [34]. During this process, the total database is split into k smaller equal-sized datasets. For each of the k consecutive iterations, the following procedure is applied: we train the model on k-1 subsets and then, we validate the resulting model on the remaining testing subset. As a result, at the end of the k iterations, results can be considered on the whole database, as the gathering of the results obtained on each testing data subset. Depending on the dataset size and thus the number of splits, cross-validation can suffer from bias and variance effects. When increasing the number of splits and therefore the size of the training sets, bias is reduced in the testing set, but we also reduce the number of test data so the output of the classifier is less certain. The variance of the classifier is thus said to be high. It is especially true if outliers happen to be selected in the limited testing set. On the contrary, the classifier has a lower variance by testing the model on more data. This implies a lower number of splits. In our method (called $AC$ for "automated classifier"), the parameter k is set to 10 as an acceptable trade-off between both bias and variance optimization. We ensure the data split in each dataset respects the proportions of malformed and non-malformed alevins previously described in Sections 4.2 and 4.3.

As for the human classifier ($HC$), the same cross-validation process cannot be applied, as it is not possible for the expert to forget what they have learned during a previous iteration. Iterations would not be independent. For this reason, expert results are obtained in a single run by observing the whole database. The optimistic assumption behind this is that human observations have inherent low bias.

For both methods, the results are presented in Section 5 in the form of confusion matrices. A confusion matrix [35] is defined as a classifier validation tool that represents distribution of correct and wrong classifications. Each column shows the number of occurrences for a predicted label whereas each line refers to the number of appearances of a true label. A predicted label is considered to be correct when it is the same as the true label according to the microscope-based ground truth (true negative $TN$ or true positive $TP$). Otherwise, it is considered to be incorrect (false negative $FN$ or false positive $FP$). See Table 3 for standard representation of a confusion matrix.

| Results<br>Ground truth | No axial malformation | Axial malformation |
|---|---|---|
| No axial malformation | $TN$ | $FP$ |
| Axial malformation | $FN$ | $TP$ |

*Table 3. Result presentation in the form of confusion matrix for the method under study. TN, TP, FN and FP respectively denote the true negative, the true positive, the false negative and false positive resulting with the considered method.*

Performance criteria are derived from this matrix. We calculate the percentages of true negatives, true positives, false positives and false negatives as follow:

(15) $$\text{specificity} = \text{true negative rate} = 100 \times (\frac{TN}{TN+FP}),$$

(16) $$\text{sensitivity} = \text{true positive rate} = 100 \times (\frac{TP}{TP+FN}),$$

(17) $$\text{FPR} = \text{false positive rate} = 100 \times (\frac{FP}{TN+FP}),$$

(18) $$\text{FNR} = \text{false negative rate} = 100 \times (\frac{FN}{TP+FN}).$$

We specifically call sensitivity the rate of true positives and specificity the rate of true negatives. According to these definitions, true negative and false positive rates amount to 100% and represent the totality of negative data in the dataset according to ground truth. Symmetrically, true positive and false negative rates also amount to 100% and represent the totality of positive data in the dataset according to ground truth.

For both classifiers $AC$ and $HC$, the percentage accuracy is measured from the accuracy score previously described in Section 1: $\text{accuracy percentage}(y, \hat{y}) = \text{accuracy}(y, \hat{y}) \times 100$. This scoring metric corresponds to the percentage of correct classifications among the total number of images in the database. It is also a performance criterion for the validation of our method.

## 5. Experimental results and discussion

Based on the setup described in the previous section, we present the results of the $AC$ and $HC$ methods. We assess their accuracy, before presenting the robustness, the quality control of early malformations detection and finally discussing our results.

### *5.1. Accuracy of the spine detection assay*

We now present the results of classifiers $AC$ and $HC$ compared to the microscope-based ground truths. A result is considered incorrect if it detects a spine malformation that is not present in the ground truth (false positive), or on the contrary, if it does not return a malformation when a spine abnormality is visible in the ground truth (false negative). Table 4(a) and (b) show the confusion matrices obtained for $AC$ and $HC$ respectively, on the 1,459 tested images of the database. Performance criteria are then derived from the confusion matrices and reported in Table 4(d).

