Imaging of the fish embryo model and applications to toxicology

Diane Genest

LABORATOIRE D'INFORMATIQUE
GASPARD-MONGE

Sous la co-tutelle de :
CNRS
ÉCOLE DES PONTS PARISTECH
ESIEE PARIS
UPEM • UNIVERSITÉ PARIS-EST MARNE-LA-VALLÉE

UNIVERSITÉ PARIS-EST

Université Paris-Est

Ecole doctorale Mathématiques et Sciences et Technologies de l'Information et de la Communication

Thèse presentée pour l'obtention du titre de

Docteur de l'Université Paris Est

Spécialité Informatique

Imaging of the Fish Embryo Model and Applications to Toxicology

par

Diane GENEST

soutenue publiquement le 20/05/2019 devant le jury composé de :

President of the jury	Cédric Wemmert	Enseignant-chercheur, Université de Strasbourg
Directeur de thèse	Hugues Talbot	Enseignant-chercheur, Centrale Supelec
Directeur de thèse	Jean Cousty	Enseignant-chercheur, ESIEE Paris
Superviseur	Noémie de Crozé	Ingénieure Recherche Avancée, L'Oréal Recherche et Innovation
Rapporteur	Jesús Angulo	Maître de Recherche, Les Mines Paritech - CMM
Rapporteur	Stefan Scholz	Chercheur, Helmholtz Center for Environmental Research - UFZ
Examinateur	Annelii Ny	Innovation Manager, Laboratory for Molecular Biodiscovery, KU Leuven

Acknowledgements

I would like to start by thanking **Dr. Jesús Angulo** and **Dr. Stefan Scholz** for accepting to referee my manuscript and for being part of my defence committee. I also thank **Dr. Annelii Ny** and **Dr. Cédric Wemmert** for being part of my defense committee. It is a great honor for me to receive feedback of my work from such respected members of the research community.

Je tiens à remercier mes encadrants, en commençant par mon ancien professeur, devenu mon directeur de thèse **Hugues Talbot**, avec qui tout a commencé au cours de mes études d'ingénieur, et grâce à qui j'ai découvert cette discipline. Merci à mon second directeur de thèse **Jean Cousty** pour son suivi, ses précieux conseils, ses idées et surtout sa disponibilité au cours de cette thèse. Merci à **Noémie De Crozé** pour son encadrement, sa sympathie, son professionnalisme et ses précieuses connaissances qui m'ont permis d'en apprendre plus dans les domaines de la toxicologie et de l'embryologie.

Un grand merci à **Marc Léonard** sans la confiance de qui ce projet de thèse ne se serait pas concrétisé. Merci de m'avoir permis d'intégrer ton équipe.

Je remercie tous les membres du REDD (Recherche Environnementale et Développement Durable), et en particulier son directeur **Laurent Gilbert** pour m'avoir accueillie dans son département.

Cette expérience a été l'occasion pour moi d'intégrer une équipe formidable au sein de l'équipe de Recherche Environnementale de L'Oréal. Merci à tous ses membres qui ont contribué à faire de chaque jour un moment de convivialité et de bonne humeur. **Noémie, Claire** et **Typhaine** : mon équipe, la fameuse équipe poisson ! **Cyril** et **Laurent** : pour ces blagues et moments de dérision qui ont pimenté et égayé les petits déj, et pas que ! **Aurore, Inès, Norma** et **Vincent** : pour ces rigolades, et l'enthousiasme des afterworks passés ou à venir. **Fabienne** et **Delphine** : pour votre douceur et votre gentillesse qui étaient toujours des réconforts. **Corinne** : pour tes attentions et notre super entente. **Patricia** : pour nos super séances sportives. **Catherine** : pour ta bonne humeur au quotidien. Merci également à **Elsa**, **Jacques**, **Jean-Charles**, **Jean-Pierre**, **Laurène**, **Sacha**, mais aussi **Delphine** et **Laurent**, plus récemment arrivés.

Un merci vraiment tout particulier à mon acolyte de toujours **Typhaine**, sans qui l'aventure de la thèse n'aurait certainement pas été la même. Comme dirait Noémie, ce sont des sortes de clônes qui se sont rencontrés pendant cette thèse : les mêmes goûts sur à peu près tout, une entente démesurée, des moments de complicité irremplaçables.

Je tiens également à remercier les personnes grâce auxquelles l'aventure L'Oréal va continuer : merci à **Baptiste Boudard** ainsi qu'à ma nouvelle équipe, en particulier **Rui Pereira** et **Ismaël Bounouard**. Ismaël, j'ai énormément apprécié ta disponibilité et nos échanges qui m'ont permis de me projeter dans ce nouveau cadre. J'ai hâte de commencer cette nouvelle aventure !

Côté ESIEE également, les rencontres ont été nombreuses. Le Laboratoire d'Informatique Gaspard Monge a vu passer de nombreux visages de personnes en stages, CDD, thèses, sur des projets de tout type. Des liens ont été tissés. Je remercie les récemment partis ou encore présents : **Clara**, **Julie**, **Sylvain**, **Bruno**, **Stefan**, **Ketan**, et les plus anciens avec qui cette aventure a commencé : **Elodie**, **Odyssée**, **Eloïse**, **Laurent**, **Matthieu**, **Geoffrey**, **Ali**, **Yann**. Je garde un merveilleux souvenir des moments passés avec vous. Le prochain remerciement va devoir traverser l'océan car il va à **Thibault**, loin là-bas en Amérique. Le bureau était quand même un peu vide sans mon acolyte de l'ESIEE durant cette dernière année, mais quels bons souvenirs je garde de nos moments de bonne humeur (et de bonne boustifaille) partagés !

Merci à **mes parents** d'être présents pour moi à ma soutenance. Votre présence me touche énormément et je suis fière de pouvoir vous donner un aperçu de mon travail et de ce qu'ont été pour moi ces trois dernières années.

Enfin, je remercie mon fiancé et amour **Guillaume** pour son soutien indéfectibe tout au long de cette expérience. Certains moments auraient certainement été beaucoup plus durs sans ta chaleur, tes conseils et ton réconfort.

Abstract

Numerous compounds are used and produced in the world for various applications. Aside from the assessment of their efficacy, industries that produce these chemicals have to assess the safety of their chemicals for human. Toxicological assessment of these compounds is performed to reveal their potential toxic effect. Among the potential toxicities that need to be detected, the developmental toxicity (teratogenicity), meaning the chemical ability to provoke abnormalities during the embryonic development, is crucial. Moreover, in accordance with the Russel and Burch's 3Rs rule that recommends to Replace, to Reduce and to Refine tests performed on laboratory animals, more and more industries are interested in developing alternatives to animal testing for the toxicological assessment of compounds. In compliance with the European regulation that forbids the use of animal testing for the safety assessment of cosmetics, the toxicological assessment of chemicals must rely on a series of techniques including *in silico* and *in vitro* assays. Assays performed on alternative models are also required to replace the regulatory *in vivo* tests made on laboratory animals. For now, no alternative method has been validated in the field of developmental toxicology. The development of new effective alternative methods is thus required. Furthermore, the use of most cosmetics and personal care products inevitably leads to their rejection in waterways after washing and rinsing. This results in the exposition of some aquatic environments (surface waters and coastal marine environments) to chemicals included in cosmetics and personal care products. Thus, environmental toxicological assessment of cosmetics and of their ingredients is also necessary, which requires the knowledge of their toxicity on organisms that are representative of aquatic food chains.

In this context, the fish embryo model, considered ethically acceptable for the toxicological assessment of cosmetics according to international regulations, presents a dual advantage for the cosmetics industry. First, as a model representative of aquatic organisms, it is useful for assessing the environmental impact of chemicals. Second, as a vertebrate, key mechanisms of the embryonic development are conserved between fish and human, making this model promising for the assessment of the teratogenic effects of chemicals on human.

In this dissertation, a test is presented for the assessment of the chemicals teratogenic potential based on the analysis of exposed medaka fish embryos (*Oryzias latipes*). This test relies on the calculation of a teratogenicity index, which is the ratio between two indices: LC50 is the concentration which is lethal for 50% of embryos, and EC50 is the concentration that causes an effect in 50% of the embryos, including malformations and lethality. LC50 calculation is based on a binary classification between alive and dead embryos. Similarly, a classification between healthy and malformed embryos leads to EC50 calculation. The final teratogenicity index allows to draw a conclusion on the teratogenic effect of the chemical. Currently, this test is manually performed, meaning it relies on observations of embryos made under a stereomicroscope and annotations of the embryos anomalies. This process is time-consuming and prone to error, as observations coming from several operators may differ. Computerized classification methods could help to improve the efficacy and the objectivity of the teratogenicity test. Thus, the objective of this project is to automate the test, by using image processing and machine learning classification methods (random forest).

All methods developed during this project rely on (i) the identification, from image or video, of a region of interest which depends on the anomaly we want to detect, (ii) image or video characterization, *i.e.*, extraction of features which are representative of the anomaly, and (iii) automated classification of embryos according to features analysis. A first method is developed aiming at automatically detect embryo heartbeats from short video of embryos. This detection is based on the time analysis of the pixel intensity variation in the heart region. An accuracy of 98,5% is obtained compared to videos observation. A second method is developed to detect malformations in the spine of embryos after hatching (eleutheroembryo). The method automatically extracts morphological features from images and uses a random forest classifier to perform the classification. This leads to an accuracy of 85% compared to the gold standard of interactive microscope-based observations. Finally, a third method also relies on a random forest classifier to classify embryos according to the presence or the absence of a swim bladder, that leads to 95% of accuracy compared to the gold standard. These three assessments demonstrate the feasibility of automatically performing the functional and morphological assessment of embryos, and to validate the relevance of the used features for characterizing the studied anomalies.

Résumé

De nombreuses substances chimiques sont produites et utilisées dans le monde pour des applications diverses. En dehors de la nécessité d'évaluer leur efficacité, l'industrie se doit surtout d'évaluer la sécurité de leurs substances pour l'humain. L'évaluation toxicologique des substances chimiques est réalisée dans le but de révéler un potentiel effet toxique de la substance testée. Parmi les effets potentiels que l'on doit détecter, la toxicité du développement (tératogénicité), c'est-à-dire la capacité d'une substance à provoquer des anomalies lors du développement embryonnaire, est fondamentale. De plus, en accord avec la règle des 3R de Russel et Burch qui recommande de Remplacer, Réduire et Raffiner les tests sur animaux de laboratoires, de plus en plus d'industries s'intéressent au développement de nouvelles méthodes alternatives à l'expérimentation animale pour l'évaluation toxicologique des produits chimiques. Conformément à la législation européenne qui interdit à l'industrie cosmétique d'avoir recours à des tests sur animaux de laboratoire pour l'évaluation toxicologique de leurs substances, cette évaluation se base sur les résultats de tests *in silico* et *in vitro*. Des tests développés sur modèles alternatifs sont également requis pour remplacer les tests réglementaires *in vivo* réalisés sur animaux de laboratoire. Pour le moment, aucune méthode alternative n'a été validée d'un point de vue réglementaire pour évaluer la toxicité du développement. Le développement de nouvelles méthodes alternatives s'avère donc nécessaire. D'autre part, l'usage de la plupart des produits cosmétiques et d'hygiène corporelle conduit, après lavage et rinçage, à un rejet à l'égout et donc dans les cours d'eau. Il en résulte que les environnements aquatiques (eaux de surface et milieux marins côtiers) sont parfois exposés aux substances chimiques incluses dans les formules cosmétiques. Ainsi, l'évaluation toxicologique environnementale des cosmétiques et de leurs ingrédients est également nécessaire. Celle-ci nécessite de connaître leur toxicité sur des organismes représentatifs de chaînes alimentaires aquatiques.

Dans ce contexte, le modèle embryon de poisson, considéré éthiquement acceptable par la législation pour l'évaluation toxicologique des produits cosmétiques, présente un double avantage pour l'industrie cosmétique. Premièrement, ce modèle est représentatif des organismes aquatiques. Il est donc pertinent pour évaluer la toxicité environnementale des substances chimiques. Deuxièmement, en tant que vertébré, les mécanismes clefs du

développement embryonnaire sont conservés entre le poisson et l'humain, faisant de l'embryon de poisson un modèle prometteur pour évaluer l'effet tératogène de substances chimiques chez l'humain.

Ce manuscrit présente un test d'évaluation de la tératogénicité de substances chimiques, basé sur l'analyse d'embryons de medaka (*Oryzias latipes*) exposés à des substances. Ce test repose sur le calcul d'un indice tératogène, qui est le ratio de deux indices : la CL50 est la concentration létale pour (qui provoque la mort de) 50% des embryons exposés, et la CE50 est la concentration qui cause un effet chez 50% des embryons exposés, ce qui inclut les malformations et la létalité. Le calcul de la CL50 est basé sur une classification binaire des embryons morts et vivants. Le calcul de la CE50 se base sur une classification binaire des embryons selon qu'ils présentent ou non une malformation. L'indice tératogène permet de tirer une conclusion quant à l'effet tératogène de la substance testée. Pour le moment, ce test est réalisé de façon manuelle, ce qui implique qu'un opérateur observe les embryons sous une loupe binoculaire et les annotent en fonction de leurs potentielles anomalies. Ce processus est long et sujet à erreur, dans le sens où différents opérateurs peuvent ne pas avoir les mêmes observations sur un même embryon. Des procédures automatisées pourraient aider à améliorer l'efficacité et l'objectivité du test d'évaluation de la tératogénicité. Ainsi, l'objectif de ce projet est d'automatiser le test en ayant recours à des procédures de traitement d'images et de classification par apprentissage automatique (forêts aléatoires).

Les méthodes développées durant ce projet reposent toutes sur (i) l'identification d'une zone d'intérêt spécifique de l'anomalie que l'on cherche à détecter, (ii) la caractérisation des images et vidéos par extraction de descripteurs caractéristiques de l'anomalie que l'on cherche à détecter, et (iii) la classification automatique des embryons basées sur l'analyse de ces descripteurs. Une première méthode est développée et sert à détecter automatiquement les battements cardiaques d'embryons de medaka à partir de courtes séquences vidéos. La détection des battements cardiaques repose sur l'analyse de la variation d'intensité de pixels dans la zone du cœur et permet de classer les embryons en vivants et morts. On obtient un taux de classifications correctes de 98,5% comparé aux observations faites sur les vidéos. Une seconde méthode est développée pour détecter les malformations axiales des embryons après éclosion (alevins vésiculés). Des descripteurs morphologiques représentatifs de cette anomalie sont extraits des images et utilisés par un classificateur de type forêt aléatoire pour classer les images. Il en résulte un taux de classifications correctes de 85% par rapport aux observations des alevins faites au microscope, qui constituent la norme de référence. Enfin une troisième méthode repose également sur l'utilisation d'un classificateur de type forêt aléatoire pour classer les alevins selon qu'ils présentent ou non une vessie natatoire. Le taux de succès de la classification est de 95% comparé à la norme de référence. Ces trois évaluations permettent de mettre en évidence la faisabilité de l'automatisation et de valider la pertinence des descripteurs utilisés pour caractériser chacune de ces anomalies.

Publications

Parts of this project have appeared in the following publications:

International journal

E. Puybareau, D. Genest, E. Barbeau, M. Leonard, H. Talbot. “An automated assay for the assessment of cardiac arrests in fish embryo”, in *Computer in Biology and Medicine*, pp 32-44, 2017.

D. Genest, E. Puybareau, M. Léonard, J. Cousty, N. De Crozé, H. Talbot. “High throughput automated detection of axial malformations in Medaka embryo”, in *Computer in Biology and Medicine*, pp 157-168, 2018.

International conferences

E. Puybareau, D. Genest, N. de Crozé, M. Leonard, H. Talbot. “Automated image analysis of fish embryo for toxicology and teratology assays: a state of the art”, in *International Symposiumon on Fish and Amphibian Embryos as Alternative Models in Toxicology and Teratoly*, L’Oréal, Aulnay-sous-Bois, France, 2016 (oral presentation).

E. Puybareau, D. Genest, E. Barbeau, M. Léonard, H. Talbot. “An automated assay for the assessment of cardiac arrests in fish embryo”, in *World Congress on Alternatives and Animal Use in the Life Sciences*, Seattle, United States, 2017 (poster).

(submitted) D. Genest, M. Léonard, J. Cousty, N. De Crozé, H. Talbot. “Atlas-based automated detection of swim bladder in Medaka embryo”, in *International Symposium on Mathematical Morphology*, Saarbrücken, Germany, 2019 (oral presentation).

Contents

List of Figures

List of Tables

Preamble

In the last decades, the demographic growth and the evolution of the way of life has led to the increase of chemical compounds production. Among numerous possible applications, compounds are often considered as solutions for sustainability purposes. They are useful for energetic storage and for reducing energetic costs by increasing the performance of thermal insulation or by lightening vehicles structures for example. Synthesis chemistry allows the preservation of natural resources, by proposing alternatives to the extraction of natural substances from plants or animals. In food industry, as purchases are more and more spaced, the use of packaging and preservatives are required for obvious logistic and sanitary reasons. In agriculture, it seems unrealistic to feed the almost nine billion of individuals present on the earth without resorting to chemical compounds for crops protection. Synthetic textiles also take a dominating place in the fields of dressing, furniture, soundproofing and thermal insulation. Finally, compounds are necessary for the development of new medicines. Nonetheless, the increase in the number of produced compounds also causes an increase of the pollution, as many compounds are rejected on the environment after use or manufacture. Among them, some are biodegradable, some are accumulated in organisms or oceans, some are recyclable, some can be eliminated in safe and controlled conditions, other are not destructible. Thus, while the chemicals production increased, new interrogations have appeared concerning the potential harmful effect of these chemicals, and their future after their rejection on the environment.

These suspicions were reinforced by several environmental and health scandals that occur during the twentieth century. An example of the most impacting environmental contaminations is the mercury poisoning at Minamata, Japan. After the opening of the Chisso Corporation chemicals factory in 1908, waste products resulting from the manufacture of chemicals were released into Minamata Bay, through the factory wastewater. During the following decades, thousands of cases were identified in the local population as presenting serious alterations of the central nervous system. Searching for the cause of these new disease, wastewater was revealed as containing several heavy metals in concentrations sufficiently high to bring about serious environmental degradation. The poisoning of the population was caused by the consumption of large quantities of fish and shellfish living in Minamata Bay and its surroundings, the major causative agent being organic mercury compound. The neurological syndrome caused by mercury poisoning was named Minamata disease. A congenital form of the Minamata disease also affects fetuses in the womb, and leads to the birth of malformed babies. This disease was a striking example of an environmental contamination that also had consequences on human health. It revealed the importance of assessing the impact of chemicals on environment. *Ecotoxicology* is a relatively recent science that appeared to face this new problematic. This term was firstly introduced by Pr. Truhaut in 1969 and is derived from the words "ecology" (the study of environmental spheres and of their relations with living organisms) and "toxicology" (the study of harmful effects of chemicals on living organisms).

Another mediatized health scandal is the thalidomide case which occurred during the fifties. At this period, the assessment of developmental toxicity, meaning the impact of chemicals on the embryonic development, was only performed on mouse that did not allow to reveal a toxic effect of thalidomide. It was commercialized in many countries as an over-the-counter sedative and marketed to pregnant women who suffered from nausea. In the following years, over 10,000 babies were born that present serious morphological abnormalities, essentially in the limbs. Further tests performed on rabbit and monkey allowed to reveal that thalidomide causes alterations on the embryonic development, which is responsible for the appearance of malformations on children of the exposed mothers.