For $AC$, we achieve a sensitivity of 40.4% and a specificity of 96%. False positive and false negative rates are 4.0% and 59.6% respectively. The corresponding percentage accuracy is 85.7%. For $HC$, a sensitivity of 47.4% and a specificity of 97.8% are measured, for a false positive percentage of 2.2% and a false negative ratio of 55.6%. The corresponding percentage accuracy is 88.5%. Without any model retraining, the results of $AC$ vs. $HC$ were also compared, leading to a third confusion matrix. In this case, accuracy is equal to 91.2%, FPR and FNR are equal to 5% and 40.1% respectively, and sensitivity and specificity are equal to 59% and 95% respectively.

| Classifiers<br>results<br>Ground truth | $AC$ | | $HC$ | |
|---|---|---|---|---|
| | No axial malformation | Axial malformation | No axial malformation | Axial malformation |
| No axial malformation | 1142 | 47 | 1163 | 26 |
| Axial malformation | 161 | 109 | 142 | 128 |

a

| $AC$ Results<br>$HC$ results | No axial malformation | Axial malformation |
|---|---|---|
| No axial malformation | 1240 | 65 |
| Axial malformation | 63 | 91 |

b

| Performance criterion $f$ | $f_{AC}$ | $f_{HC}$ | $f_{AC\ vs\ HC}$ |
|---|---|---|---|
| Specificity (%) | 96.0 | 97.8 | 95.0 |
| Sensitivity (%) | 40.4 | 47.4 | 59.0 |
| False Positive (%) | 4.0 | 2.2 | 5.0 |
| False Negative (%) | 59.6 | 52.6 | 41.0 |
| Accuracy (%) | 85.7 | 88.5 | 91.2 |

c

*Table 4. Results obtained by the automated classifier $AC$ and the human classifier $HC$ on the complete database of 1,459 images. The tables represent the confusion matrices of alevins with and without a spine malformation according to the $AC$ and $HC$ results compared to the microscope-based ground truth after 10-fold cross validation in a, the confusion matrix of $AC$ vs. $HC$, without any retraining in b, and the classifier comparison metrics in c.*

It can be seen, for both the $AC$ and $HC$ classifiers, that specificity is maximized. On the other hand, we can see that sensitivity is low for both classifiers. Taking human observations as a gold standard, the error metrics of $HC$ gives an insight into the amount of information loss between interactive observations under a microscope and what is achievable using only 2D images. The overall accuracy of $HC$ is 88.5%, which is quite high. This result suggests that spine deformation can be detected with an acceptable accuracy from 2D images only, which has considerable implications for the automation of this test. Moreover, specificity is high, meaning very few false deformations are detected (2.2%). Concerning $AC$, very similar results are observed, when compared to the human observer, with an accuracy of 85.7%. This comforts us in the intermediate conclusion that automating the spine deformation assay is indeed feasible. The FPR of $AC$ is 4.0%, which is twice as much as the human observer but is still acceptable. The comparison of $AC$ vs. $HC$ shows an accuracy of 91.2%. This can be interpreted as saying that humans and computers do not make exactly the same mistakes but that they make them in similar numbers. In particular, $AC$ agrees in 95% of the cases when $HC$ detects no axial deformation, and $AC$ agrees in 59% of the cases when $HC$ does detect an axial deformation. This latter number may seem low, but axial deformations are relatively uncommon, so overall few errors are made. True negatives, true positives and a false positives of $AC$ results are illustrated in Figure 8.