Such societal scandals led to urgent research in toxicological field and initiated an evolution in the regulation of environmental and safety assessment of chemicals. Nowadays, regulations require from chemicals industries to assess the impact of their chemicals on both the environment (*ecotoxicological assessment*) and human health (*safety assessment*). In particular, for the safety assessment, which includes assessment of developmental toxicity, regulations now require to perform tests on two different animal species, usually mouse and rabbit, for increasing the sensitivity of the tests.

However, in the meanwhile, the consideration of animal welfare increased. It was formalized in 1959 by W.M.S. Russell et R.L. Burch through the 3Rs rule, which establishes guidelines for animal testing. The aim of the 3Rs rule is to Replace, to Reduce and to Refine the tests currently performed on animals. This concept still constitutes the basis of the ethical approach on animal testing applied in Europe. It is increasingly adopted by scientific institutions and industries that try to develop alternative methods for the assessment of their chemicals. One of these alternative methods is the subject of this dissertation.

However, at digital age, one is no longer satisfied with only manual assessment methods which are often considered as time-consuming, less accurate, operator-dependent (subjective) and consequently less robust. In particular, for screening methods that involve the analysis of a large number of data, automation appears necessary. Among scientific fields that help biologists for the automation or the semi-automation of their screening and assessment methods, image analysis is widely used. *Image analysis* is the process that consists of extracting a characteristic information related to objects of interest that are present in the image. For example, image analysis allows to count or to study the shape of cells in a histological image (*i.e.* a microscopy image representing cells and structures of living tissues). This research field followed the development of image acquisition devices since their appearance in the sixties. At this period, the rise of computer science led to the increasing production of *digital images,* also called *discrete images*, referring to the set of points with different brightness (pixels) aligned on a grid. With digital image proliferation came the modern problematics of image analysis, such as image segmentation and classification. *Image segmentation* refers to the gathering of pixels which share a same predefined property. Pixels of a same group form a partition of the image. For example, pixels of a same object can be separated from the background. Image segmentation is used for extracting information (features) from images and describe them. According to the obtained image description, objects or images can be classified into several classes. For example, a segmented shape corresponding to a cell in a histological image can be described in terms of size (number of pixel in the segmented shape), of color, etc. Cells which are identified as dark will be classified as dead cells (necrosis), while the light ones will be classified as living cells. Conclusion can thus be made on the health status of the considered tissue depending on the number of dead and living cells in the image. Nonetheless, in some cases, object or image description is complicated, meaning the number of features used to describe images is too important to make possible a direct interpretation. In such case, human often needs to refer to more data that correspond to one class or to the other, in order to compare them, and to deduce rules from them to predict the class of a future data. In other words, human needs to train on a set of images previously and reliability labeled by an expert. This is the principle also used by the so-called supervised machine learning methods to perform complex classification problems.

Machine learning also raised in the second half of the twentieth century and is part of artificial intelligence. Nowadays, this scientific field is one of the most flourishing. It aims to elaborate algorithms able to learn from experience, without being explicitly programmed for that. In other world, a machine learning algorithm will not be programmed to analyze a specific feature, such as cell color, at a specific moment of the image analysis. Instead, supervised machine learning will scan many already classified data and learn from them the best way of considering each feature to reach the most accurate classification in the training sample. The resulting learned classification process could then be applied to classify future data.

In this dissertation, automated methods based on image analysis and machine learning are used to automate a test currently developed at L'Oréal to detect the teratogenic effect of chemicals, *i.e.*, their ability to provoke developmental anomalies. The test is performed on an alternative model which is the medaka fish embryo (*Oryzias latipes*), and relies on the classification of medaka images according to the presence or the absence of an abnormality. This project is a collaboration between the cosmetics company L'Oréal and the Laboratoire d'Informatique Gaspard Monge (LIGM) based in ESIEE Paris, in the context of a CIFRE Ph. D contract.

The topic of this dissertation is introduced in Section 1, which begins by introducing the regulatory context of the chemicals assessment for industry, in particular for the cosmetics industry. The alternative fish embryo model is then presented, before describing the use of fish embryo model for assessing toxicity. In particular, the teratogenicity test is described. As this test is based on the classification between healthy and abnormal embryos, the objective is to develop robust and efficient computerized procedures for embryos images classification. Section 1.2 introduces the computerized aspects of automated classification. The main principles of data classification are exposed, before focusing on machine learning classification. As all classifications rely on data characterization, some image processing tools used for features extraction from images are also described. The problematics of this work are presented in Section 1.3.

We then focus on three main points which constitute the three following sections. Section 2 describes the method we developed to automatically classify embryos into two classes: alive and dead. In this section, the method developed to detect dead embryos based on the detection of cardiac arrests is exposed. In Section 3 and 4, we take an interest to the automated detection of some types of malformations in hatched eggs (also called alevins), using machine learning. More precisely, Section 3 describes the method developed to automatically classify alevins according to the presence or the absence of an axial malformation. Section 4 describes the methodology that allows to classify alevins according to the presence or the absence of a swim bladder. Results are then discussed in Section 5.

1. Introduction

This Section introduces the topic of this dissertation, beginning with the presentation of general aspects related to the toxicological assessment of chemical compounds in Section 1.1. The international regulations relative to the safety assessment of compounds for human and for environment are described in 1.1.1. This section explains the need of the scientific community to develop alternative methods to animal testing for the toxicological assessment of chemicals. The fish embryo model, that promises to be a relevant alternative to animal models, is then presented in 1.1.2. In 1.1.3, we focus on the existing tests based on fish embryos to assess toxicity. Regulatory tests are presented. Then, the test currently developed to assess the teratogenicity of chemicals is exposed. Section 1.2 introduces computerized classification procedures that can be used for automating the test. It begins with the definition of general principles of data classification in 1.2.1, before focusing on some machine learning-based approaches in 1.2.2. Image processing tools used for image characterization are then described in 1.2.3. Section 1.3 formalizes the problematics. Existing automated methods for toxicological assessment performed from images are presented in 1.3.1. Finally, the objectives and challenges of the project are exposed in 1.3.2.

1.1. Toxicological assessment of chemicals

1.1.1. Regulatory context

1.1.1.1. Toxicity assessment of chemicals for industry

Chemical compounds are widely used in the world for many applications (agriculture, pharmacology, textile, etc.), making people and environment continuously exposed to chemicals. This societal problem makes the toxicological assessment of chemicals necessary, both for human (*safety assessment*) and for environment (*ecotoxicological assessment*). Toxicological assessment is regulated. The REACH (Registration Evaluation and Assessment of CHemicals) European regulation (n° 1907/2006/CE) was designed in 2007 in order to protect human and aquatic organisms against the hazards of chemicals commercialized in the European Union. This regulation requires from industries to perform safety and ecotoxicological assessments of the chemicals they use and produce. If the ecotoxicological assessment concerns every chemical agent or compound produced at the rate of 100 metric tons per year or more, safety assessment is required for every chemical that is produced at more than 1 metric ton per year. To perform toxicological assessment, industries need to gather all available data for example from epidemiological, *in silico* studies (mathematical modelling), *in vitro* and *in vivo* tests, in a so-called *read-across* process [1]. If gathered information is not sufficient, the REACH regulation requires complementary tests to be conducted in order to assess the toxicity of chemicals on both human and aquatic organisms. The type and number of tests are determined depending on the chemical tonnage. They are conducted according to experimental protocols that respect guidelines made by the Organization of Economic Cooperation and Development (OECD). Most of these tests require animal testing (see Section 1.1.3.1).

Toxicological assessment must be distinguished from *toxicological screening* which is not formally regulated. The term *screening* describes the practice of subjecting a high number of new chemicals to one or several tests designed to detect particular properties of these chemicals. Screening allows to sort these chemicals and to directly eliminate some of them according to the results of the tests. In the development chain of a new product, *toxicological screening* is thus a first-line process performed on a high number of compounds in order to alert and to eliminate those which have a detected toxicological effect. Chemicals which are not eliminated during toxicological screening will then be subject to a regulatory toxicological assessment, including the read-across process and potential complementary regulated tests.

Several toxicity effects are analyzed during toxicological screening and assessment of chemicals, including acute, chronic and reproductive toxicities. *Acute toxicity* refers to the adverse effects that occur following administration of a single dose of a chemical, or multiple doses given within 24 hours. Acute toxicity consists of measuring the short-term toxicity of the tested chemical (in a few minutes or hours). In contrast, *chronic toxicity* describes the adverse health effects that occur following repeated exposures to a lower dose of the tested chemical, over a longer time period (months or years). Finally, *reproductive toxicity* refers to the potential of some chemicals to interfere with normal reproduction. We define a *reprotoxic chemical* as a substance or a preparation which, by inhalation, ingestion or skin penetration, can adversely affect sexual function and fertility of adult males and females, or cause non-hereditary developmental abnormalities in the offspring. The potential of a chemical to cause such congenital malformations is called *teratogenicity*. This term has its origin in the Greek *τερ́ας* teras, meaning “monster”.

In international regulations, a particular attention is paid to the assessment of the teratogenic potential of chemicals after several scandals as the one of Thalidomide [2]. Thalidomide was commercialized in the fifties and prescribed to pregnant women who suffered from nausea. It caused the birth of thousands of malformed babies. This scandal revealed the need of reinforcing the evaluation of the developmental toxicity of chemicals as previous teratogenicity assessment tests performed only on mouse did not allow to reveal a teratogenic effect of Thalidomide. Further studies have revealed that mouse was less sensitive to Thalidomide than non-human primates or rabbits for example. For this reason, the teratogenicity assessment required by regulations now involves tests made on two models species: a rodent and a non-rodent (usually mouse and rabbit). The final analysis of all the available reprotoxicity data must allow to calculate the dose which has no effect on fertility and the dose which has no effect on development. Teratogenicity is also assessed for aquatic organisms and aim to propose a safe estimate of the highest concentration which has no effect on survival, on fertility and on development of the tested population.

The REACH regulation also asks to industries to reduce animal testing as much as possible. Alternative models and methods should be preferably used, in compliance with the Russel and Burch’s 3Rs rule, that recommends to:

- Replace *in vivo* methods performed on laboratory animals by *in vitro* (studies on human cells or animal cells) and *in silico* methods;
- Reduce the number of laboratory animal used in experimentation;
- Refine protocols in order to limit the pain and the suffering of laboratory animals, while preserving the quality of produced information [3, 4].

The ambition of the REACH regulation is to progressively exclude animal testing from the safety assessment of chemicals, by encouraging industries to develop and validate new alternative methods. For now, a few alternative methods are developed for the toxicological screening of chemicals, and none have been validated for the developmental toxicity assessment. As regulatory tests related to reprotoxicity and teratogenicity assessment are the costliest and the most demanding of animal testing, the development of alternative methods for the prediction of teratogenic effects of chemicals would be a benefit for scientific and industrial society.

1.1.1.2. Specific case of the cosmetics industry

The cosmetics industry is part of the chemical industry. Consequently, it falls not only under the jurisdiction of the REACH regulation, but also to some specific regulations, such as the regulation n° 1223/2009/CE relative to cosmetics [5]. A cosmetic is defined by this regulation as a substance or a mixture, intended to be in contact with the superficial layers of human body (epidermis, hair system, nails…), or with the teeth and the buccal mucosa, with the exclusive or principal aim to clean them, to perfume them, to change their aspect, to preserve or protect them or to correct body odors. This regulation forbids the use of carcinogenic, mutagen or reprotoxic chemicals and requires the safety assessment of chemicals used in cosmetics (through the Cosmetic Products Safety Report CPSR). Moreover, since 2013, this regulation totally forbids tests on animals that are protected by the directive 2010/63/UE of the European parliament and of the council relative to the protection of animals used for scientific purposes [6]. This directive defines animals used for scientific purposes as:

1. live non-human vertebrate animals, including:

   (i) independently feeding larval forms; and

   (ii) foetal forms of mammals as from the last third of their normal development;

2. live cephalopods

These tests are forbidden for final products, for ingredients or ingredients combinations. In contrast, some regulations related to other industrial fields, such as pharmaceutical, require animal testing for their safety assessment. As many cosmetics ingredients are also included in the composition of pharmaceutical products, data obtained from animal experimentations that were performed with respect to regulations not related to the cosmetics industry may be used for the safety assessment of cosmetics if they are relevant and the data quality is proved. Nevertheless, in such context, the cosmetics industry remains limited to the use of chemicals that are already used for other applications, making the innovation perspectives also limited. Thus, developing new methods for the assessment of chemicals toxicity, according to the 3R rules, appears to be a competitive and innovation challenge for the cosmetics industry.

In a context where animal testing is forbidden for the required safety assessment of cosmetics, the cosmetics industry is especially interested in developing screening and assessment methods based on alternative models. Moreover, even if animal testing is not forbidden for the environmental toxicity assessment of cosmetics, the company L'Oréal decided, in compliance with the 3R principle, to not use animal testing for ecotoxicological assessment of their chemicals. In the following section, we introduce one possible alternative model which is the fish embryo model.

1.1.2. Fish embryos as an alternative model for toxicological assessment

1.1.2.1. Model description and advantages

In the definition of a laboratory animal, as presented in the European directive n°2010/63/UE, "autonomous larval stages" refers to the development stages where the larva is autonomously feeding. The larva is the development stage following the embryonic stages. Thus, this definition does not include the embryonic stages of organisms such as of fishes and amphibians.

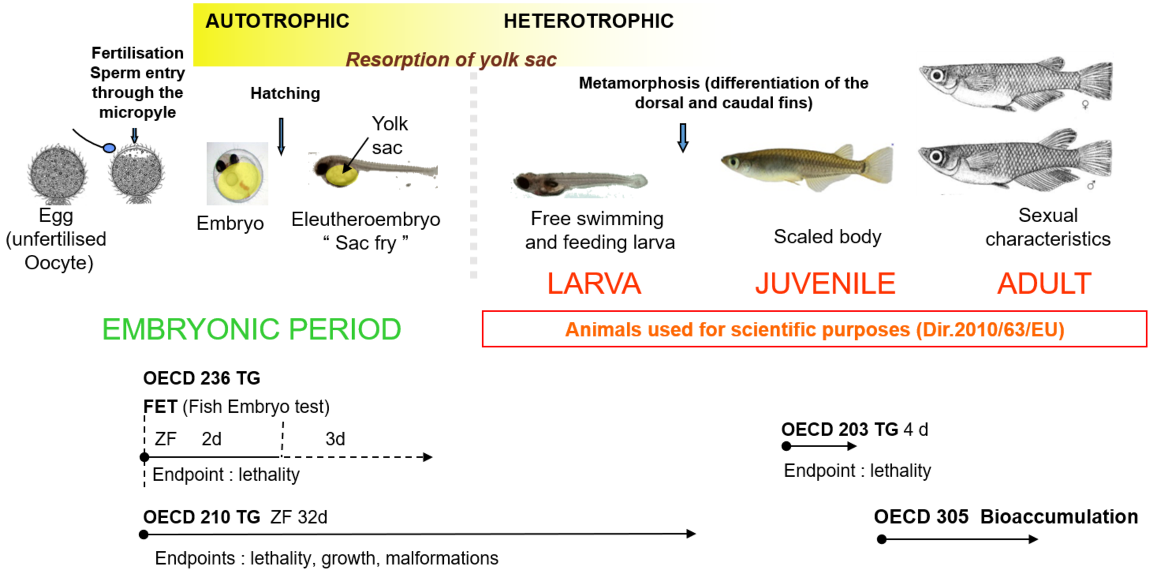


Figure 1. Main stages of the development of fish with corresponding regulatory OECD tests. The embryonic period finishes with the resorption of the yolk sac. Unless the FET (acute toxicity test), all OECD tests lasts during the larval, juvenile or adult stages which are concerned by the definition of a laboratory animal according to the Directive 2010/63/EU.

More precisely, embryonic development in fish continues after hatching through the *eleutheroembryo* stage. At this stage, the energetic supply to the developing organism is provided by the yolk. The transition to the larval stage starts with the onset of exogenous feeding (Figure 1) [7, 8, 9]. Eggs (before hatching) and eleutheroembryos do not meet the European regulatory definition of animals used for scientific purposes and are therefore considered an alternative to (adult) animal testing [10, 11, 12]. In the present manuscript, eleutheroembryos are referred to as *alevins* and we will refer to both eggs and alevins as *embryos*.

The fish embryo model presents several advantages. As an aquatic organism, this model is useful for assessing waterway pollution levels. Thus, it can provide relevant information regarding the environmental impact of chemicals [13, 14]. Moreover, fishes are vertebrates and key mechanisms of embryonic development are conserved throughout evolution from fishes to human [15, 16]. Several studies based on the classification between teratogenic and non-teratogenic substances have demonstrated a high correlation between zebrafish and mammalian developmental toxicity (overall correlation of 72-92%) [17, 18, 19, 20]. For this reason, fish embryos are considered a relevant model for studying the potential impact of chemicals on human embryonic development [21, 22]. Among fish species that are widely used in developmental biology and to study vertebrate organogenesis, zebrafish (*Danio rerio*) and medaka (*Oryzias latipes*) are two well-established models. Not only their embryonic development is well documented [23], but their early developmental stages are transparent, which simplifies the direct observation of their organogenesis without requiring invasive procedures. Compared to mammalian models, they develop quickly so tests can be performed in a shorter time. They do not require difficult husbandry techniques so they are cheap and easy to obtain from farms. For all these reasons, fish embryo models are widely used in both environmental and human toxicology as well as for assessment of chemicals efficacy [10, 13].

Most toxicological and pharmaceutical studies are conducted on zebrafish, a freshwater fish, native from the Himalayan region, and belonging to the minnow family (*Cyprinidae*) of the order of Cypriniformes. The zebrafish model is highly documented [24]. Another useful and widely used model is medaka. Medaka is a fish belonging to the *Adrianichthyidae* family, native from East Asia. It is a member of the genus *Oryzias,* the only genus of the subfamily Oryziinae. These two small fishes (up to 4-5 cm) live in rice field, marshes, ponds, slow-moving streams and tide pools. Even if comparatively fewer studies are conducted in medaka embryo, it presents some advantages compared to zebrafish. To begin with, medaka is more resistant than zebrafish to temperature changes. In particular, this point is an advantage in cases where medaka embryos are transported from the farm location to the location where experiments are conducted. Moreover, the autotrophic period, *i.e.* the duration before yolk resorption, is longer for medaka (9 days at 27°C) than for zebrafish (5 days at 27°C). Medaka organism is also more developed when yolk resorption occurs. The longer embryonic development of medaka enables the exposure of more stages of development and during a longer time period compared to zebrafish, that may improve the likelihood of detecting an adverse effect.

To conduct the study presented in this manuscript, the medaka fish embryo model was chosen (Figure 2). The following section will thus describe in more detailed the embryonic development of medaka.

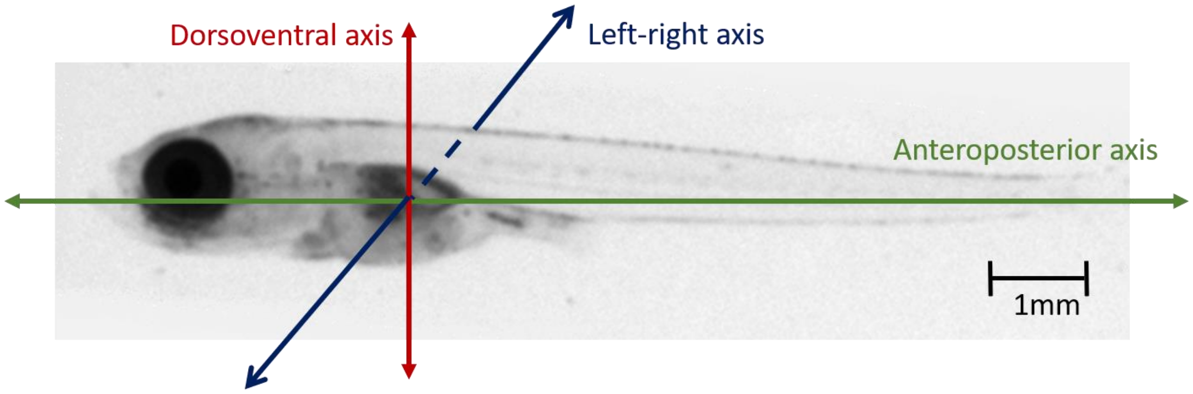
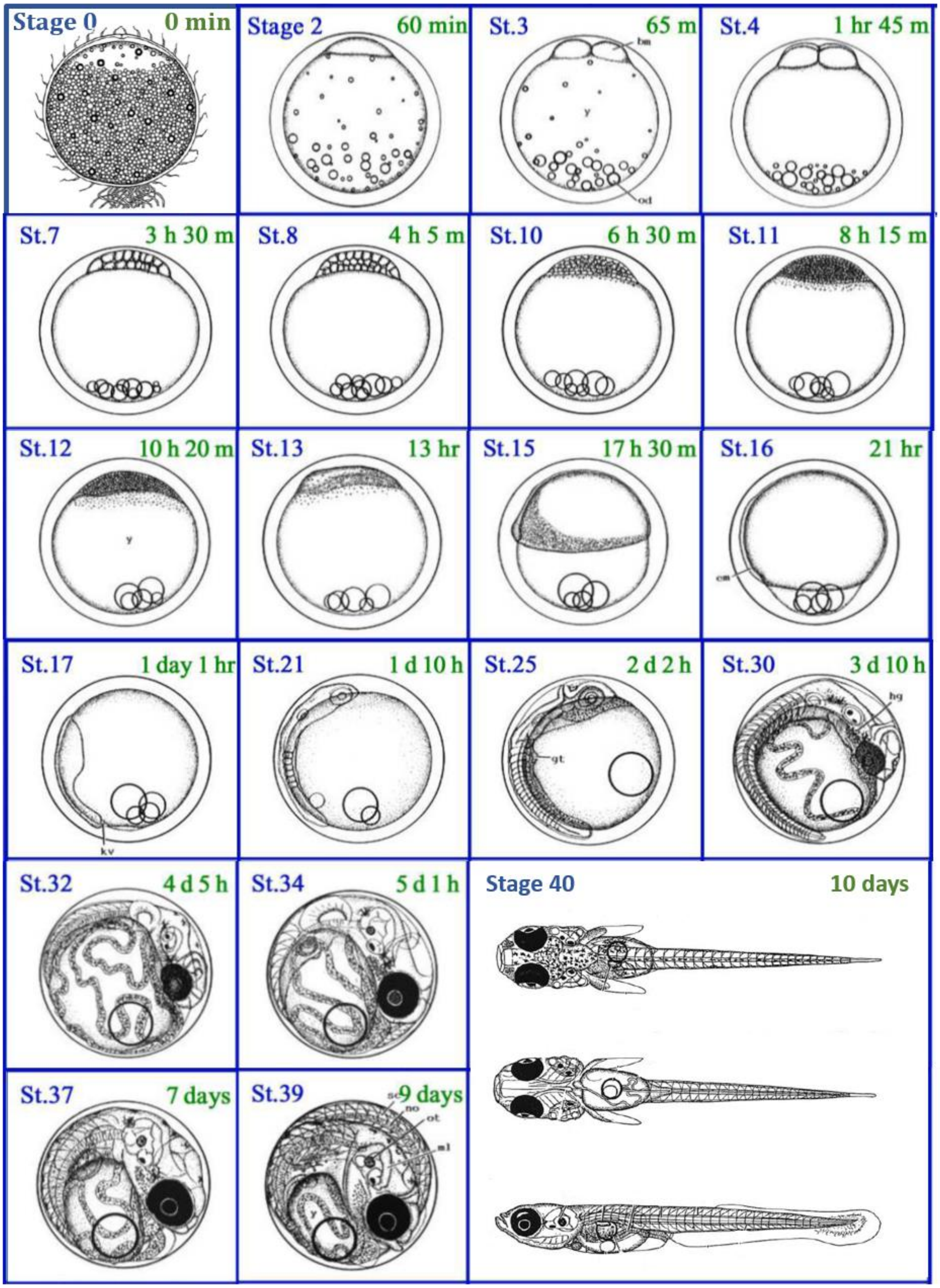


Figure 2. Medaka alevin at 9 days post-fertilization (dpf).



Adapted from Takashi Iwamatsu, Zoological Science 11, 825:839 (1994)
and from Takashi Iwamatsu, Mechanisms of Development 121, 605:618 (2004)

Figure 3. Stages of the medaka embryonic development. Organs development last from the beginning of neurulation at stage 17 until hatching at stage 40.

1.1.2.2. The medaka embryonic development

The duration of the medaka embryonic development depends on the incubation temperature. At 27°C, medaka embryos hatch between 7 and 8 days. This embryonic period is divided into several main steps that are illustrated in Figure 3 [23].

- **Fertilization (stage 1):** this stage is characterized by the ovum activation and the fusion of both nuclei from male and female gametes (ovum and spermatozoon).
- **Segmentation (stages 2 to 11):** the fertilized egg is divided into two undifferentiated cells (blastomeres), that carry on successively dividing until there are a few thousands of cells (stages 10 and 11).
- **Gastrulation (stages 12 to 16):** this stage is characterized by the invagination of the blastomeres and their repartition into three different layers: ectoderm, mesoderm and endoderm.
- **Neurulation (stages 17 to 20):** the set-up of the central nervous system begins.
- **Somitogenesis (stages 21 to 32):** somites is a transitional structure that forms along the dorsoventral axis and that latter will give rise to the three essential structures which are the vertebra, striated skeletal muscles and skin connective tissues.
- **Heart development (stage 36).**
- **Formation of the pericardial cavity (stage 37).**
- **Spleen formation (stage 38).**
- **Hatching (stage 40).**

When studying developmental toxicity, we are interested in analyzing the development of organs. For this reason, we especially take an interest in stages 17 to 38.

1.1.3. The fish embryo model to assess toxicity

As explained in Section 1.1.1, the safety and ecotoxicological assessments of chemicals are required for the cosmetics industry. In particular, as most of cosmetics are used by potentially pregnant women, it is essential for the cosmetic industry to eliminate any chemical that could have an effect on the embryonic development, by performing teratogenicity assessment. In a context where animal testing is forbidden for the safety assessment of chemicals, the fish embryo model appears promising for the development of alternative methods. For now, a few regulatory tests allow to assess the environmental toxicity on alternative models, and none exist to assess the teratogenicity.

1.1.3.1. Regulatory assays for ecotoxicological assessment

As previously said, the Organisation for Economic Co-operation and Development (OECD) provides standardized international directives for the regulatory toxicological assessment of chemicals. Among the directives that are required by the REACH regulation for the toxicity assessment of aquatic organisms, several tests are performed on fish. Depending on the tests, the specie can be specified or it can be chosen by the industry which performs the test, among four recommended species: zebrafish *(Danio rerio)*, medaka *(Oryzias latipes)*, rainbow trout *(Oncorhynchus mykiss)* and fathead minnows *(Pimephales promelas).*

The directive n°305 on *bioaccumulation* allows to measure the bioconcentration potential after an exposure phase (uptake) and a post-exposure phase (depuration) in adult fish, using an aqueous or a dietary exposure. The directive n°203 is an acute toxicity test, that relies on the lethality assessment on adult stages of fishes. Another test is performed to assess the acute toxicity in fish embryos: the Fish Embryo Test (FET) is described by the directive n°236. Finally, the directive n°210 consists of assessing malformations in fish. It is performed on the first life stages of fish and lasts from the day of egg fertilization until all controls could autonomously feed.

Except the FET, all these tests require the use of laboratory animals according to the definition of the European Directive on animals used for scientific purposes (Figure 1). As L’Oréal chose to not use laboratory animals, even for the ecotoxicological assessment of their chemicals, these tests cannot be used. Thus, to predict the toxicity on adult fish, indicators must be found during the embryonic development, which is equivalent to study the developmental toxicity on fish embryo. Moreover, for now, the only regulatory tests used for assessing teratogenicity on human are tests performed on laboratory animals. Some alternative methods are currently developed and concern *in vitro* tests. Such tests are limited by their inability to model effects that occur in a complex living organism. Studying the developmental toxicity on fish embryos could have a dual advantage, allowing the prediction of the ecotoxicity without the use of adult stages of fishes, and raising teratogenic alerts that could be indicators of a potential teratogenic effect on human.

1.1.3.2. Development of a teratogenicity assessment test

In this dissertation, we describe a test developed to study the developmental toxicity on medaka fish embryos. Several reasons explain the choice of medaka instead of zebrafish for this study. First, as l'Oréal does not have any farming in their premises, fish embryos must be delivered from an external society. As explained in Section 1.1.2.1, medaka are more resistant than zebrafish to the transportation. Second, since medaka development is slower, it can be exposed for a longer time period to the tested chemical. This may improve the likelihood of detecting an adverse effect, and makes medaka more suited for the analysis of the chemical impact on its embryonic development.

This test relies on the calculation of a Teratogenicity Index TI, defined as follow:

$$\mathrm{TI} = \frac{\mathrm{LC50}}{\mathrm{EC50}}. \tag{1.1}$$

where LC50 (*Lethal Concentration 50*) corresponds to the concentration that causes the death of 50% of the exposed embryos, and EC50 (*Effective Concentration 50*) refers to the concentration that induces a response, including malformations appearance and death, in 50% of the exposed embryos. According to these definitions, LC50 and EC50 calculations are based on a series of binary classifications of medaka embryos according to a lethality or a malformation criterion, for different tested concentrations. More precisely, the classification between alive and dead embryos at each tested concentration allows to draw a dose-response curve that represents the percentage of mortality depending on the concentration in the tested sample of embryos. The LC50 index is measured from this dose-response curve. Alive embryos are also classified into two classes: healthy and malformed. At each concentration, the classification between (i) alive healthy embryos and (ii) dead embryos and alive malformed embryos allows to draw the dose-response curve representing of the percentage of effect, and thus to calculate EC50 (Figure 4).

The value of LC50 is always higher or equal to EC50, and so the teratogenicity index TI is greater or equal to 1. TI value expresses the degree of the chemical teratogenicity. A chemical is considered highly teratogenic if it causes malformations appearance for a wide range of concentrations without provoking death. Thus, the higher the teratogenic effect is, the wider the gap between LC50 and EC50 is, and so the greater TI we have. In contrast, a TI value close to 1 means that LC50 is close to EC50, and that there is no concentration that causes malfor-

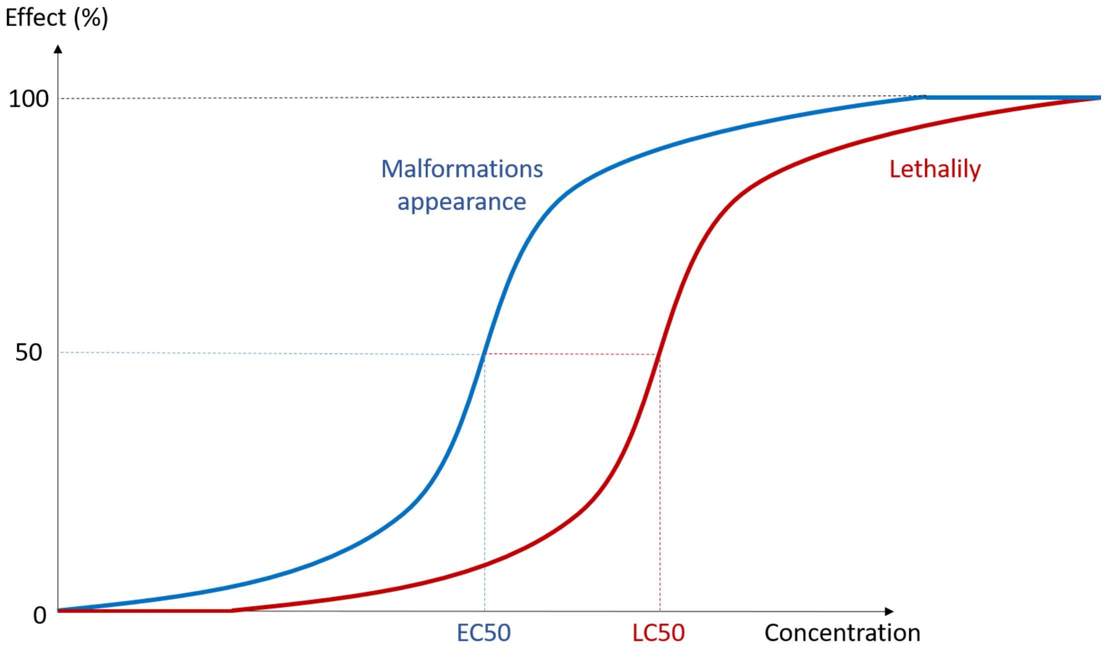


Figure 4. EC50 and LC50 calculation from dose response curves of the lethality and of the malformations appearance. LC50 is the concentration that provoks death in 50% of the tested population. EC50 is the concentration that provoks an effect, including malformations and death, in 50% of the tested population.

mations appearance without also provoking the death of exposed embryos, so the chemical is judged as non-teratogenic. Thus, to conclude about the teratogenic effect of the tested chemical, a threshold must be set on TI.

As this test is expected to be used for alerting one about the potential teratogenic effect of a chemical on human, a TI threshold was determined according to the results obtained on a list of reference chemicals, *i.e.*, a list of chemicals whose teratogenic effects are known on human. In a context of toxicological assessment or screening, this test is also expected to include a series of tests in an Integrated Testing Strategy. An *Integrated Testing Strategy* (ITS) is defined as the process of combining results from different tests to take a decision about the tested chemical. For example, the combination may not be formally structured and apply different weights depending on the tests (*weight of evidence*), a conclusion can be raised if any of the test is positive (*battery of tests*), or tests may be applied or not, depending on the results of the previous tests (*tiered strategy)*. According to the strategy of L'Oréal, the objective of the teratogenicity test is to enable the assessment of a high amount of chemicals and to eliminate chemicals with a high teratogenicity hazard. Then, further tests will be able to detect toxic chemicals that were not detected at this step. This implies to fix some constraints related to the test performance.

Assessing the efficacy of a toxicological screening assay implies to pay attention to both its sensitivity and its specificity. *Sensitivity* is the capacity of a test to indicate a correct positive result. For a test intended to detect chemicals that have a toxic effect on human, this corresponds to the proportion of toxic chemicals correctly detected. *Specificity* refers to the ability of a test to correctly indicate a negative result, *i.e.* the ratio of chemical correctly detected as having no toxic effect. The *overall accuracy* is the average of both numbers weighted by their population. In our case, the teratogenicity assay is expected to be very specific, in order to avoid wrongly eliminating non-teratogenic chemicals, and to be sensitive enough to detect strong teratogenic chemicals. For this reason, the test conducted on references chemicals was defined with a specificity of 100%, allowing to fix a TI threshold of 2.5. The assay result allows to directly take a decision: in the case of a chemical with a measured TI higher than this threshold, it is highly suspected to have a teratogenic effect and it will be eliminated. On the contrary, if the measured TI is lower than this threshold, the result is not sufficient to conclude about the safety of the chemical. Thus, the chemical development will continue through further assessment tests [1].

In practice, the teratogenicity test relies on the visual assessment of medaka embryos state, after a 9-day incubation (authorized stages of development) in increasing concentrations of the tested chemical, including none for control. Medaka embryos are individually incubated into wells of a 24-well plate (Section 6.1 in the Appendix). The day of analysis, each fish embryo is anesthetized. Then embryos are observed under a microscope to detect potential alterations of the embryonic development. Microscope observations allow to manipulate the embryo and to see it under all possible orientations (dorsoventral, lateral and all intermediary orientations). The expert annotates each embryo as presenting an abnormality or not, and which ones if any. The developmental alterations currently considered are the following:

- **functional abnormalities** (presence or absence of heartbeat) allow to calculate LC50 and EC50;
- **morphological abnormalities** (pigmentation, alevin size, presence of eye malformations, edemas or spine malformations, absence of swim bladder) allow to calculate EC50. Examples are shown in Figure 5.

Currently, this test is manually performed at l'Oréal, meaning the embryo classification is manually performed by the expert. It takes around 1 minute in order to rigorously analyze each functional and morphological endpoint and label it. Since each chemical assessment requires the analysis of 144 embryos, this means that more than 2 hours are necessary to screen only one chemical, monopolizing an operator during all this time. Moreover, this assessment is subjective, as it involves visual analysis that depends on the experience and of the accuracy of the operator. An experiment was performed to assess the subjectivity between operators on a same sample, also called *inter-operator subjectivity*. Three different operators observed the same sample of 143 medaka embryos. This revealed a difference rate on observations of 20%. This difference has a direct impact on the TI calculation and on the result of the test. Observations of a same operator can also lead to different conclusions depending on the conditions, on the operator's experience or on its fatigue for example. For example, an embryo with a small malformation may be detected if it is surrounded by many healthy embryos, while the same embryo may not be detected if it is surrounded by many strongly malformed embryos (Figure 6). This is what we call *intra-operator subjectivity*. As computer is not influenced but programmed to do a specific task always the same way, it could help to improve this process which is time-consuming and prone to error. This is what constitutes the subject of this work.

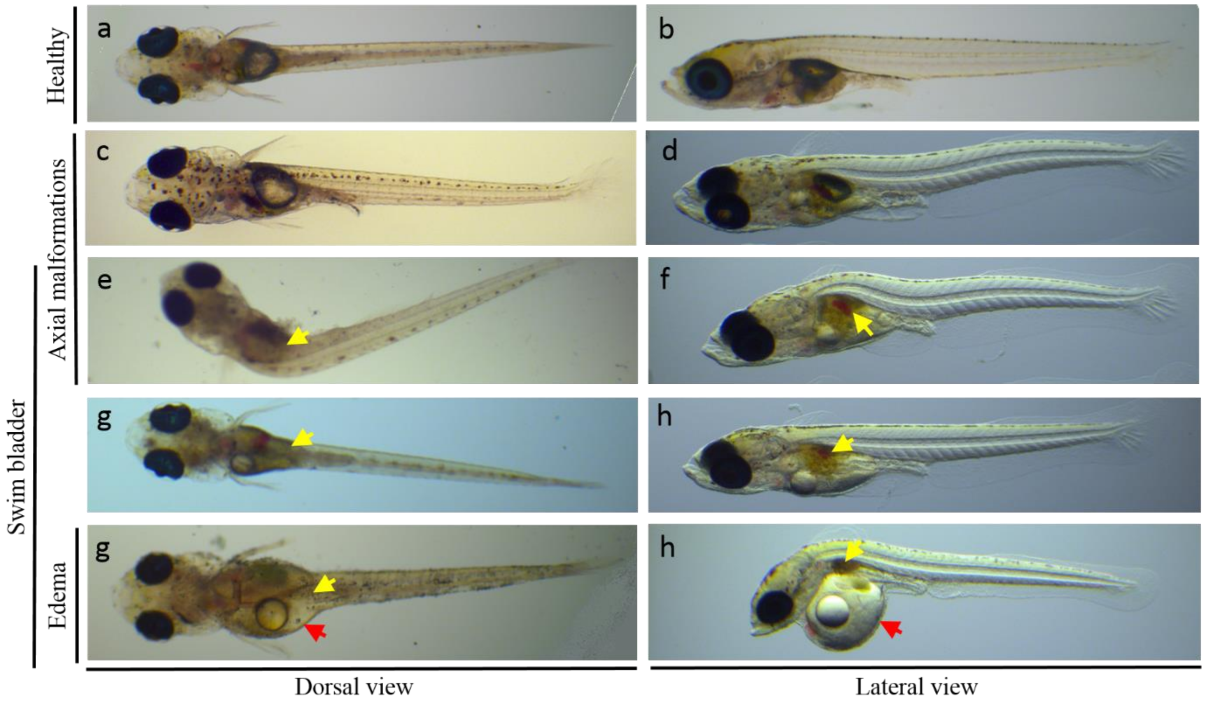


Figure 5. Examples of morphological abnormalities observed under a stereomicroscope for alevins seen in dorsal and lateral view. a and b: healthy alevins. c to f: alevins with an axial malformation. In c, the alevin shows an axial torsion visible as the head appears in dorsal view whereas the tail appears in lateral view. In d, e and f, the alevins show a curved spine. e to h: alevins without a swim bladder. The yellow arrows indicate the location where a swim bladder should be present. g and h: alevins with a large edema on the yolk, indicated by red arrows. The alevin in h also shows a curved spine.