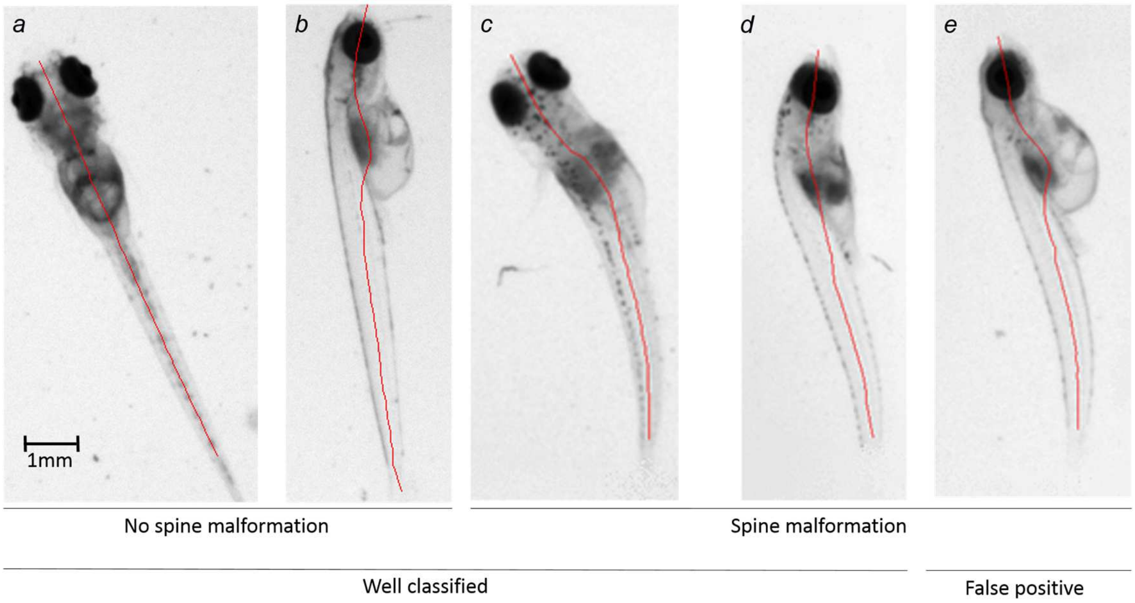


Figure 8. Results of alevin spine classification. The red line represents the result of the spine segmentation. The method leads to proper classification (a and b: no spine malformation; c and d: spine malformation) or to a false positive (e: false detection of a spine malformation). a and c are presented in dorsal view while b, d and e are presented in lateral views.

### *5.2. Robustness of the method and time efficiency*

With machine learning, the results of classification models currently vary depending on the partitioning data selected to train and test the model. Thus, assessing the robustness of our model means estimating the variability in the performance criteria obtained for several successive iterations of training and testing steps made on randomly determined splitting. For our purpose, two aspects are considered. Through the 10 iterations of the cross-validation, 10 different estimators are built and tested on 10 different subsets that do not overlap. We begin by testing the variance of the models results by calculating the standard deviation of the percentage accuracy. In our experiments, the $AC$ percentage accuracy varies between 81.5% and 91.0%, for an average of 85.7% over the 10 iterations and a standard deviation $\sigma_X$ of 2.6. Such a low variability is acceptable.

For scaling up, close attention is paid to analysing the change in the program results over 100 new 10-fold cross-validations. Each time, a new partitioning is made, splitting the total dataset into 10 subsets and a new cross-validation is applied. The corresponding true negative, true positive, false positive and false negative ratios are calculated according to the Equations 16 to 19. All the ratios were remarkably stable and argues that the cross validation principle applied in this validation process minimizes the partition's influence on the results.

### *5.3. Quality control of early data sorting*

As previously explained in Section 4.2, some images were excluded before applying the spine malformation detection test. On the dataset of 1,459 images, 12 are detected early as not being representative of our method on a direct orthonormal system due to the presence of a hook in the spine. However, referring to our ground truth, only 4 of them actually present a hooked spine. The other 8 cases detected were therefore wrongly excluded from the learning-based classification process due to the presence of impurities in the well that causes alevin segmentation errors during pre-processing. Segmentation improvements in pre-processing are worth taking into consideration, but were not implemented yet since the resulting improvement would be insignificant when taking into account the whole dataset (around 0.5%).