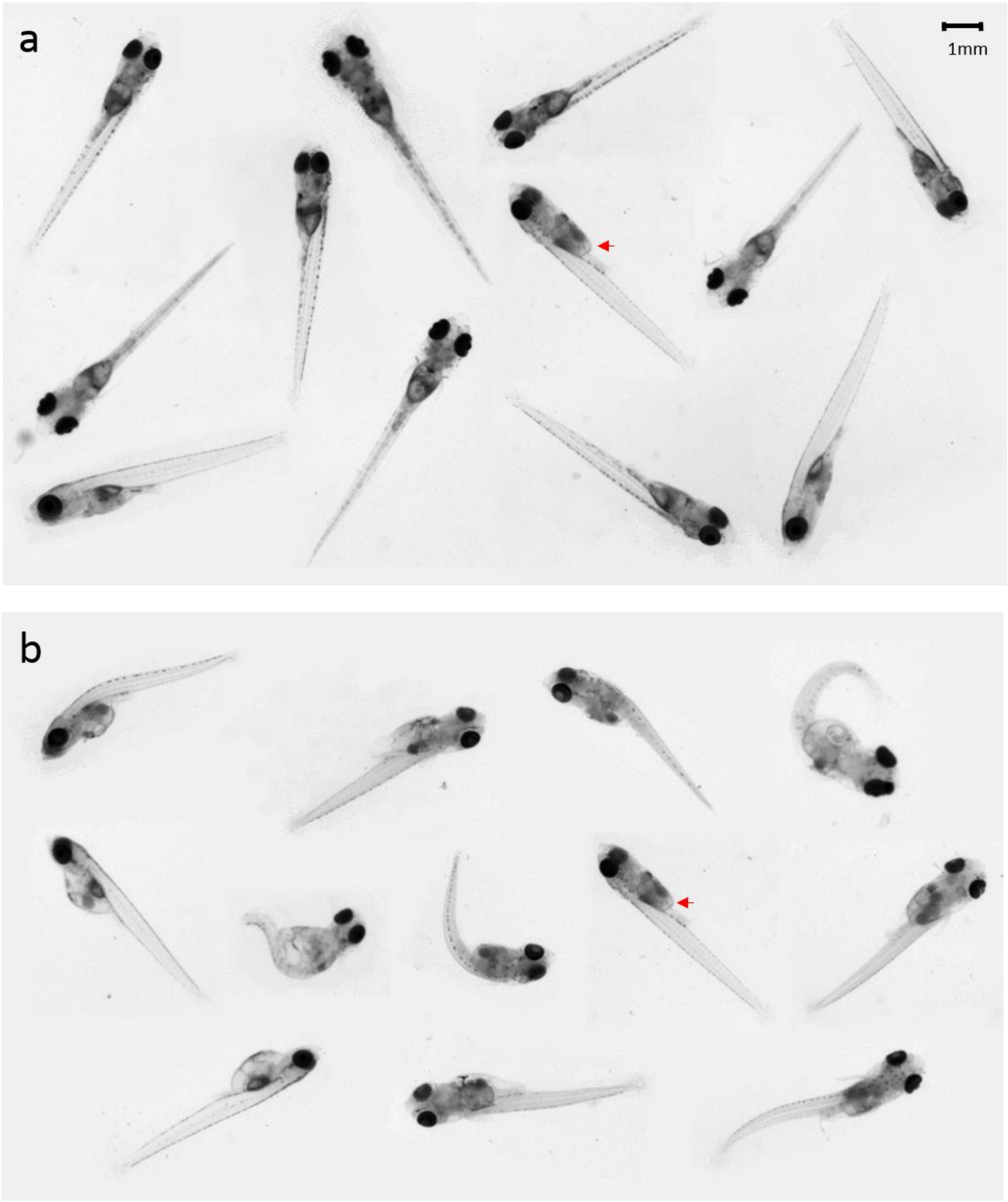


Figure 6. Demonstration of subjectivity. The same embryo with an edema marked with a red arrow is shown surrounded by healthy embryos in a, and surrounded by strongly malformed embryos in b. In b, malformed embryos especially show large edemas and important malformations of the spine. The malformation of the marked embryo does not appear as visible in b as in a.

1.2. Automated classification for toxicological assessment

Computerized procedures could help to overcome the limitations linked to the duration and the subjectivity of the teratogenicity test, by automating the visual assessment made on medaka embryos. As explained in the previous section, the visual assessment and the TI calculation rely on classification of embryos according to functional and morphological criteria. Existing classification procedures appear promising for detecting abnormalities in medaka embryos and so, for automating the teratogenicity assessment test.

1.2.1. Classification main principles

Classification is a difficult task for both human and computer, but for different reasons. Training human to perform delicate classifications is difficult and expensive, even experts may not always agree, and human get tired and tend to make mistakes. In contrast, computers typically never tire and they give results that are more robust and sometimes even better than human. Nevertheless, as computer classification is often based on human expertise, it may inherit biases and limited expertise. Typically, computers are less able to generalize from few examples and require much more data and annotations to perform as well as human. For this reason, automated object classification according to some predefined criteria is an important and challenging task within the field of computer vision and artificial intelligence. It is used in various application fields including biometry, vehicle and robot navigation, remote sensing, and biomedical imaging [25, 26, 27].

Intuitively, classification is the process of grouping individuals with the same or similar characteristics (features) into a consistent set: e.g. domestic animals may be classified into dogs, cats, guinea pigs, goldfish, etc. A class is a set defined by a certain property stemming from the data features, and consequently all data having features that respect this property are deemed elements of this class. In order to build a classification, one must find the property of each class (e.g. cats have retractable claws). Classification algorithms can model a problem differently depending on the input data and the desired outcome. We can categorize classification methodologies in three main classes:

- **Unsupervised**: This category concerns all approaches that operate on unlabeled training data (*i.e.* a dataset without any information on classes of the data). The aim of these approaches is to cluster them in different classes or categories by extracting patterns from the inherent structure of the input data without explicitly-provided labels (e.g. an untrained operator who tries to distinguish healthy and malformed embryos in a sample). Unsupervised learning is useful for exploratory analysis as it may highlight the inherent structures in the data. Hierarchical cluster analysis (HCA) and K-means are common examples of unsupervised classification algorithms [28, 29, 30].

- **Supervised**: The main difference with unsupervised classification is that supervised algorithms rely on ground truth, *i.e.* on data that are labeled according to the desired result, and that constitute the training database (e.g. a human can be trained to assess embryos malformations by observing 100 training embryos under a microscope that were previously labeled as “malformed” or “healthy” by an expert). In other words, the training database is composed of a set of pairs (x_i, y_i) where x_i is the input vector that contains features of the data i and $y_i = f(x_i)$ is the associated desired result (label), linked to the input x_i by the mapping f. In supervised classification, finding each class property is equivalent to finding the relationship between all inputs and associated labels of the training set, by searching for the function that best approximates the target mapping f. It is used for prediction purposes. Once the mapping is found based on a set of training data, this mapping is then applied to new data of a testing set in order to predict to which class these new data belong (if the operator notices on training embryos that malformed embryos have smaller eyes than healthy ones, the operator will classify new unlabeled embryos of the testing set according to the size of their eyes). Thus, supervised classification is one of the two key principles used for prediction purposes, the regression being the second. The distinction between both is made on the type of the predicted result which is categorical for classification (e.g.: presence or absence of a malformation), and real-valued for regression (orientation degree of the alevin).

- **Semi-supervised**: semi-supervised learning is a newer form of machine learning where only some of the data are labeled, possibly with noisy ground truth. Semi-supervised uses a combination of supervised and unsupervised learning techniques as well as exploration techniques for first finding the good labels in the ground truth, and then reinforces these by using the clustered unlabeled data near the labeled ones as new training data [31].

When classification is performed in a supervised way, two methods can permit to determine the mapping function f that links the input vector of data features and the desired outcome: the empirical and the machine learning-based method. The first one consists of building its own rules in an empirical way on training data. Most of the time, this method is equivalent to find accurate thresholds to apply on features in order to correctly classify the data in the expected classes (which size must the eyes have to be considered too small, which size must the swim bladder have to be considered inflated?). This method is suited for classification tasks that involve a low number of features, or features that are easy to discriminate with data visualization. However, for most classification tasks, we often note an increase in the data size, in the number of classes, or in the features dimension, that make the classification rules difficult to build in an empirical way, and thus affect the classification accuracy. Such classification process built with explicit rules tend to handle problems with a high complexity relatively poorly. For this reason, we extensively rely on the second method which involves machine learning-based approaches to solve complex classification problems. The principles of machine learning-based approaches are described in the next section.

1.2.2. Learning rules from the data: machine learning approaches

1.2.2.1. Machine learning principle

Machine learning (ML) is defined as a subset of artificial intelligence (AI) that allows computers to learn rules, decisions making processes and other cognitive tasks without these being explicitly and specifically programmed in. Instead, machine-learning algorithms and methods manufacture these directly from input training data. ML is used in data analysis in order to automatically develop analytical models. This technology is adaptive, in the sense that machine learning-based programs can provide answers in the presence of new, previously unseen data. Machine learning related methods have been extensively used for classification problems in biology and biomedical purposes [32, 33, 34]. For example, machine learning can help doctors for medical diagnosis or for predicting the evolution of a disease [35, 36]. In toxicological screening, machine learning can help interpret the large amount of data that results from the combination of mainly *in vitro, in vivo* and *in silico* assays performed on a chemical. Of course, there are limitations. ML algorithms do not learn like humans: they need a lot of well-labeled training data to achieve good performance. However, in the last few decades and particularly since the early 2010s with the advent of popular, effective frameworks for what is known as “deep learning”, ML has made tremendous progress, that have sometime exceeded human-level performance [37].

With classification tasks, machine learning approaches are used in order to automatically build a decision-making model [26]. In the case of supervised classification described in the previous section, this is equivalent to building a mapping function f making a correspondence between the input features vector and a desired output. For this purpose, supervised learning relies on optimizing the function parameters with respect to a distance between the function output and the expected answer. Depending on the form of the approximated function, classification methods can also be grouped in two categories:

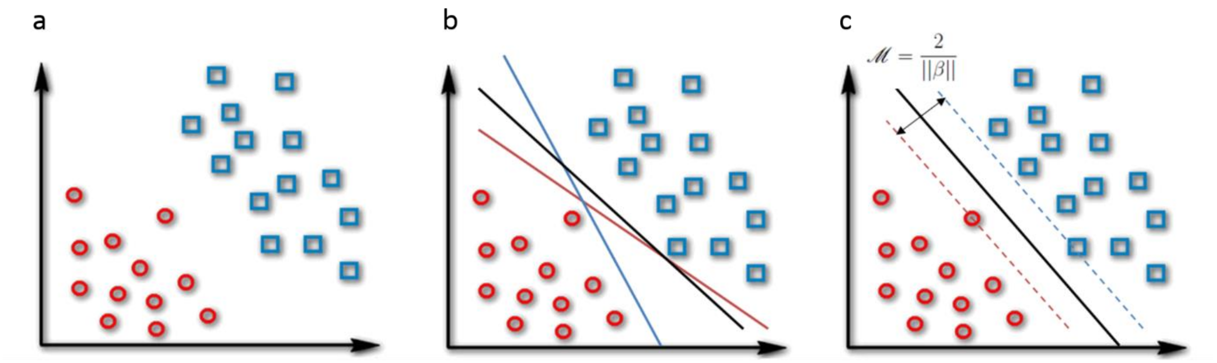
- **Parametric model.** This refers to classification methods that simplify the function into a mathematical form that depends on a known number of parameters. Resolving such problems consists of selecting a form for the mapping f, and then learning the parameters of the function from the training data. Linear Support Vector Machines are a typical example. Making an assumption on the form of the function f can greatly simplify the process of function approximation. Moreover, these parametric classification models are quick and can learn from little data. This makes them useful for classification problems where the data are easy to characterize, as less features are necessary to characterize them and less data are necessary for representing the entire population. They are often resistant to data bias, meaning that any sufficiently representative training data will give robust results. However, the choice of mapping clearly constrains the solution. In some complex classification problems, this implies limits on what can be learned and may result in comparatively poor results, as the problem may not be approximated by a well-known parametric function. It is often the case in image classification problems, because it is difficult to propose a generic image model.

- **Non-parametric model.** In contrast, a non-parametric machine learning model refers to models that do not make strong assumptions about the form of the mapping function f. The number of parameters of f can be very large, potentially infinite. These methods are flexible, as they are free to learn any functional form from the training data, and the complexity of the model can grow with the size of the training data. If more training data are available to estimate the mapping function, this generally results in higher performances. K-nearest neighbors, Decision trees, RBF Kernel Support Vector Machine and Neural Networks are examples of non-parametric learning algorithms. While it may seem that these methods are superior, they do require more data to train, and they often learn a biased model, meaning that classification results on new, unseen data may be poor, if the training data is insufficient or does not represent the entire population well.

The difference between supervised/unsupervised; parametric/non-parametric models and the various sources of errors and biases give rise to many important issues in machine learning that would take too long to report here. The interested reader will find an excellent reference in [38]. In the following, we present two of the most documented supervised learning classification methods, which are also the most frequently used in the context of toxicological assessment of chemicals: Support Vector Machines, which are typically used for the classification of basic phenotype such as dead, hatched of unhatched embryo [39, 40], and Random Forest, which is used for the recognition of more complex morphological phenotypes such as axial or yolk malformations [41].

1.2.2.2. Support vector machines

Among existing supervised learning methods, Support Vector Machines (SVM) are suitable for recognizing multivariate patterns and thus are one of the most widely used methods for classification and regression purposes [42, 43, 44, 45]. Considering training data that are characterized by a set of features and classified into two known categories, the binary SVM classifier aims at assigning one of these two categories to new unlabeled data based on the comparison between the new data features and training data features. For this purpose, the SVM algorithm relies on the projection of training data in the features representation space (see an example in a 2D features space in Figure **7**a). The SVM result model is defined by a *hyperplane* of the features space that separates the two classes of training data with the widest possible gap between the hyperplane and each of the two categories. For that, we define the *margin* as the distance between the hyperplane and the closest training examples of each class (called support vectors). Maximizing this margin allows to ensure a high generalization of the classifier. Thus,



Adapted from Imen Melki, ''Towards an Automated Framework for Coronary Lesions Detection and Quantification in Cardiac CT Angiography'', Thesis Manuscript, 2015

Figure 7. Support vector machines classifier. a: a 2D example of linearly separable data. b: the two classes could be separated using multiple lines. But being too close to some training data such lines are more sensitive to noise and will not generalize correctly. c: the better hyperplane separating the two training subsets is the one maximizing the margin distance as it will be able to better classify new samples that are close to the current decision boundary.

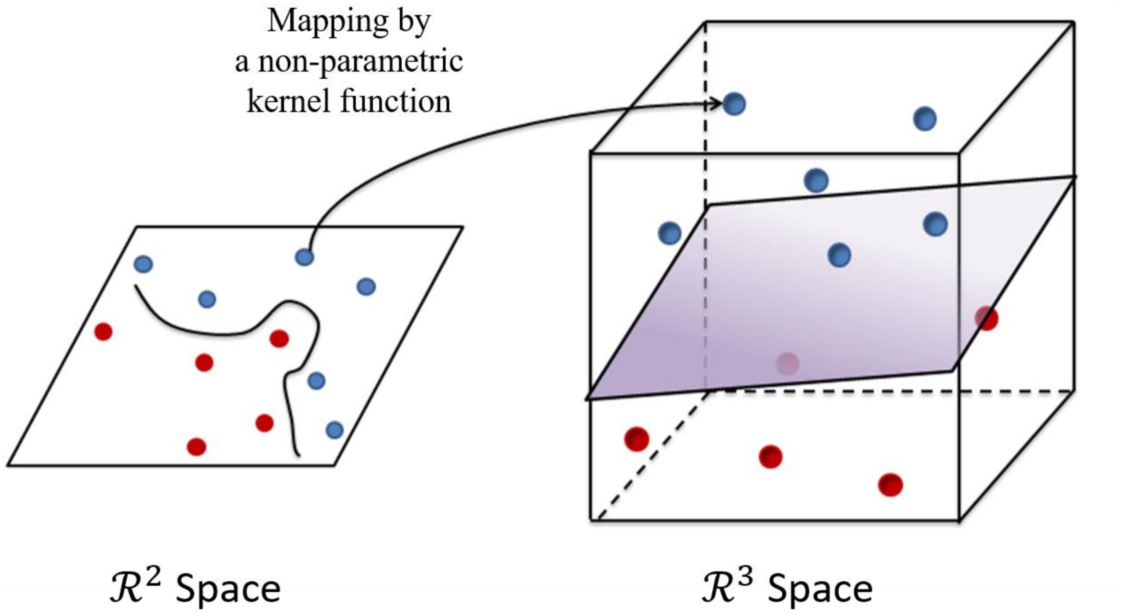


Figure 8. Non-linear classification with a support vector machines classifier. The data are not linearly separable in the 2D representation space. A non-parametric kernel function is used to map each data from the 2D representation space to a 3D space in which data are linearly separable by a hyperplan shown in grey.

when classifying new unlabeled data, these data are mapped into the same features space and then, each data is assigned a label that depends on the side of the hyperplane this data falls on. In case of linear discrimination problem, the result hyperplane h can be formally defined as the linear combination of the features vector of training data $x = (x_1, \dots, x_n)^T$ and of the weight vector of training data $w = (w_1, \dots, w_n)^T$:

$$h(x) = w_0 + w^T x, \tag{1.2}$$

where n is the number of features and thus the dimension of the representation space, and w_0 is the bias of the hyperplane h (Figure 7b). The margin is equivalent to twice the distance to the closest training data and is expressed by $\frac{2}{\|w\|}$ (Figure 7c). The SVM algorithm aims to find w_0 and w that define the optimal hyperplane, meaning such that the margin is maximized. It is equivalent to minimizing the constraint function $C(w)$:

$$\min_{w,w_0} \frac{1}{2} \|w\|^2, \text{ subject to } y_i(w_0 + w^T x_i,) \geq 1, \ \forall i, \tag{1.3}$$

where y_i is the label of the training data x_i. Lagrange multipliers can be used to solve this problem. Figure 7 shows an example of SVM application in a 2D example of linearly separable training data (so $n = 2$).

The SVM classification principle was also extended to nonlinear classification (Figure 8). This relies on the transformation, defined by a non-parametric kernel function, of the data representation space into a space of higher dimension (potentially into an infinite-dimensional space). The probability of finding a linear separation hyperplane in such a higher dimension space increases dramatically [42], but may introduce data bias into the method.

1.2.2.3. Decision trees and random forests

Decisions trees are hierarchical piecewise constant models that allows to make a final decision from input data based on multiple variable analysis. They are often used as predictive models for classification purposes in supervised learning [46].

A decision tree is a directed binary tree where *non-leaf nodes* carry decision rules and where the *leaves* carry target labels. The *decision rules* associated with each node take the form of a Boolean test function pointed toward the children of the considered node. 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 result, the label of this leaf (Figure 9).

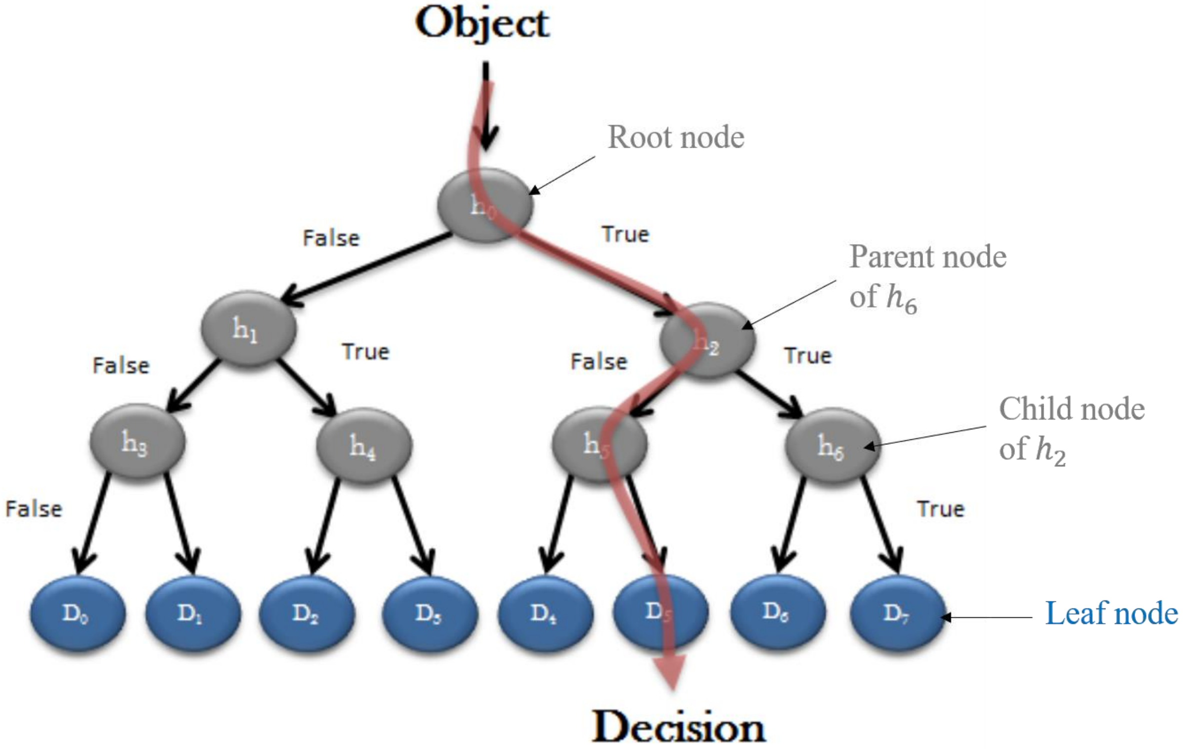
The accuracy of a decision tree-based algorithm is assessed on a data sample by comparing the predicted values on this sample with a corresponding set of correctly labeled data. On a sample of size n_{sample}, we respectively call y and $\hat{y}$ the series of labeled and predicted values. If y_i is the label of the ith data and $\widehat{y_i}$ is the corresponding predicted value, then we calculate the accuracy 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:

$$accuracy(y, \hat{y}) = \frac{1}{n_{\text{sample}}} \sum_{i=1}^{n_{\text{sample}}} 1(y_i, \widehat{y_i}), \tag{1.4}$$

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

Fitting a Boolean test function to a training set of labeled 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 met. Examples of stop criteria are reaching a desired accuracy or a 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 of Random Forests (RF).



Adapted from Imen Melki, ''Towards an Automated Framework for Coronary Lesions Detection and Quantification in Cardiac CT Angiography'' , Thesis Manuscript, 2015

Figure 9. Decision tree composed of a set of hierarchically organized nodes (in grey) with ending leaves (in blue). Each node corresponds to a test function (h_i) used to decide whether to send the input data to the left child node or to the right one. Each path on the tree leads to a leaf point corresponding to a final decision D_i (prediction) on the input object.

Random Forest 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 weaker decision trees can produce a stronger and more reliable classifier. The Random Forest algorithm was first formulated by Leo Breiman and combines two levels of randomness : each decision tree is computed (node split functions are defined) (i) on a random subset of the training dataset, according to the general technique of bootstrap aggregating, or bagging, and (ii) using a randomly selected set of features [47]. This technique allows to reduce misclassification error due to single application of the partitioning clustering procedure by ensuring a low correlation level between the trees [48, 49]. Moreover, the selection of a subset of training data for each tree building allows the algorithm to reduce the search space dimension and thus to enhance efficiency.

As any other supervised learning classification technique, the Random Forest algorithm relies on two successive steps which are the training (or learning) step and the testing step.

- **Training step.** During this step, the parameters of the Random Forest model are optimized. 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 labeled 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, the standard entropy criterion can be used. 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 labeled data is defined by:

$$H(S) = -(p_{L_-}\log_2 p_{L_-}) - (p_{L_+}\log_2 p_{L_+}), \tag{1.5}$$

where p_{L_+} and p_{L_-} are respectively the relative frequencies of the positive label L_+ and the negative label L_- in S. For example, in a binary sample containing 20% of malformed embryos (label L_-) and 80% of healthy embryos (label L_+), where each data is equally weighted, the relative frequencies p_{L_-} and p_{L_+} are 0.2 and 0.8 respectively. 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 Φ and a threshold ϑ, the splitting function s associated with Φ and ϑ is the map $s_{\Phi,\vartheta}$ from the set of data S and into the set $\{\text{True}, \text{False}\}$ such as $s_{\Phi,\vartheta}(x) = \text{ True}$ whenever the feature $\Phi(\text{x})$ is higher than the value ϑ *i.e.* $\Phi(\text{x}) > \vartheta$.

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

$$\text{Gain}\big(S, s_{\Phi,\vartheta}\big) = H(S) - [\frac{n_{\text{True}}}{n} \times H(S_{\text{True}}) + \frac{n_{\text{False}}}{n} \times H(S_{\text{False}})], \tag{1.6}$$

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

Some parameters control the size and the complexity of the trees. One can specify the 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. Optimization algorithms allow to determine the most accurate combination of parameters according to a predefined criterion (accuracy optimization for example).

A weighting system can be used in order to favor one of the two labels. Such weighting intervenes in the calculation of the label's 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:

$$p_{L_-} = \frac{w_{L_-} \times n_{L_-}}{(w_{L_-} \times n_{L_-}) + (w_{L_+} \times n_{L_+})} ; p_{L_+} = \frac{w_{L_+} \times n_{L_+}}{(w_{L_-} \times n_{L_-}) + (w_{L_+} \times n_{L_+})}. \tag{1.7}$$

For example, a sample containing 20% of malformed embryos (label L_-) and 80% of healthy embryos (label L_+) can be balanced by applying a weight of $w_{L_-} = 4$ to malformed embryos. We obtain the relative frequencies $p_{L_-} = p_{L_+} = 0.5$. The training step results in the construction of a forest where each optimized tree represents a hierarchical test to apply to new unlabeled data to classify them during the testing step.

- **Testing step.** A new sample of unlabeled data is sent to the root of all trees of the forest. For each tree, this sample is pushed through the inner nodes. At each node, the input data is sent either to the left or to the right child node depending on the result of the associated test function, until reaching a leaf node. Parallelizing this process reduces its computational cost. For each testing data, we obtain as many predictions as the number of trees in the forest. In order to choose the final prediction for the input data, a decision function is applied to all the predictions. Most commonly, the mode of all predictions is taken as the final decision.

While many computerized methods exist to automatically classify data, all these methods essentially rely on data characterization, meaning associating characteristic features to these data. These features, also called descriptors, are used as input to the classification algorithm, which will look for the relationship between the input features and the final output prediction. There are many ways to obtain features from the data. They can be manually generated by experts during a manual annotation process, or it can be automatically extracted from the data itself. When the data to classify are images, the classification process includes object detection and segmentation, features extraction, and object classification [27]. Some basic principles of image processing used for object detection, segmentation and features extraction are provided in the next section of this manuscript.

1.2.3. Automated features extraction for image classification

A *greyscale image* is defined as an application $f: \{(x, y) \in \mathbb{Z}^2, 0 \leq x < M, 0 \leq y < N\}$ in a set of values V, where M and N are two integers called *width* and *height* of the image respectively. Each element of an image is called a pixel. The corresponding *digital image* is defined as the numerical representation of the image into a two-dimensional matrix, where each element is identified by its spatial coordinates and represents a pixel of the image. With each pixel is associated a value for a scalar image, or a vector for a multivariate image (such as a color image), This value can represent anything, but most often is associated with the amount of light received on the image sensor at the location of the pixel, also called the *intensity*. The pixel intensity is equivalent to a shade of grey from darkest (black) to lightest (white). The *pixel depth* is defined as the number of possible values that a pixel intensity can have. It is equivalent to the size of the set V. The pixel depth is determined according to the number of *bits* (binary digit) on which each pixel of the image is encoded: an image encoded in x bits means each pixel can take 2^x different values. A 1-bit image (also called *binary image*) is a digital image where each pixel can take $2^1 = 2$ possible values: 0 or 1. These values are typically represented in black and white respectively. In a 8-bit image, the most common type used in this dissertation, each pixel can take $2^8 = 256$ different values.

In a similar way, a video can be interpreted as a 3D signal, with the third dimension representing a temporal variable (also called $2D + t$ image). It represents the evolution of a greyscale image through time. A video is numerically represented as a three-dimensional matrix for which each plane corresponds to the digital image at time.

In this dissertation, we mostly use 8-bit greyscale images and 8-bit $2D + t$ images. Features extraction from such images requires shape extraction (segmentation) and shape description.

1.2.3.1. Shape extraction

In image analysis, *shape extraction* refers to an ensemble of techniques, including filtering and segmentation methods, that consist of identifying and representing patterns on the analyzed image. This section introduces the main operators used in this dissertation for shapes extraction.

In computer vision, image segmentation refers to the process of partitioning a digital image into several regions, called segments, with similar characteristics or semantic content, allowing to simplify the image representation. Resulting segments are expected to be meaningful and suited for further analysis. Image segmentation is used to locate object boundaries (edge detection) and thus is used for shape extraction in image processing. Depending on the characteristics of the shape we aim to segment, specific techniques can be used.

Thresholding is the simplest method used for image segmentation. This process allows us to create binary images from greyscale images, by defining a threshold value and by comparing each pixel intensity to this threshold value. If the pixel intensity is greater than the fixed threshold, then the pixel is attributed the value 1 in the resulted thresholded image, corresponding to a white pixel. If the pixel intensity is lower than the fixed threshold, then the pixel is attributed a zero intensity in the resulted thresholded image, corresponding to a black pixel. Formally defined, let I be a $M \times N$ pixels grey level image, taking 8-bit discrete values, *i.e.* $I: \{1, \dots, N\} \times \{1, \dots, M\} \to \mathbb{Z} \cap [0, 255]$. The thresholded image above value θ is denoted $(I)_{\geq\theta}$:

$$x \in [1, N] \times [1, M], (I)_{\geq\theta} = \begin{cases} 1\ if\ I(x) \geq\ \theta \\ 0\ if\ I(x) <\ \theta \end{cases}. \tag{1.8}$$

The threshold value can be empirically or automatically determined. A thresholding algorithm is called adaptive when it modifies the fixed threshold depending on the content of the considered image. A typical an effective example of an adaptive thresholding algorithm is Otsu's method, used to automatically cluster an image into two classes of pixels corresponding to the foreground and the background. This method relies on the analysis of the image histogram, assuming that this histogram is bi-modal (the mode being the value with the highest number of occurrences) (Figure 10b). Considering a threshold θ that results in two classes of pixels in the image (0 and 1), the intra-class variance σ_w^2 is defined as the weighted sum of pixel intensity variances of the two classes:

$$\sigma_w^2(\theta) = w_0(\theta)\sigma_0^2(\theta) + w_1(\theta)\sigma_1^2(\theta) \tag{1.9}$$

where w_0 and w_1 refer to the probabilities of being in the classes 0 and 1 respectively; and σ_0^2 and σ_1^2 are greyscale variances of the pixels that belong to the classes 0 and 1 respectively. The algorithm calculates the optimal threshold such that the intra-class variance is minimal. It can be shown that this algorithm also maximizes the inter-class variance (Figure 10) [50].

Image segmentation generally implies a filtering step, *i.e.*, a process to alter image characteristics as the size, shading or morphology. A frequently used method for blurring, sharpening (edge enhancement) or for edge detection in images consists of modifying the value of each pixel according to the values of its local neighbors. A *window* is used to define the neighborhood. The *median filter* replaces each pixel with the median value in its local window. The median filter is often used to reduce the number of pixels with extreme values and thus to denoise images (Figure 11c). Many filtering methods rely on the *convolution* between a *kernel* and the image. A kernel is simply a collection of values associated with a window. Generally, these values are normalized so that they sum to 1, in order to keep the image overall contrast

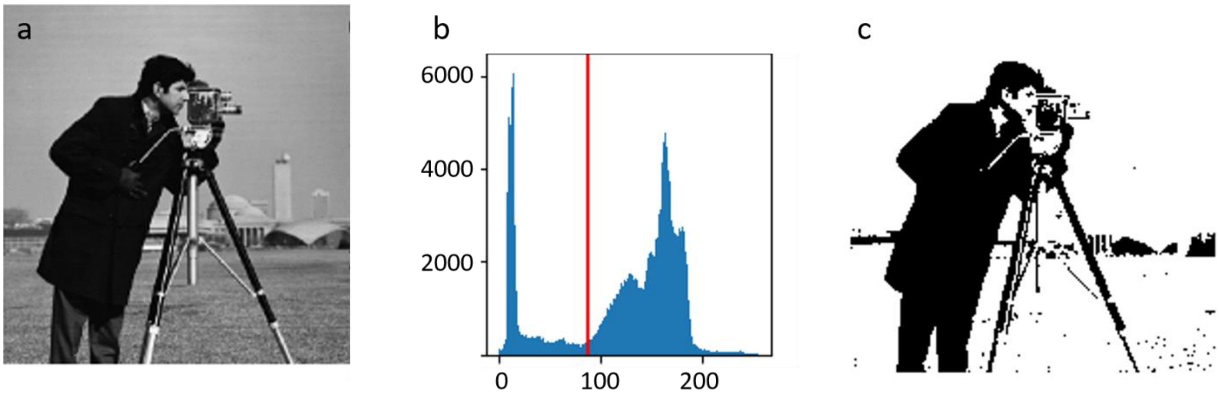


Figure 10. Otsu threshold. a: original greyscale image. b: the histogram of the image shows the distribution of the pixel intensities with two relative modes. The red line shows the threshold which best discriminate the two modes of the histogram. c: resulted binary thresholded image.

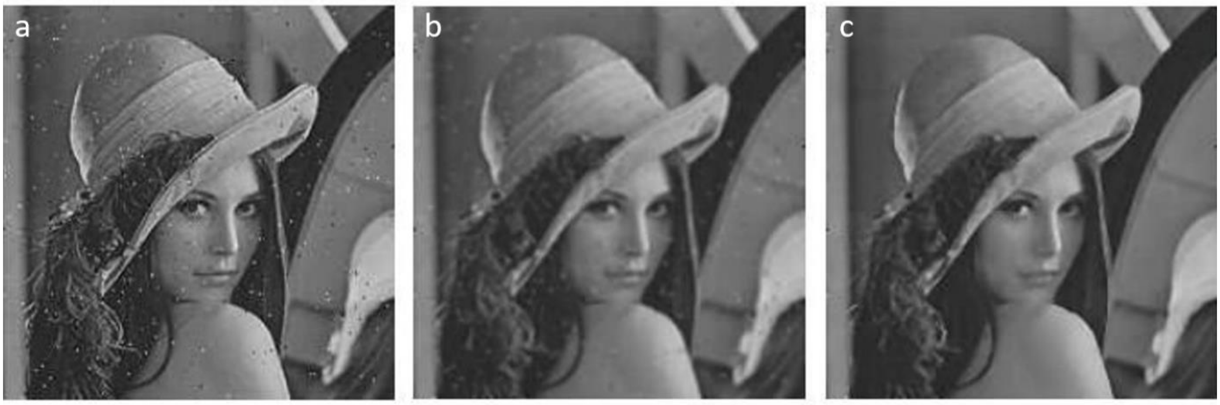


Figure 11. Examples of image filtering with Gaussian and median filter. a: original noisy image of Lena. b: image filtered with a Gaussian kernel. c: image filtered with a median filter.

unchanged. *Convolution* is the process of summing each element of the image to its window, multiplied by the values in the kernel. Here is presented the general expression of a convolution:

$$g(x,y) = (\omega * f)(x,y) = \sum_{s=-a}^{a} \sum_{t=-b}^{b} \omega(s,t) f(x-s, y-t), \qquad (1.10)$$

where $f(x,y)$ is the original image, $g(x,y)$ is the filtered image, ω is the filter kernel. Depending on the desired result, various kernels can be used to filter the image. The simplest kernels are fixed for the entire image, such as the Gaussian kernel, which yields a low-pass filter (*i.e.* a filter that keeps the low frequencies in the image, meaning the region with low contrast), good for filtering noise at the cost of introducing blurring in the image. We define the image $G_\sigma(I)$ as the image filtered with a zero-mean Gaussian kernel of standard deviation σ, defined by $G_\sigma(p) = \frac{1}{\sigma\sqrt{2\pi}} exp \frac{-p^2}{2\sigma^2}$. Other, more complex filters use kernels that vary depending on the image content. For example, the bilateral filter is an edge-preserving and noise-reducing smoothing filter [51].

Morphological filtering is also often used in image processing to grow, shrink, remove or fill image regions based on shape characteristics. To perform morphological filtering, Mathematical Morphology provides relevant tools. If *Morphology* refers to the study of shapes, *Mathematical Morphology* refers to the theory based on mathematics and informatics that describes shapes using sets which is particularly used in image processing purposes [52, 53, 54, 55]. The most basic morphological operators are *erosion* and *dilation*, that aim to describe the interaction between the considered image and a *structuring element*, at each possible position of the structuring element on the studied image. In practice, the structuring element is a neighborhood window as described above, associated with the max operator for the dilation, and the min operator for the erosion, as for the median or convolutions above. This structuring element is often small compared to the studied image. Erosion and dilation are often used for removing noise or artefactual structures from images. The principle of erosion can basically be described as the removal of points of the object where the structuring element does not fit, allowing to remove noise or concavities of the object. Dilation is the dual operator that removes points of the background where the structuring element does not fit, allowing to fill convexities of the object. For binary images, these operators as defined as follow. Considering a set X of a binary image, and a structuring element $\mathit{\Gamma}$, both subsets of $\mathbb{Z}^2$. The *erosion* of X by $\mathit{\Gamma}$ is defined by:

$$\epsilon_{\mathit{\Gamma}}(X) = \{x | \mathit{\Gamma}_x \subset X\}, \qquad (1.11)$$

where Γ_x is the structuring element Γ translated by the vector $\overrightarrow{Ox}$, where O is the point of coordinate (0,0). The dual operation yields to the *dilation* of X by the structuring element Γ defined by:

$$\delta_\Gamma(X) = \{x | \breve{\Gamma}_x \cap X \neq \emptyset\}, \tag{1.12}$$

where $\breve{\Gamma} = \{-s | s \in \Gamma\}$ is the symmetric of Γ. Erosions and dilations are dual to one another, in the sense that if we refer to $\underline{X}$ as the background, or complement of X, *i.e.* $\underline{X} = \mathbb{Z}^2 \backslash X$, then $\delta_\Gamma(\underline{X}) = \underline{\epsilon_\Gamma(X)}$, and vice-versa. Most morphological operators come in dual pairs in this fashion (Figure 12).