### *5.4. Inter-expert variability*

This last part of our study concerns inter-expert variability on a single data subset due to subjectivity. Indeed, as for microscope and for image-based observations, annotations from several experts can differ from each other. Several reasons can explain this fact, including operator fatigue and degree of expertise. For a single dataset observed by a unique operator, results can also differ depending on the data previously observed. For instance, a malformed alevin can appear healthy for an operator who previously saw an important number of highly abnormal alevins. On the contrary, when comparing to healthy alevins, an expert can sometimes interpret a slight curve due to natural positioning on the well as a malformation. For these reasons, quantifying inter-expert subjectivity is considered to be relevant. Practicality aspects make the assessment complicated to perform on microscope observations. As the latter can take place only on the day of data acquisition, they require the presence of several available experts on the same day, unlike images that can be registered and analysed later. For this reason, our inter-expert assessment is performed on 2D images. Among the 1,459 images annotated by our main expert, named Expert 1, a subset of 200 images was annotated by two additional experts, named Expert 2 and Expert 3. In this subset, the 2D observations of Expert 1 exactly match those made under the microscope. In this sense, we can consider this dataset as non-ambiguous. On such a dataset, we could reasonably expect Experts 2 and 3 to concur with the microscope. However, we note in Table 5(b) that Experts 2 and 3 recorded errors at a respective rate of 11.5% and 5.5%. This is comparable with the 8.0% percentage error by the proposed automated method on this data.

The subjectivity rate is defined as the percentage of images on which experts disagree. In our case, discrepancies are observed on 28 images, for a subjectivity rate of 14%. In addition, nearly all discrepancies are false positives. This rate is close to the programed error rate of 14.5% calculated on the whole database. This observation enables us to argue that operator subjectivity is a significant problem, which in particular may call into question the reliability of our ground truth. In addition, on the 200 data sample, we note that the error rate of our proposed method is 8.0%, which is in between the Experts 2 and 3 respective error rates of 5.5% and 11.5%. We also note that the results distribution in Table 5(a) shows that errors made by the program are more balanced between false positives and false negatives. These results can be considered to be acceptable.

| Results<br>Ground truth | Expert 2 | | Expert 3 | | Program | |
|---|---|---|---|---|---|---|
| | No axial malformation | Axial malformation | No axial malformation | Axial malformation | No axial malformation | Axial malformation |
| No axial malformation | 154 | 23 | 167 | 10 | 169 | 7 |
| Axial malformation | 0 | 23 | 1 | 22 | 9 | 15 |

a

| | Expert 2 | Expert 3 | Program |
|---|---|---|---|
| Percentage errors between expert observations and ground truths on the 200 data samples | 11,5 | 5,5 | 8,0 |

b

*Table 5. Results and error rates obtained for each observer and for the automated classifier versus the microscope-based ground truths during subjectivity assessment on a sample of 200 images.*
*a: distribution of alevins with and without a spine malformation according to the results of Experts 2 and 3 compared to the microscope-based ground truths. We report in b the percentage error calculated for each expert and for the automated classifier on this 200 data sample.*

### *5.5. Execution time*

The program is executed on a standard computer with a 3.60 GHz Intel® Core™ i7-4790 CPU and 32 GB of RAM. The classifier training step can be repeated as much as necessary in about one second. Our program then classifies an image in only about 1 second, meaning that the test can be performed at the same time as the next well image is being acquired.

## ***6. Discussion***

This work aimed to develop an automated image processing-based assay for the detection of spine malformations in Medaka alevins. Since a screening test was the target, the emphasis was put on the overall accuracy of the test. As shown in Section 5.1, we reached our objective by achieving a false positive rate of only 4% and a total accuracy of 85.7%. Nevertheless, optimizing overall accuracy first and specificity second inevitably implies lowering sensitivity, which is defined as the assay's ability to correctly detect a malformation. In our assay, only 40% of the actual spine malformations are detected according to what is visible under a microscope. Since $HC$ results are a little better at 47%, this seems to imply that many of these kinds of deformation cannot always be reliably detected from 2D images. To improve this, better acquisition devices would be needed, or more simply, experiments could be repeated or other deformation tests used. Eventually, the proposed assay is intended to be made part of a series of abnormality detection programs (including eyes, oedemas and swim bladder abnormalities) that could improve the sensitivity of the whole detection assay. Thus, in spite of these shortcomings, this program remains relevant and useful as a screening tool with regard to its high specificity.