Binary mathematical morphology can also be generalized to greyscale images (Figure 13). In this case, the *erosion* of a point x by a structuring element Γ that delimits a neighborhood is defined as the minimum of the neighbor points:

$$[\epsilon_\Gamma(I)](x) = \min\{I(y) | y \in \Gamma_x\}. \tag{1.13}$$

Similarly, the *dilation* of a point x by a structuring element Γ that delimits a neighborhood of x is defined as the maximum of the neighbor points:

$$[\delta_\Gamma(I)](x) = \max\{I(y) | y \in \breve{\Gamma}_x\}. \tag{1.14}$$

Many morphological operators derive from these elementary operators. Among them, we introduce the *morphological gradient* of the image I by a structuring element Γ. This operator is commonly used for edge detection, and is defined by:

$$grad_M(I) \equiv \delta_\Gamma(I) - \varepsilon_\Gamma(I). \tag{1.15}$$

An illustration is shown in Figure 12 for binary images. In this dissertation, the morphological gradient is used with a disk-shape structuring element Γ_{r_1} of radius $r_1 = 1$.

The *morphological opening* of X by the structuring element Γ is defined as the composition of erosion and dilation by the same structuring element Γ: $\gamma_\Gamma(X) = \delta_\Gamma(\varepsilon_\Gamma(X))$. Openings are used to suppress small bright structures in a dark background. The *morphological closing* of X by a structuring element Γ is defined as the composition of dilation and erosion of X by the same structuring element Γ: $\varphi_\Gamma(X) = \varepsilon_\Gamma(\delta_\Gamma(X))$. Morphological closing allows to remove small

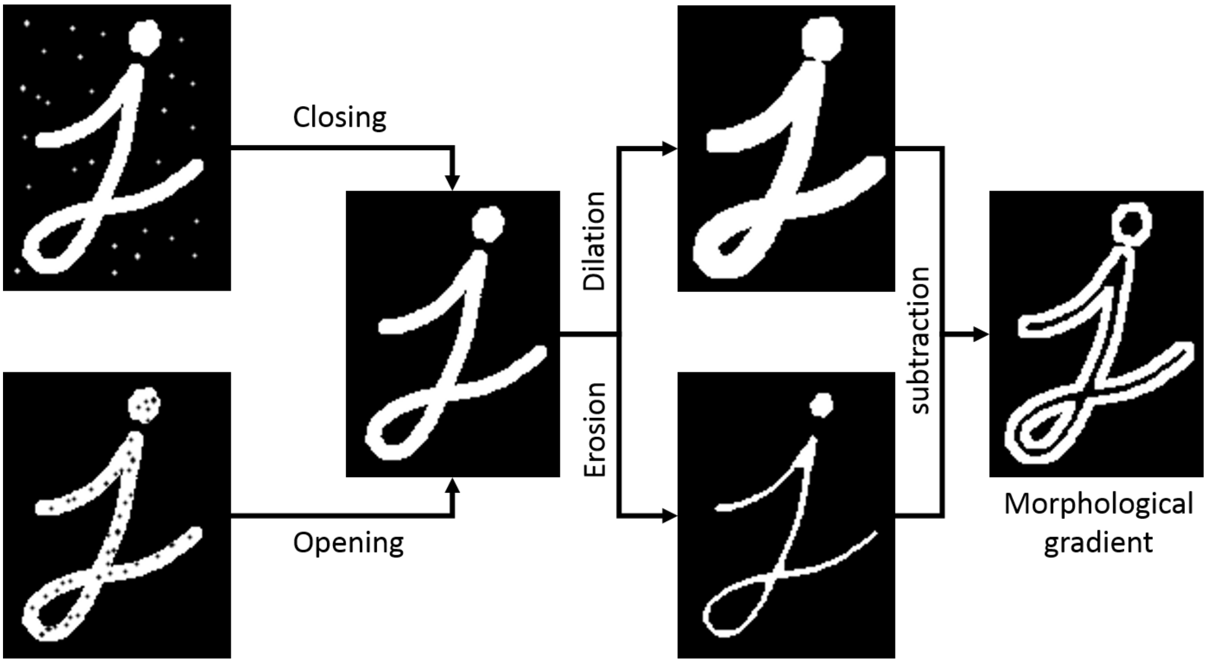


Figure 12. Binary morphological operators. A morphological closing is applied to an image with white noise (up left), allowing to remove the white artefact pixels. A morphological opening is applied to an image with black noise (down left), allowing to fill the black holes. Both lead to the centered denoised image. This image is then either dilated or eroded. The subtraction of the dilation and the erosion leads to the morphological gradient (right).

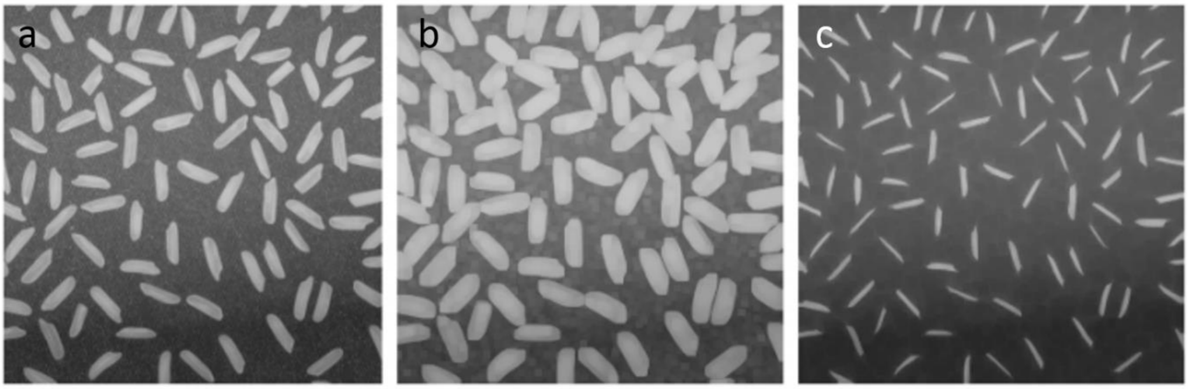


Figure 13. Greyscale dilation and erosion. a: the initial greyscale image. b: the image dilated by a disk-shaped structuring element. c: the dual eroded image.

dark objects over a light background. Openings and closings with the same structuring elements are dual to one another like erosions and dilations (Figure 12).

Algebraic openings are an extension to the notion of morphological openings, referring to a transformation that has the three properties of being *increasing* (that is, they preserve the order of the sets they are applied on), *anti-extensive* (that is, the transformed image is less than or equal to the original image), and *idempotent* (that is, multiple applications of a same operation do not change the result of the initial application). Algebraic openings can always be interpreted as the supremum (or union) of morphological openings with several structuring elements. Algebraic closings can similarly be interpreted as the infimum (or intersection) of morphological closings. Several types of algebraic openings are used in this dissertation to filter or extract particular structures from images. The *area opening* of X with area parameter α is denoted $\gamma_{\lambda}^{\alpha}(X)$ and is an algebraic opening that removes from the image all the light structures that are smaller than the area α [56]. We now present the notion of *radial opening* [57]. Let ρ_{ϑ}^{τ} be a line segment of length τ and orientation ϑ. The radial opening γ_{τ}^{ρ} is the algebraic opening obtained by taking the supremum (*i.e.* the pointwise maximum) of all the morphological openings $\gamma_{\rho_{\vartheta}^{\tau}}$ using ρ_{ϑ}^{τ} as structuring element, with ϑ varying between 0 and π:

$$\gamma_{\tau}^{\rho}(I) = \vee_{\vartheta \epsilon [0,\pi]} \gamma_{\rho_{\vartheta}^{\tau}}(I). \tag{1.16}$$

Intuitively, this opening preserves all structures in the image that can contain at least one segment of length τ in at least one orientation. When I is a binary image, the supremum operator $\vee$ reduces to the set union $\cup$.

Algebraic closing refers to a transformation that has the three properties of being increasing, *extensive* (that is, the transformed image is greater than or equal to the original image), and idempotent. Among them, we introduce the *convex hull* of a component of a binary image that corresponds to the smallest convex set containing this component [94].

Derived from morphological opening and closing, we then introduce the *top hat* transform by a structuring element Γ, defined by: $TopHat_{\Gamma} = I - \gamma_{\Gamma}(I)$, and its dual *bottom hat* transform, defined by: $BottomHat_{\Gamma} = \varphi_{\Gamma}(I) - I$. These operators are used to extract small bright objects and small dark objects respectively, from an image.

Such morphological filtering and thresholding are sometimes not sufficient to extract interesting edges and shapes from images, in particular, when luminosity and contrast are not the same everywhere on the image. In this case, a more robust transformation need to be applied. The *watershed* transformation was firstly introduced in the seventies as a tool for greyscale image segmentation [58]. This operator is intuitively defined as in hydrology, in a greyscale image whose intensity can be assimilated to a three-dimensional terrain where valleys correspond to dark areas of the image, and peaks correspond to the light areas of the image. This representation of the image as a topographic relief allows to take the intensity gradient into account for edges detection. Shapes borders (watershed line) can be intuitively considered as the set of points that delimit adjacent catchment basins. Thus, a drop of water falling in this line may flow down towards these adjacent catchment basins. The morphological gradient is often taken as the relief on which to compute the watershed. The intuitive reason for this is that the morphological gradient is an edge detection method, and the watershed can be used to form closed contours based on this detection. Watershed transformation is considered as a fundamental tool used for many segmentation procedures [59, 60].

Among process that derive from basic morphological operators, skeletonisation consists of thinning an initial shape, until obtaining a one-pixel-thick shape composed of an ensemble of curves that are centered in the initial shape. This ensemble of curves is called the *skeleton* of the initial shape. Skeletons are often used as a simplification of the original data, which facilitates shape recognition or registration. Thinning operators can be defined as similar to erosions, but with a topological constraint. Thus the resulting skeleton has the same topological structure as the input object, *i.e.* the same number of holes. With this simple definition, a unique finite connected component without any hole in 2D is reduced to a single point (ultimate homotypic thinning, also called *ultimate skeleton*). Additional constraints can be used to modify the aspect of the skeleton, such as defining a constraint set which is preserved from the thinning

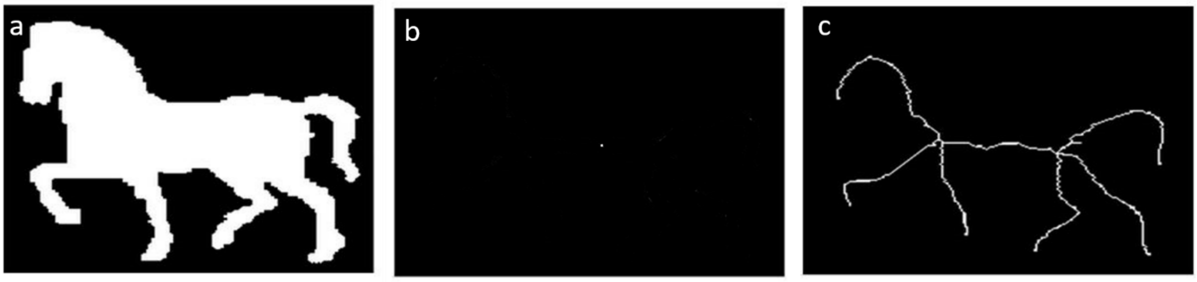


Figure 14. Skeletonisation process. a: initial binary image representing one connected component without any hole. b: the ultimate skeleton of the initial image is reduced to a single point. c: curvilinear skeleton, that preserves the protrusions of the initial image.

process. For example, a constraint set can be used to keep curve extremities to create a *curvilinear skeleton*. This skeleton preserves, in addition to the topology, the geometrical characteristic of the object, such as protrusions [61]. An illustration of an ultimate skeleton and of a curvilinear skeleton is shown in Figure 14 for a binary image.

1.2.3.2. Shape description

Once shapes of interest are identified on the image, they delimit a region from which information can be extracted. This is the principle of *shape description* or *characterization*. We call *features* (or *descriptors*) the measured parameters that are representative of the information we aim to extract. Thus shape description involves the identification and measurements of these parameters.

Shape characterization can rely on geometric features. Parameters related to size, such as *area*, *length*, *perimeter* can be measured directly from the pixels of the region on interest. The *circularity* descriptor is defined by: $circularity = \frac{4\pi \times area}{perimeter^2}$, and can be used to characterize the elongation of a shape. This ratio is maximal and equal to 1 for a disk, and decreases as the elongation becomes more pronounced. When analyzing a one-pixel-thick shape, such as the skeleton, approximation by a parametric function with a known shape (parabolic or sinusoidal for example) can help to characterize the curve shape. Regularity can be described by angles between two segments. Convexity and concavity can be characterized with some morphological operators, such as convex hull described in the previous section: the shape is convex if it is identical to its convex hull.

Relevant features can be extracted from the intensity distribution of an image I or of a considered shape in an image I. Among them, we define the *scalar average* of an image I: $average(\{p, \forall p \in I\})$ and the *scalar median* of an image I: $median(\{p, \forall p \in I\})$. The *scalar variance* of an image I is denoted $variance(\{p, \forall p \in I\})$ and measures the spread of the intensity distribution of an image. It is defined as the average of the squared deviations from the average intensity. When working on images only, these three parameters are referred to as *average*, *median* and *variance* respectively. When working on a $2D + t$ video sequence $\mathcal{V}$ (as in Section 2 of this dissertation), an intensity distribution is obtained for each pixel of the video through time, allowing to measure a scalar average, a scalar median or a scalar variance for each pixel. The respective resulting 2D images are referred to as the *sequential average*:

$$\forall p, SeqAverage(p) \;=\; average(\{\mathcal{V}_i(p), \mathcal{V}_i \in \mathcal{V}\}), \tag{1.17}$$

the *sequential median*:

$$\forall p, SeqMedian(p) \;=\; median(\{\mathcal{V}_i(p), \mathcal{V}_i \in \mathcal{V}\}), \tag{1.18}$$

and the *sequential variance*:

$$\forall p, SeqVariance(p) \;=\; variance(\{\mathcal{V}_i(p), \mathcal{V}_i \in \mathcal{V}\}). \tag{1.19}$$

Computerized procedures have been introduced to work on image classification purposes. We will see in the following section how such procedures can be applied to embryos classification, and thus to the development of automated toxicological assessment methods on embryos.

1.3. Automating embryos classification

As said before (Section 1.1.3.2), teratogenicity assessment of chemicals consists of classifying alevins according to the presence or the absence of anomalies and is performed manually by observing each embryo under a microscope. This process is time-consuming and subjective. The core problematic of this work is to propose ways to improve these assays by automating them. To do so, image analysis appears promising.

1.3.1. Image analysis applied to fish analysis: state of the art

Automated image classification is a complex issue frequently met in toxicological screening. In cells analysis for example, various simple segmentation tools combined with mathematical morphology operations have been widely used in the context of histopathology, with cells aggregates segmentation, quantification and clustering [62, 63, 64, 65, 66, 67]. Recently, new image processing software packages have been developed in order to simplify some of the most tedious biologists tasks as cells counting and differentiating, determination of tissue topology, wound healing, etc. [68]. Nevertheless, such methods concern the analysis of cell, which is transparent, unchanging regardless to orientation and thus simple. As a complete organism, the fish embryo model is more complex. Three axes allow to characterize it: dorsoventral, anteroposterior and left-right (Figure 2). All parts of the model do not have the same optical properties. In particular, some organs are more or less transparent, making analysis of the fish embryo model by image processing more challenging. In the last decades, progress in image processing enabled the development of new automated methods for fish embryo analysis. Nonetheless, the literature shows that almost all automated fish embryo-based toxicological studies are only performed on the zebrafish model [92]. Thus, these protocols will have to be adapted before using them on the medaka model.

1.3.1.1. Fish analysis using fluorescence microscopy

While developmental toxicology assessments on fish embryos are still often manually performed [69, 70, 71], image processing tools and pattern recognition have been increasingly used in alevins studies [72]. Many functional and structural studies focus on fluorescence imaging. This involves the use of transgenic lines of fish embryos that express fluorescent protein in a specific cells population of a specific organ. This process facilitates observations of the structure of interest and makes it easier to identify chemicals that modulate gene expression [73]. Some examples are the quantification of neural patterns in the spinal cord of zebrafish, or the detection of chemicals that cause yolk malabsorption [74, 75]. In [76], the authors provide a semi-automated imaging pipeline for the analysis of zebrafish kidney. The process requires manual positioning of embryo in a custom designed tool made in agarose, which limits throughput due to increased handling complexity. Many studies focus on the analysis of the development of the cardiovascular function and on angiogenesis. Blood vessel morphology is studied for example by quantifying intersegmental vessel [77, 78, 79]. This requires a high resolution that our experimental setting do not provide. The performance of the circulatory system is also measured by detecting heart beats or studying the circulation based on video motion analysis [78, 80, 81, 82]. In [78] and [83], authors propose a complete pipeline to assess body length, heart rate, intersegmental vessel area, circulation, pericardial area and circulation. Nevertheless, the processing of most of these features are not fully automated, which implies to take time to manually correct the data. A whole system is also described in [84] for high throughput screening of zebrafish embryos, including embryo dispensation, compound delivery, incubation, imaging and analysis.

From a general point of view, fluorescence studies are complicated to automate. They require costly fluorescence microscopes and high-sensitivity cameras for fluorescence imaging. The study must be conducted in a dark room. In particular, these studies are limited to the use of transgenic lines of zebrafish embryos. Transgenic lines of medaka embryos are more complicated to provide than zebrafish, as this model is less frequently used. For these reasons, fluorescence studies are not adapted to our experimental settings. For our purpose, we need to develop automated methods based on bright-field microscopy.

1.3.1.2. Fish analysis using bright-field microscopy

For now, only a few studies were published that propose automated phenotype recognition of fish embryo using bright-field microscopy, avoiding fluorescence and staining. These studies mostly focus on the analysis of basic phenotypes such as the lethality [39, 85], hatching [39], eyes [40], and pigmentation [40, 86]. Based on histogram analysis for the extraction of image color and texture-related information, the authors of [39] propose a phenotype recognition model for high throughput screening based on the classification of zebrafish images according to three basic phenotypes which are hatched, unhatched and dead. Nevertheless, when performing toxicological screening, other more complex phenotypes are visible, such as axial malformations, swim bladder alterations or edemas, that are not handled with these approaches.