Several methodologies have been published in the context of alevin spinal cord analysis using image processing. Most were conducted on zebrafish embryos. Stegmaier *et al* assessed the development of specific neuron population by extracting a quantitative information from fluorescent proteins labelled spinal cord neurons in transgenic zebrafish [18]. Al-Saaidah *et al* described an automatic system for the detection of abnormal curvature zebrafish tail. However, the study is limited to the classification of obvious abnormalities in tail curvature (up or down) [17]. Jeanray *et al* proposed a way to classify multiple zebrafish phenotypes, including tail abnormalities, by applying supervised machine learning. This approach does not need features characterization as it is based on the extraction of dense random subwindows their description in raw pixel values and classification by extremely randomized tree. If the study shows result with a good correlation with that from experts on nine different zebrafish phenotypes, the error rates do not take into account the information loss from manual observations under microscope to those on 2D images, as every ground truth is obtained by looking directly on acquired images. In particular, in these two latest studies, the analysis is limited to the detection of defects specifically visible on the lateral side of the zebrafish, that implies to pay a particular attention to embryo positioning [16]. Contrary to these techniques, the methodology proposed in this article relies on a simple experimental setup, compatible with the high-throughput screening related constraints. The day of image acquisition, each alevin remains in its growing medium and the image is recorded without manual positioning of the alevin, minimizing human manual intervention. The test is then based on a morphological analysis of the alevin on brightfield images, and was validated on more than 1400 images.

## 7. Conclusion and perspectives

In this article, a fast and automated procedure was proposed to detect malformations in the spinal cord of Medaka alevins with minimal operator interaction, maximum speed and reliability. The objective of this procedure is to devise an image-based waterway pollution and toxicology assay. Based on mathematical morphology, our image-processing pipeline best approximates the spine of alevins in order to extract representative features. Based on these, a Random Forest model is trained to detect the presence or the absence of a spine malformation. This work illustrates the main difficulties linked to ground truth definition and the limitations of the data acquisition device to obtain a reliable automated process.

The main contributions of this article are the following:

- A dataset of 1,459 alevin images with associated ground-truth was constructed. To establish the ground truth, each alevin was screened interactively under a microscope by a trained operator prior to imaging, and the presence or absence of a malformation was carefully recorded.
- Separately, the presence or absence of a malformation was assessed by a different, trained operator on the resulting 2D alevin images. The information loss leading to erroneous deformation estimations due to the interactive 3D to fixed 2D images transition was investigated.
- Additional observations and labelling were performed by three experts on a subset of 200 images, to establish inter-expert subjectivity on 2D images;
- An image analysis pipeline consisting essentially of mathematical morphology operators for characterizing alevins malformations by developing a number of relevant descriptors was proposed;
- We have shown that alevins can be classified automatically as normal or deformed using the proposed descriptors and a Random Forests classifier. We have shown that our RF classifier can reach accuracy similar to image-based human classification and with time efficiency (about one second to process each image) that is compatible with its use in a high throughput industrial context.

The resulting alevin deformation image-based assay is intended to be a part of a more comprehensive morphological and functional abnormality detection test. For such a context, at the highest possible level of overall precision, it is more important that individual tests be specific rather than sensitive, since the sensitivity of the global assay is likely to increase by analysing other criteria, such as eye abnormalities, absence of swim bladder or presence of oedemas. Consequently, our assay proposal is designed with this in mind and favours specificity, as described in the main text.

As for future work, quantitative studies of the cardiovascular system (heart frequency, blood flow estimation etc.) as well as the detection of swim bladder deformations, eye deformations and oedema are currently undergoing development. This method can also be transferred to the analysis of zebrafish organisms, also of interest to toxicological studies.

## Acknowledgments

We thank the Association Nationale de la Recherche et de la Technologie (ANRT) for providing the CIFRE financing that allowed us to produce this work.

## References

[1] "Directive 2010/63/EU of the European Parliament and of the Council of 22 September 2010 on the protection of animals used for scientific purposes," *Official Journal of the European Union*, pp. 33–79, 2010.

[2] "Regulation (EC) No 1223/2009 of the European Parliament and the Council of 30 november 2009 on cosmetics products," *Official Journal of the European Union*, pp. 59–209, 2009.

[3] Marlies Halder, "Regulatory Aspects on the Use of Fish Embryos in Environmental Toxicology," *Integr. Environ. Assess. Manag.*, vol. 6, pp. 484–491, 2010.

[4] U. Strähle, "Zebrafish embryos as an alternative to animal experiments - A commentary on the definition of the onset of protected life stages in animal welfare regulations," *Reproductive Toxicology*, pp. 128–132, 2011.

[5] C.B. Lovely, Y. Fernandes, and J.K. Eberhart, "Fishing for Fetal Alcohol Spectrum Disorders: Zebrafish as a Model for Ethanol Teratogenesis," *Zebrafish*, vol. 13, no. 5, pp. 391–398, 2016.

[6] A. Jaja-Chimedza, K. Sanchez, M. Gantar, P. Gibbs, M. Schmale, and J.P. Berry, "Carotenoid glycosides from cyanobacteria are teratogenic in the zebrafish (Danio rerio) embryo model," *Chemosphere*, pp. 478–489, 2017.

[7] M. R. Embry *et al.*, "The fish embryo toxicity test as an animal alternative method in hazard and risk assessment and scientific research.," *Aquatic Toxicology*, pp. 79–87, 2010.

[8] R. (Nagel), "DarT: The Embryo Test with the Zebrafish Danio rerio - a General Model in Ecotoxicology and Toxicology," *Altex*, pp. 38–48, 2002.

[9] S.E. Belanger, E.K. Balon, and J.M. Rawlings, "Saltatory ontogeny of fishes and sensitive early life stages for ecotoxicology tests," *Aquatic Toxicology*, pp. 88–95, 2009.

[10] E.K. Balon, "Types of feeding in the ontogeny of fishes and the life history model, Environ. Biol. Fishes.," *Environmental Biology of Fishes*, pp. 11–24, 1986.

[11] L.V. Dishaw, D.L. Hunter, B. Padnos, S. Padilla, and H.M. Stapleton, “Developmental Exposure to Organophosphate Flame Retardants Elicits Overt Toxicity and Alters Behavior in Early Life Stage Zebrafish (Danio rerio),” *Toxicological Sciences*, pp. 445–454, 2014.

[12] A. Yamashita, H. Inada, K. Chihara, T. Yamada, J. Deguchi, and H. Funabashi, “Improvement on the evaluation method for teratogenicity using zebrafish embryos,” *J. Toxicol. Sci.*, vol. 39, no. 3, pp. 453–464, 2014.

[13] S. Xia, Y. Zhu, X. Xu, and W. Xia, “Computational techniques in zebrafish image processing and analysis,” *J. Neurosci. Methods*, vol. 213, no. 1, pp. 6–13, 2013.

[14] Pylatiuk C. *et al.*, “Automatic Zebrafish Heartbeat Detection and Analysis for Zebrafish Embryos,” *Zebrafish*, 2014.

[15] M. Schutera *et al.*, “Automated phenotype pattern recognition of zebrafish for high-throughput screening,” *Bioengineered*, vol. 7, pp. 261–265, 2016.

[16] N. Jeanray *et al.*, “Phenotype Classification of Zebrafish Embryos by Supervised Learning,” *Plos One*, 2015.

[17] B. Al-Saaidah, W. Al-Nuaimy, M. Al-Taee, I. Young, and Q. Al-Jubouri, “Identification of Tail Curvature Malformation in Zebrafish Embryos,” in *8th International Conference on Information Technology (ICIT)*, Toronto, Canada, 2017, pp. 588–593.

[18] J. Stegmaier *et al.*, “Automated prior knowledge-based quantification of neuronal patterns in the spinal cord of zebrafish,” *Bioinformatics*, vol. 30, no. 5, pp. 726–733, 2014.

[19] E. Puybareau, D. Genest, E. Barbeau, M. Léonard, and H. Talbot, “An automated assay for the assessment of cardiac arrest in fish embryo,” *Comput. Biol. Med.*, pp. 32–44, 2017.

[20] M. Couprie and G. Bertrand, “Discrete Topological Transformations for Image Processing,” in *Digital Geometry Algorithms*, vol. 2, Springer, 2012, pp. 73–107.

[21] R. Kresch and D. Malah, “Skeleton-Based Morphological Coding of Binary Images,” *Transaction in Image Processing*, pp. 1387–1399, 1998.

[22] L. Breiman, “Random forests,” *Mach. Learn.*, vol. 45, no. 1, pp. 5–32, 2001.