A methodology is proposed to detect two different types of tail malformations: up and down [87]. This method performs classification with an accuracy of 95%. Nevertheless, this method is limited to the analysis of embryos seen in lateral view, which implies to manually position each embryo in the well. This process takes time and is not compatible with our experimental constraints. Indeed, when embryos are immerged, they generally maintain a balanced orientation, that cannot be changed during manipulation. The software Cellomics® Zebratox V4 BioApplication of Thermofisher is a commercialized tool developed for the morphological analysis of fish embryos. This application is used in [88] to detect a wide range of malformations, by combining the endpoints of a basic visual assessment made by humans with the Cellomics® data parameters. This results in a linear regression model named Computational Malformation Index used for the detection of alevin's malformations in tested images. The software allows to reduce the 30-minute long detailed visual assessment to a brief 10-minute long basic visual assessment per 96 well plates, which is promising. Nevertheless, visual inspection remains a necessity with this protocol, which implies to assess gross abnormalities in head/eye, pectoral fins, swim bladder inflation, organ definition and well as position (i.e., floating, upside down, on side). In addition, this studie implies to fix embryos after killing. Fixation is a technique used for conserving and stabilizing cells and tissues. It requires the embryos incubation in a solution composed of highly toxic agents (paraformaldehyde) during a whole night at 4°C. Embryos delivery is not adjustable as it depends on the husbandry conditions, which is external. Thus, the analysis must be performed on Friday, which prevents from extending the manipulation duration.

A very recent study proposes the first automated method for assessing heartbeats of multiple zebrafish per well, at 2 days post-fertilization, in bright-field images. The ability of the method to analyze several alevins in a same well allows to increase the screening throughput, while circumventing both the fluorescence and anesthesia required for previous studies made on heartbeat detection. Nevertheless, this method implies to record 10-second-long high-resolution videos. Such data are heavy and require a high storage capacity. In particular, our current settings do not allow to record video with such a high resolution (pixel size in our images is about 12µm, instead of 2µm in [89]).

Recently, the FishInspector software was developed aiming to assess several functional and morphological markers of the zebrafish development using the VAST BioImager™ Platform Overview from Union Biometrica for alevin manipulation and orientation control. The Vertebrate Automated Screening Technology (VAST) system is a high-throughput platform designed for cellular-resolution *in vivo* screening of zebrafish alevins. In particular, it is designed for the automation of zebrafish alevins manipulation. The system loads each alevin from reservoirs of multiwell plates into a capillary. The alevin is positioned and oriented into the capillary before imaging takes place. This instrument can yield images of alevins in multiple orientations as required by the user [90, 91]. However, several limitations make the use of this tool incompatible with the assay described in this manuscript. Firstly, it is designed for zebrafish and is not compatible with medaka, which are larger and cannot be loaded into the capillary. Secondly, even if the assay was adapted to zebrafish, this tool imposes the analysis of straight sedated alevins that can fit in the capillary. However, under the conditions of our teratogenicity test, some tested chemicals can cause important malformations on the alevin, such as a high spine curvature or a large edema. These strongly malformed embryos would not fit into the capillary. Even without using the VAST Bioimager for alevins manipulation, the FishInspector software can adapt to the analysis of images acquired in other experimental conditions. Adjustable parameters are included to compensate for camera or microscope-dependent differences. The software provides a user-friendly interface that allows features annotations with flexibility, and with a high future development potential. Currently, the features already proposed include alevin shape (length and area), tail curvature, eye, yolk and head size, pericard analysis for heart rate quantification, swim bladder and jaw anomalies. If the process remains supervised, in the sense that manual interaction is still frequently required to correct the detection of some features, such as pericard and jaw, it can also be conducted blind. Nonetheless, the main limitation for our purpose is that analyzed images must be acquired after a precise orientation of the embryo in lateral view, which is the less frequent orientation met in our experimental conditions [92].

Several articles have shown the efficiency of supervised learning techniques in the scope of fish embryos phenotypes classification. Among the machine learning-based classification methods that are most often used, we find Support Vector Machines (SVM), as described in Section 1.2.2.2. In [39], an accuracy of about 97% is obtained for the recognition of basic zebrafish embryos phenotypes which are dead, hatched and unhatched. The authors of [40] detect alevins without eyes with an accuracy of 89% and over-pigmented phenotypes with an accuracy of 99%. Indeed, these basics phenotypes are easy to characterize with a few and well-discriminative descriptors, that makes SVM reliable to achieve good classification results. Nevertheless, for the recognition of under-pigmented alevins for example, the presence of shadows interferes with features extraction, leading to a lower accuracy of 79% [40]. This example demonstrates the complexity of data characterization for the analysis of a specific phenotype. Thus, for the detection of more complex phenotypes, non-parametric methods could be tested for improving the results.

In [41], a supervised learning approach of extremely randomized trees is applied to distinguish a wide range of phenotypes, from dense random subwindows extracted from images. The classification is performed in two times. First, a classifier distinguishes 3 basic phenotypes which are “dead” (necrosis), “chorion” (non-hatched eggs) and “other”, with almost 100% of accuracy. Second, a binary classifier is applied on images of the “other” category, to separate images into two classes: the “normal” (healthy) phenotype, and more complex defects, including axial abnormality, necrosed yolk, edema and hemostasis (small amount of blood which can be located everywhere in the embryo). This results in a success rate between 90 and 95%. This study demonstrates the sensitivity of supervised learning algorithms for the classification of various defects, that could be used for our purpose. Nevertheless, as for most of the methods proposed to classify complex fish embryos phenotypes, it is limited to the analysis of alevins seen from a specific orientation [87, 88]. Each alevin is manually positioned and oriented on a high-viscosity support (melt of methylcellulose) before starting the image acquisition. Such process requires manual intervention and is time consuming. Moreover, validation is performed only by comparing classifier results and annotations made on the acquired images. As 2D images only represent a single point of view, it may occult some abnormalities that would be visible by observing in all possible points of view under a microscope.

1.3.2. Objectives and challenges

In this project, we aim to automate a screening test developed for the assessment of chemicals teratogenicity, by developing robust and efficient automated methods based on image processing and machine learning. These methods must result in the automated classification of embryos according to the presence or the absence of abnormalities. The automation must minimize the time taken by manual operation during the screening test, and increase the results subjectivity. Mortality assessment is required to calculate LC50, and malformations assessment to calculate EC50 (Section 1.1.3.2).

To do so, an image acquisition platform has been designed by the Environmental Research Department of L'Oréal for acquisition of bright-field images of the plates. This platform is composed of a light platform and a moving camera (Figure 15a). The plates, which contain one embryo per well, are put directly on the light platform. During acquisition, the camera records one image and one video per well, each well containing one embryo (Figure 15c). The acquired image represents the whole well (Figure 15b). Because the embryo is immerged in its medium, it may end up in any possible orientation from the dorsal to the lateral view and anywhere in the well after anesthesia. As the acquired image is in 2D, a single alevin orientation is visible on it, which is not the same for every alevin. Depending on the alevin orientation on the image, all abnormalities are not visible. For example, a malformation of the alevin's spine may be visible when looking at the dorsal view and not when looking at the lateral view. The alevin's pigmentation is not homogeneous on the whole body and may sometimes occult information like the heartbeat. This is especially true for eggs in which the embryo is strongly folded. Some anomalies may be visible but not the same way depending on the orientation. For example, the

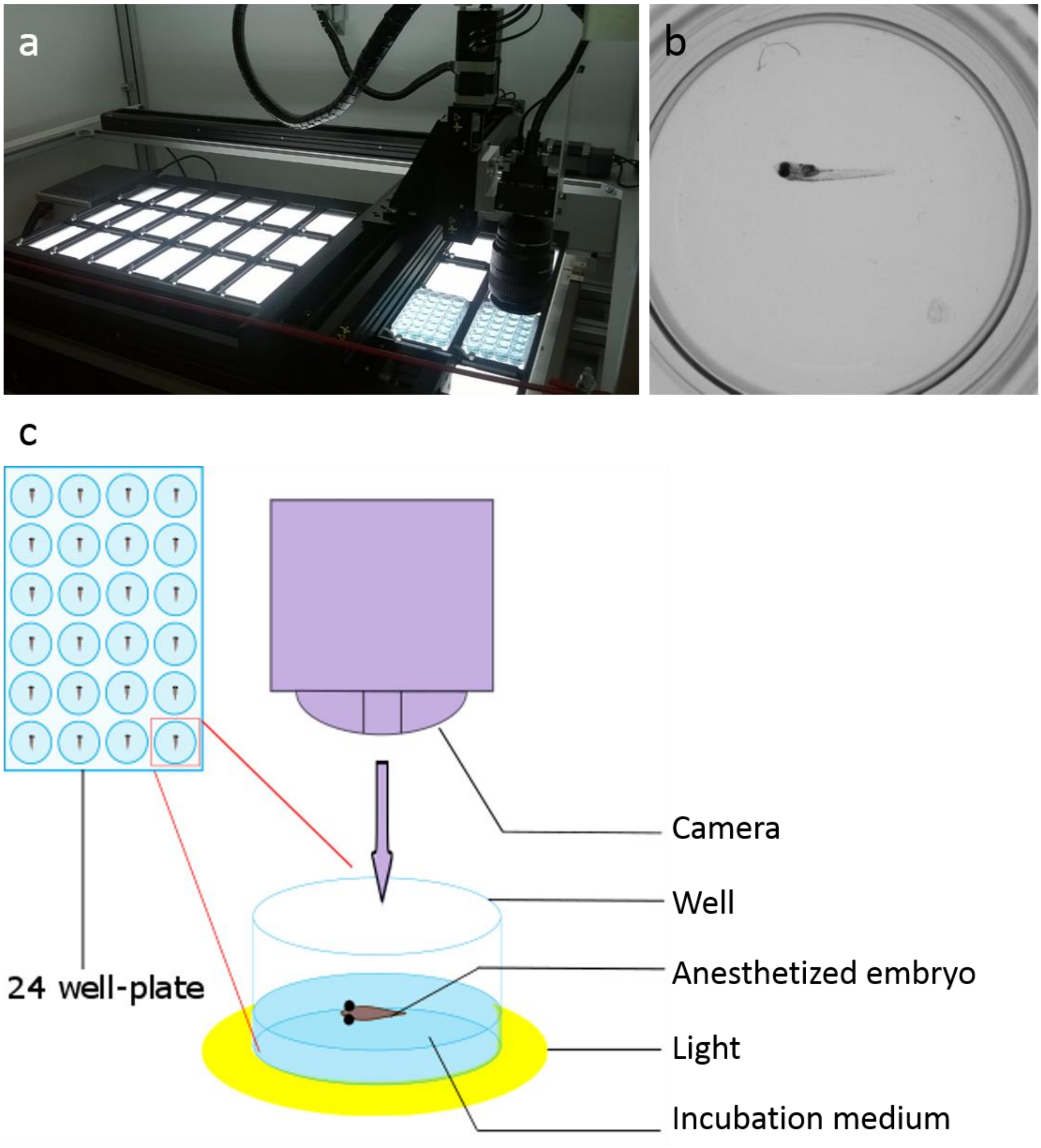


Figure 15. Acquisition set-up. a: images and video acquisition device with a light platform and a moving camera. b: example of image acquired with the acquisition device. c: diagram of the acquisition set-up. The camera moves above each well of the 24-well plates and takes a picture and a video of the whole well with one anesthetized embryo inside.

swim bladder does not have the same shape when seen in dorsal and in lateral view. In contrast, observations made under a microscope allow to see the embryo in any possible orientation, to zoom in and to focus on it. In this work, a significant challenge is to handle the issue of information loss between microscope-based observations and information which is visible on 2D images. In particular, this information loss must be quantified, and adaptive methodologies must be developed.

Another technical challenge concerns the subjectivity linked to the ground truth. Subjectivity is a reason of the teratogenicity assessment automation, but it also causes difficulties during the development of automated methods. As the ground truth relies on observations made by an operator, inter-operator and intra-operator subjectivities, as defined in Section 1.1.3.2, have an impact on the reliability of the ground truth. For this reason, subjectivity must be assessed and quantified. Moreover, ground truth will change through time with the experience of the operator that makes the observations. Thus, a strategy must be set up to adapt the program to the ground truth.

Data recording may sometimes fail, leading to an empty or a partially recorded image or video, that must be eliminated. As medium renewals are automatically performed with a device during the 9-day long incubation, some embryos may be sucked up, leading to an empty well, that must be detected. Embryos may end up anywhere in the well after anesthesia, even close to the well borders. On the image, such embryo appears partially hidden by the well borders, and is not usable. Sometimes, an embryo close to the well border may take the curved shape of the border, that could be wrongly attributed to an effect of the chemical. For these reasons, embryos which are too close to the well borders must also be detected. Moreover, the plates often contain dust and the remaining chorion after hatching, *i.e.* the membrane that covers the embryo when still in egg form. Thus, another challenge consists of developing a strong and robust pre-processing task in order to obtain a workable image of the embryo for further treatment. Such pre-processing task must take into account the sorting of images according to their usability, including the detection of empty images, of empty wells, of partially hidden embryos, and the distinction between the embryo under study and artefact objects in the well. Finally, as both eggs and hatched eggs (alevins) are present on studied images, the pre-processing also must distinguish between them to adapt further image processing methods.

As with any screening test, the efficacy of our automated classifiers must be assessed in terms of sensitivity and specificity. Sensitivity corresponds to the proportion of abnormal alevins correctly detected, and specificity is the ratio of healthy alevins correctly detected. 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, in this manuscript, both specificity and sensitivity must be optimized, which corresponds to the conventional choice of optimizing the overall accuracy. Nevertheless, as explained in Section 1.1.3.2, our assays are expected to detect strongly teratogenic chemicals, while avoiding raising alarm on other chemicals without an actual toxic effect. For this reason, at equivalent overall accuracy, we decided to favor the specificity.

2. Mortality assessment: automated classification of medaka embryos according to the detection of cardiac arrests

In this section, we describe a mortality assay for automatically classify video sequences of medaka embryos in two classes: dead or alive. After a pre-processing step that includes the detection of unusable videos, the differentiation between eggs and alevins, and the localization of the region of interest in the embryo, the algorithm relies on intensity variation on this region of interest to detect a heartbeat. From an initial dataset of 3192 videos, 660 were discarded as unusable (20.7%), 655 of them correctly so (99.25%) and only 5 incorrectly so (0.75%). The 2532 remaining videos were used for our test. Compared to videos observations, 45 errors were made, leading to a success rate of 98%.

The work presented in this section has appeared in the following publications:

- E. Puybareau, D. Genest, E. Barbeau, M. Léonard, H. Talbot. “An automated assay for the assessment of cardiac arrests in fish embryo”, in *World Congress on Alternatives and Animal Use in the Life Sciences*, Seattle, United States, 2017 (poster).
- E. Puybareau, D. Genest, E. Barbeau, M. Leonard, H. Talbot. “An automated assay for the assessment of cardiac arrests in fish embryo”, in *Computer in Biology and Medicine*, pp 32-44, 2017.

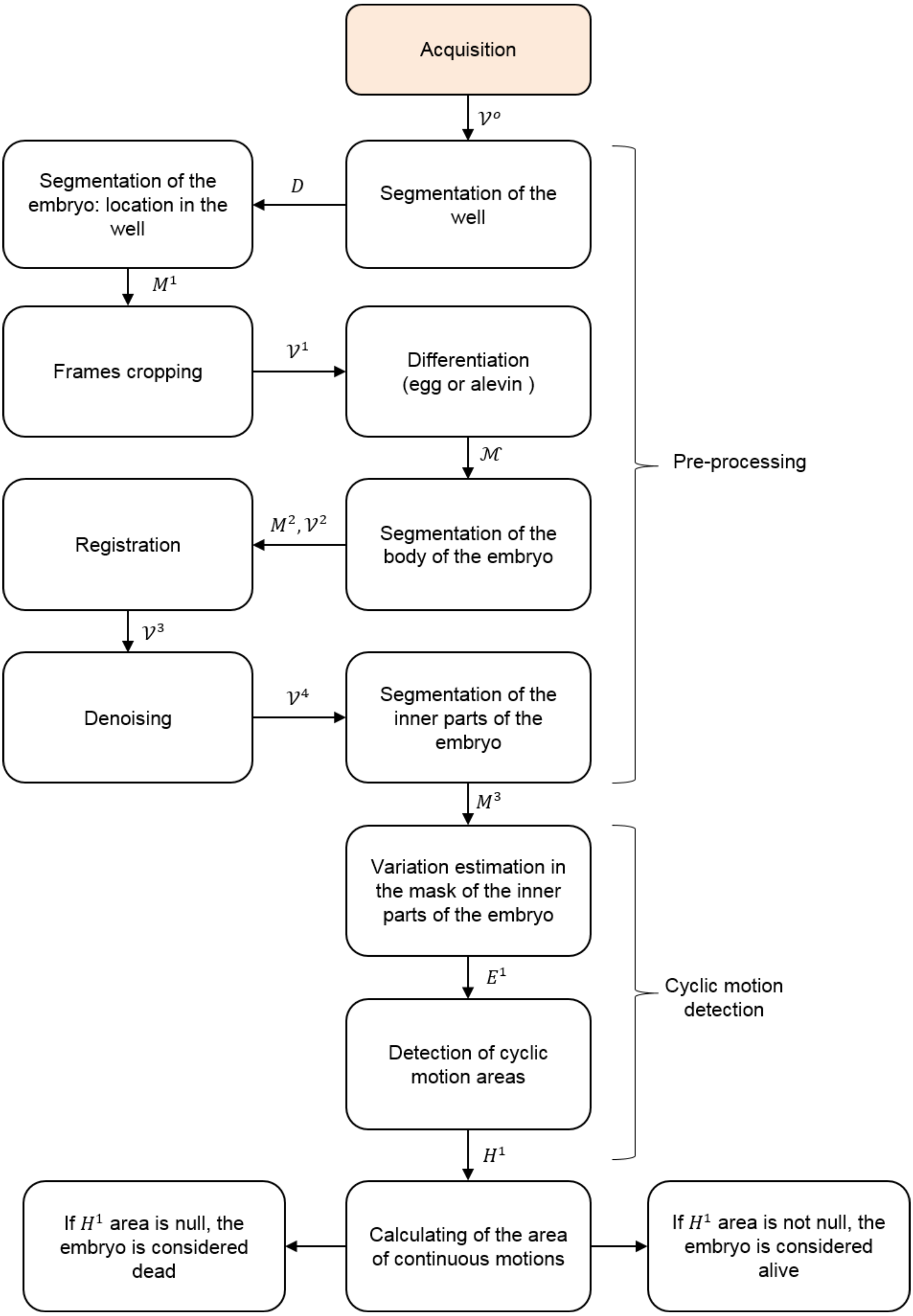

Acquisition
$\mathcal{V}^o$
Segmentation of the well
D
Segmentation of the embryo: location in the well
M^1
Frames cropping
$\mathcal{V}^1$
Differentiation (egg or alevin)
$\mathcal{M}$
Segmentation of the body of the embryo
$M^2, \mathcal{V}^2$
Registration
$\mathcal{V}^3$
Denoising
$\mathcal{V}^4$
Segmentation of the inner parts of the embryo
Pre-processing
M^3
Variation estimation in the mask of the inner parts of the embryo
E^1
Detection of cyclic motion areas
Cyclic motion detection
H^1
Calculating of the area of continuous motions
If H^1 area is null, the embryo is considered dead
If H^1 area is not null, the embryo is considered alive

Figure 16. Flowchart of the embryo mortality image processing assay

2.1. Introduction to the detection of cardiac arrests

With respect to the development of a toxicological assay based on the analysis of medaka embryos, the first considered endpoint is the viability of the embryos that will allow to calculate the LC50 described in Section 1.1.3.2. In this section, we aim to classify embryos into classes which are alive and dead embryos. As medaka embryos are transparent, their cardio-vascular system is readily visible, making possible the direct visual analysis of the heartbeat and thus, the detection of cardiac arrest. At these early stages of development, cardiac arrest may not induce an immediate death due to the blood gas exchanges that occur through skin diffusion [93, 94]. Nonetheless, we will refer to cardiac arrests as an indicator of mortality, since it can be considered as a prediction of mortality at later stages of development.