[23] L. Najman and H. Talbot, Eds., *Mathematical Morphology: from theory to applications*. UK, London: ISTE-Wiley, 2010.

[24] M. Courpie, L. Marak, and H. Talbot, “Pink image processing library,” presented at the 4th European meeting on Python in Science (Euroscipy), Paris, 2011.

[25] S. Van der Walt *et al.*, “The scikit image contributors : Scikit image : Image processing in Python,” *PeerJ*, 2014.

[26] L. Breiman, “Bagging Predictors,” *Mach. Learn.*, vol. 24, pp. 123–140, 1996.

[27] S. Dudoit and J. Fridlyand, “Bagging to improve the accuracy of a clustering procedure,” *Bioinformatics*, pp. 1090–1099, 2003.

[28] T. Hothorn and B. Lausen, “Double-bagging: combining classifiers by bootstrap aggregation,” *Pattern Recognit.*, pp. 1303–1309, 2003.

[29] J. Serra, *Image analysis and mathematical morphology*. Academic Press, 1982.

[30] T. Hirata, “A unified linear-time algorithm for computing distance maps,” *Information Processing Letters*, pp. 129–133, 1996.

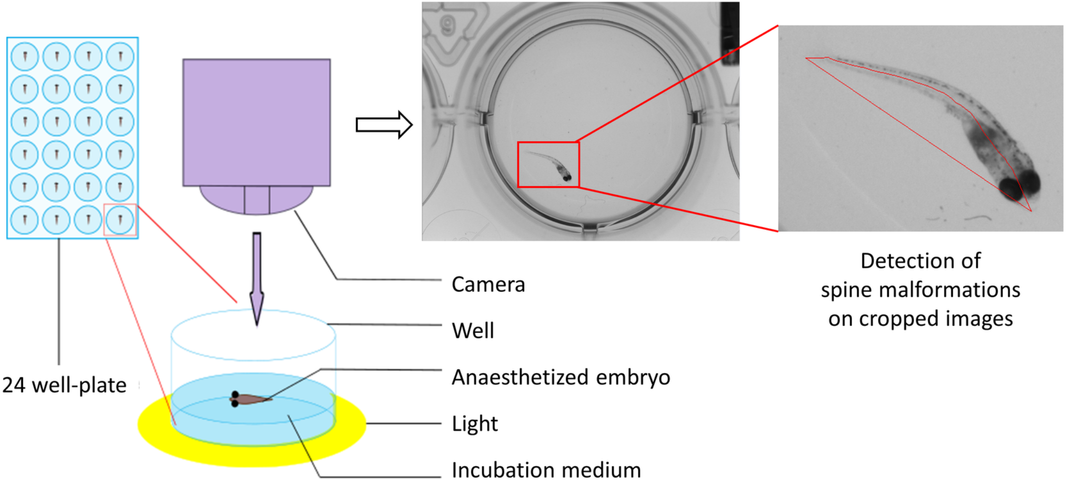
[31] A. Meijster, J.B.T.M. Roerdink, and W.H. Hesselink, “A General Algorithm for Computing Distance Transforms in Linear Time,” in *Mathematical Morphology and its Applications to Image and Signal Processing, Computational Imaging and Vision*, 2000, vol. 18, pp. 331–340.

[32] N.H. Jones, “Finding the area under the curve using JMP and a trapezoidal rule,” SAS Institute, Cary, NC, 1997.

[33] M. Kinoshita, K. Murata, K. Naruse, and M. Tanaka, *Medaka Biology, Management, and Experimental Protocols*, Wiley-Blackwell. 2012.

[34] J.M. Hancock, M.J. Zvelebil, and N. Cristianini, “Cross-Validation (K-Fold Cross-Validation, Leave-One-Out, Jackknife, Bootstrap),” in *Dictionary of Bioinformatics and Computational Biology*, Wiley Online Library., 2014.

[35] “Using the confusion matrix for improving ensemble classifiers,” presented at the IEEE 26-th Convention of Electrical and Electronics Engineers in Israel, 2010, pp. 555–559.


24 well-plate
Camera
Well
Anaesthetized embryo
Light
Incubation medium
Detection of
spine malformations
on cropped images