Here we describe an automated image-processing pipeline to detect a beating heart with minimal human interaction, maximum speed, and reliability. The proposed procedure, based on mathematical morphology [53, 54, 55], improves on a previous feasibility study [95], which had some limitations. In particular, it required a significant workload involving gel preparation and the manual positioning of embryos on the support gel. In addition, the number of plates used was limited due to a moving light platform.

Our procedure is part of a complex process for detecting morphological and functional abnormalities in fish embryos after a 9-day exposure. This endeavor imposes some experimental constraints: we have to deal with both eggs and alevins at the analysis level. Because this process is intended to be fully automated, and image processing procedures may change depending on if the analyzed embryo is an egg or an alevin, a differentiating procedure between eggs and alevins must be developed. Medaka hearts normally beat at a frequency of around 130 beats per minute (bpm) i.e. 2.2Hz [96]. However, this can vary between 0 and 300 bpm, i.e. 0-5Hz, in extreme cases. To avoid incorrect measurements, our recordings must be made using a frame rate that is high enough for our purpose. Our current camera records one-second-long videos at 30Hz, corresponding to a Nyquist cutoff frequency of 15Hz, which is sufficient. Recorded videos are 1500×1500 pixels in size, covering the whole well. Because the incubation medium is liquid, undesirable motion may be present during acquisition. Embryos may also slide to the edge of each well, rendering them partially or totally invisible. To minimize this, the platform is fixed and we use a moving camera (Section 1.3.2). The quantity of liquid in the well is also carefully adjusted, so that the embryo moves as infrequently as possible once it is placed in the well. Incomplete or otherwise corrupted videos may occasionally be acquired. Shadows and undesirable objects are also a risk. These unusable sequences must be identified at the start of processing.

Even with the above precautions and even if all of the embryos were to remain immobile under anesthesia, some residual movement is still possible. Such motion may be caused by involuntary reflex swimming or it may be induced by vibrations and shocks in the lab environment, which are easily transmitted by the liquid in the well. As we rely on variance measures to detect the heartbeat, artefactual motion caused by even the slightest uncompensated frame motion may induce areas of high variance and, in the end, generate false positives. In particular, the eyes can cause significant difficulties as they constitute the darkest part of the embryos' bodies and are not transparent. As a result, their contours have high contrast that may induce false positives on dead embryos in case of eyes vibration. Unfortunately, the heart is fairly close to the eyes in medaka, so this problem needs to be handled carefully. Moreover, while still in egg form, embryos appear tightly wound and the eyes can obscure the heart, making the detection of a heartbeat impossible. More generally, eggs have different visual characteristics compared to alevins. This is why it is important for the application to determine whether a well contains an egg or an alevin so that the processing procedures can be adjusted accordingly.

During the nine-day incubation period in the chemical compound under study, the medium is regularly changed. The ninth day, a fixed quantity of medium is removed from the well so that only 0.5mL of liquid remains during image acquisition. The medium may still contain dejection products, impurities, or even the chorion if the egg hatched during incubation. The real embryo must be carefully distinguished from these impurities during image processing.

We aim to complete the sequence analysis within the same time-frame as the acquisition, *i.e.*, in under 10 seconds. We therefore propose a robust pipeline suitable for production usage. It consists of simple operators, which are fast and, for the most part, available in off-the-shelf image analysis software packages. Figure 16 presents the flowchart of our assay. It is split into two phases: a pre-processing step for sequence stabilization and denoising, and an actual processing step for detecting significant periodic changes in the embryo assuming they are caused by its beating heart. Videos are read as raw data interpreted as grey-level values.

The video pre-processing step is described in Section 2.2. The image analysis solution developed to detect heartbeats based on the intensity variation of the video sequence is then described in Section 2.3, including the detection of the region of interest for heartbeat research, and the heartbeat detection within this region of interest. The classification method is assessed and discussed in Section 2.4.

2.2. Video pre-processing

We present our notations and image processing operators, mostly from mathematical morphology as described in Section 1.2.3.1 [52, 97]. Let I be a $M \times N$ pixels grey level image, taking 8-bit discrete values, *i.e.* $I\colon\{1,\dots,N\} \times \{1,\dots,M\} \rightarrow \mathbb{Z} \cap [0,255]$. Γ_{r_i} is a disk structuring element of radius r_i of size i; $\delta_{\Gamma_{r_i}}(I)$ is the dilation of I by the structuring element Γ_{r_i}. $\epsilon_{\Gamma_{r_i}}(I)$ is the dual erosion, $\gamma_{\Gamma_{r_i}}(I)$ and $\varphi_{\Gamma_{r_i}}(I)$ the corresponding morphological opening and closing. The notation $\gamma_{\times}^{\alpha}(X)$ denotes the area opening of the set X with area parameter α [56]. We also introduce the radial opening γ_{τ}^{ρ} defined from the line segment ρ_{ϑ}^{τ} of length τ and orientation ϑ as structuring element [57]. The binary image which is the thresholded image of I above value θ is denoted $(I)_{\geq\theta}$.

We introduce the notation $\mathcal{V}^l$ to refer to a video sequence l of n images. We note $\mathcal{V}_i^l$ the frame i of this sequence. For clarity reasons, we write $\mathcal{F}^l = \mathcal{V}_0^l$ the first frame of the video sequence $\mathcal{V}^l$.

2.2.1. Video quality control and detection of unusable videos

We begin by determining which videos present important and undesirable changes during the sequence. These changes may be due to the presence of black frames, shadows, or large uncontrolled motion. For this, on each difference d_i between two successive frames of the 30-frames long video sequence $\mathcal{V}^o$ we compute the statistical variance (as defined in Section 1.2.3.2):

$$\forall i \in [1{,}29], d_i = \mathcal{V}_i^o - \mathcal{V}_{i+1}^o, \tag{2.1}$$

$$V_i = variance\ (d_i). \tag{2.2}$$

In the case of a correctly recorded video, two successive frames should be very similar, and their pixelwise difference yields a near-zero output, so its variance remains small. On the contrary, if a large motion appears on a frame, the difference will show a high contrast. If at least one of all computed variances is higher than the experimentally determined threshold (set to 30), the video sequence is deemed unusable.

2.2.2. Segmentation of the well and selection of a region of interest

Embryo segmentation is crucial for several reasons. For speed and reduced memory usage, we crop the area of interest to a small window centered on the embryo. During this step, we also detect sequences where the embryo is not fully visible, *i.e.*, too close to the edges of the well. Moreover, motion stabilization must be performed on the embryo itself, and not on other elements in the field of view. We first isolate the region of interest by finding the disk area corresponding to the inner part of the well. This step also removes all objects connected to the edges of the well. The procedure for finding the area of this disk is as follows: edges of the disk appear dark, so we first compute a so-called bottom-hat filter: see Figure 17a and Section 1.2.3.1 for a definition. In this equation, a disk structuring element $\Gamma_{r_{20}}$ is chosen to remove small artefacts in the well. This yields image A^0 (Figure 17b), which we binarize via an automated thresholding operation to obtain image A^1 in Figure 17c. The Otsu automated thresholding, described in Section 1.2.3.1, is chosen [50], with the added constraint that the foreground area must be in the interval [20,000; 40,000]. This interval is experimentally determined to guarantee that the edges of the well are present in the foreground. This

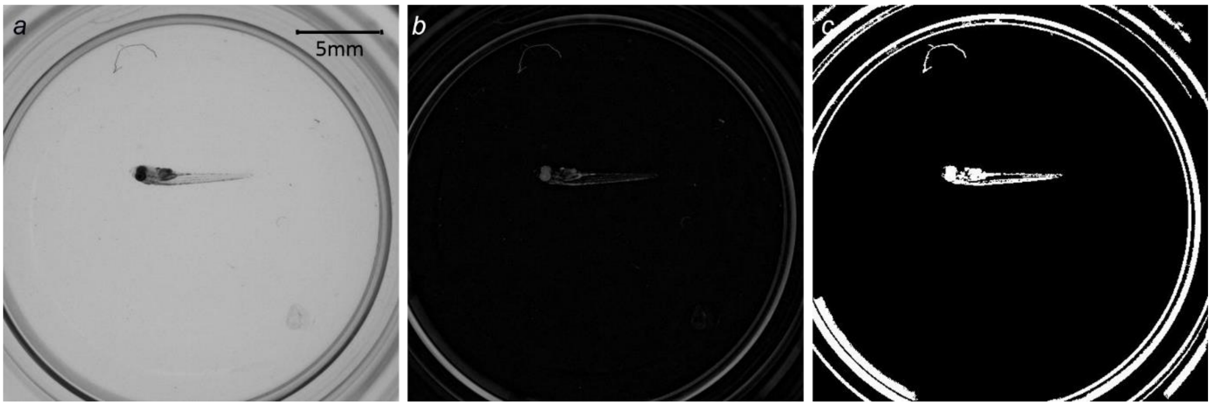


Figure 17. Bottom-hat application. a: frame $\mathcal{F}^0$ before bottom-hat. b: result of bottom-hat A^0. c: subsequent thresholded image A^1.

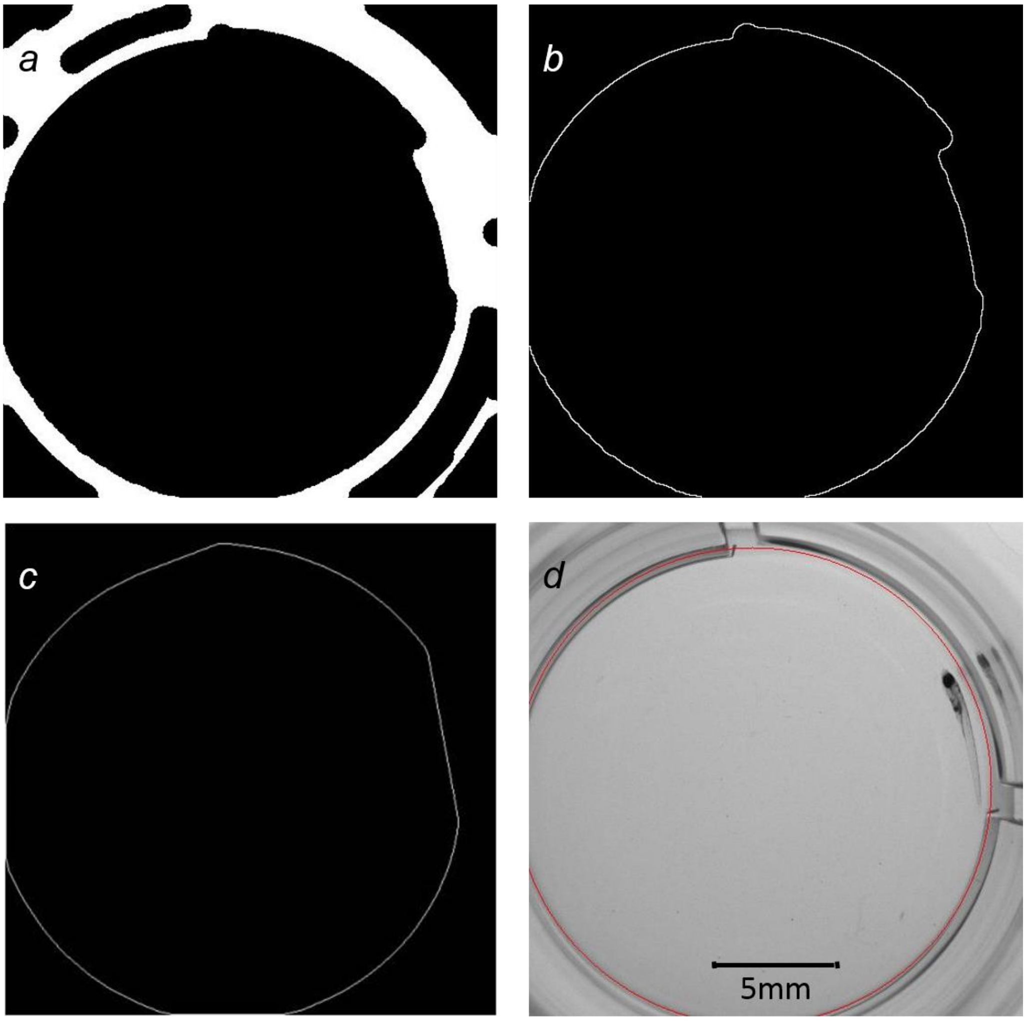


Figure 18 . Segmentation of the inner part of the well. a: the image before application of watershed algorithm (image A^2). b: result of watershed (image A^3). c: result of convex hull (image A^4). d: outline of the final result D superimposed on $\mathcal{F}^0$.

constraint is convex and easily implemented: we consider all thresholds in order from the highest to the lowest. The foreground area necessarily increases during this process. In the acceptable foreground area interval, we select the threshold with the highest Otsu criterion. We call this threshold θ_{co} (for constrained-Otsu):

$$A^O = BottomHat_{\Gamma_{r_{20}}} (\mathcal{F}^0), \tag{2.3}$$

$$A^1 = (A^0)_{\geq \theta co}. \tag{2.4}$$

We remove small components from the well with an area opening $\gamma^{\alpha}{}_{100}$ of parameter $\lambda = 100$, followed by a morphological closing with a ball $\Gamma_{r_{40}}$ to reconstruct fragmented edges of the well. Then a radial opening $\gamma^{p}{}_{100}$ with linear element p of length $\tau = 100$ is applied to remove short artefacts from the well, while retaining the thin well borders [57]. This yields image A^2:

$$A^2 = \gamma^p_{100}(\varphi\Gamma_{r_{40}}(\gamma^{\alpha}_{100}(\mathrm{A}^1))). \tag{2.5}$$

From this result shown on Figure 18a, we only want to keep the internal ring that represents the separation between the interior of the well and its edges. For this, we use a well-established morphological approach to segmentation, based on the Watershed transform [59, 98]. We compute the magnitude of the Derivative of Gaussian filter: $DoG = \nabla \star G\sigma$ using the Deriche recursive implementation of the gradient operator for speed with parameter $a = 10$ [99] (see Section 1.2.3.1 for a definition of the Gaussian filter). We then use a markers-based Watershed algorithm on the magnitude of this gradient [59]. A disk at the center of the frame is taken as internal marker m^1_{ext} and the frame corners are the external marker m^1_{ext}. We write:

$$A^3 = watershed\ (||\ DoG_{10}(A^2)||,\ m^1_{int},\ m^1_{ext}). \tag{2.6}$$

The resulting contour is shown on Figure 18b. The result may be incorrect if the embryo is too close to the edge of the well. To avoid this, we expand the contour using the smallest convex set that contains A^3 [100]. We call A^4 the resulting image (see Figure 18c) and G^4 the set of points contained in the central component of A^4. We compute the barycenter C of coordinates (a, b) and the diameter $2r$ of G^4 as the largest width or height of its bounding box. The final well segmentation is the disk D centered in C and of radius r:

$$D = \{(x, y), (x - a)^2 + (y - b)^2 \leq r^2, \tag{2.7}$$

with $(a, b) = barycenter(G^4)$ and $r = \frac{max(width\ (G^4), height\ (G^4))}{2}$.

Its contour is shown on Figure 18d.

2.2.3. Localization of the embryo in the well

Our image analysis procedure is intended to work for both alevins and eggs, but some eggs do not develop at all and differ markedly from healthy eggs and alevins (Figure 19a,c). They feature low contrast, which makes them look like empty chorions or impurities that can develop in the wells. An early pipeline challenge is to reliably detect and identify the embryo in each well. To achieve this, we begin by performing an initial segmentation adapted to all components in the well, whatever their level of intensity or variance. In the previously calculated bottom-hat image A^0 (Figure 17b), all components of interest are easy to classify as connected components located strictly inside the segmented well. We call h the *contrast significance*, understood as the intensity variation that connected components must have to be considered significant [55]. The value h is experimentally set to ignore the irrelevant intensity variations due to noise, while still detecting the dimmest components that cannot be ignored (*i.e.* the undeveloped eggs). We call p_{peak}^i the local maximum of intensity in the neighborhood of the pixel p^i. Image B^1 contains the so-called *h-maxima* of $A^0 \cap D$, defined as follows:

$$B^1 = \{p^i \in A^0 \cap D \; with \; p^i = \begin{cases} 0 \text{ if } \left(p_{peak}^i - p^i\right) > h \\ p^i \text{if } \left(p_{peak}^i - p^i\right) > h \end{cases}. \tag{2.8}$$

The image B^1 of the h-maxima can be efficiently computed using a morphological reconstruction operator, as explained in [56].

Several components can be detected in the resulting frame B^1. These components may be embryos, empty chorions, or some type of impurity. To identify the embryo, we use several criteria: presence of eyes, minimal and average intensities, variance, and circularity. A

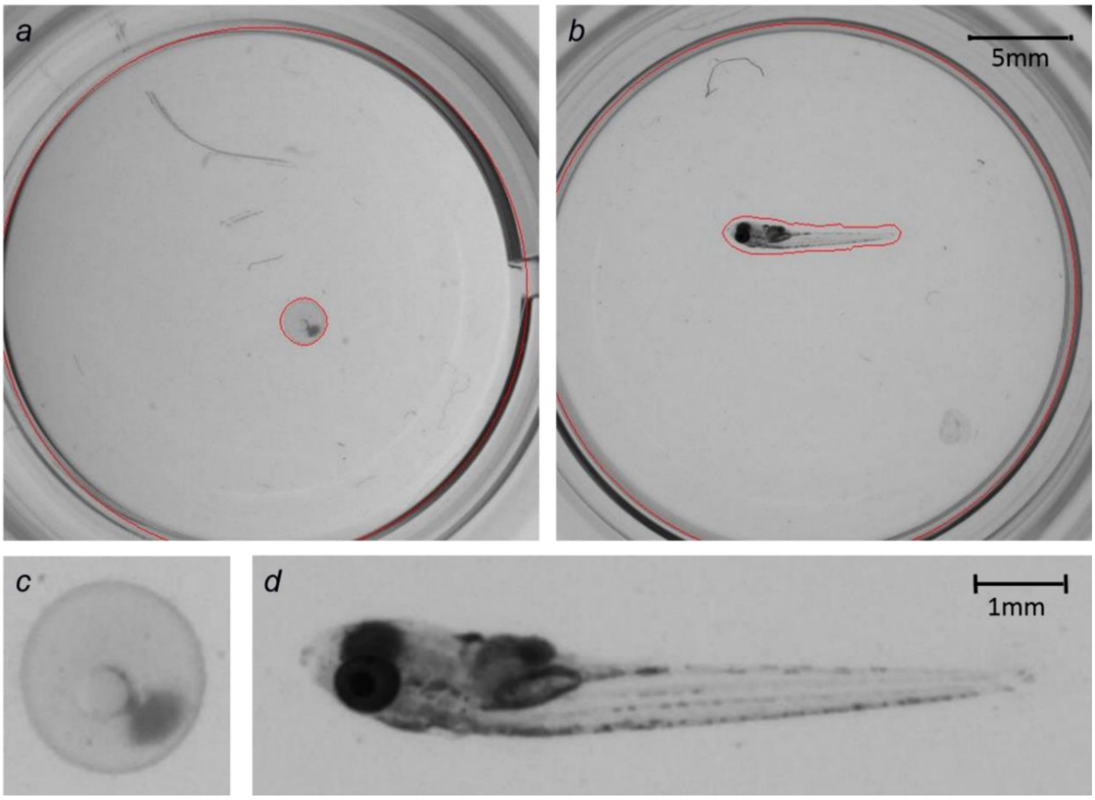


Figure 19. Segmentation of the well and location of the embryo. a and b: the red lines show the outlines of D and M^1 on the initial frame $\mathcal{F}^0$. c and d: first frames of cropped sequences $\mathcal{V}^1$.

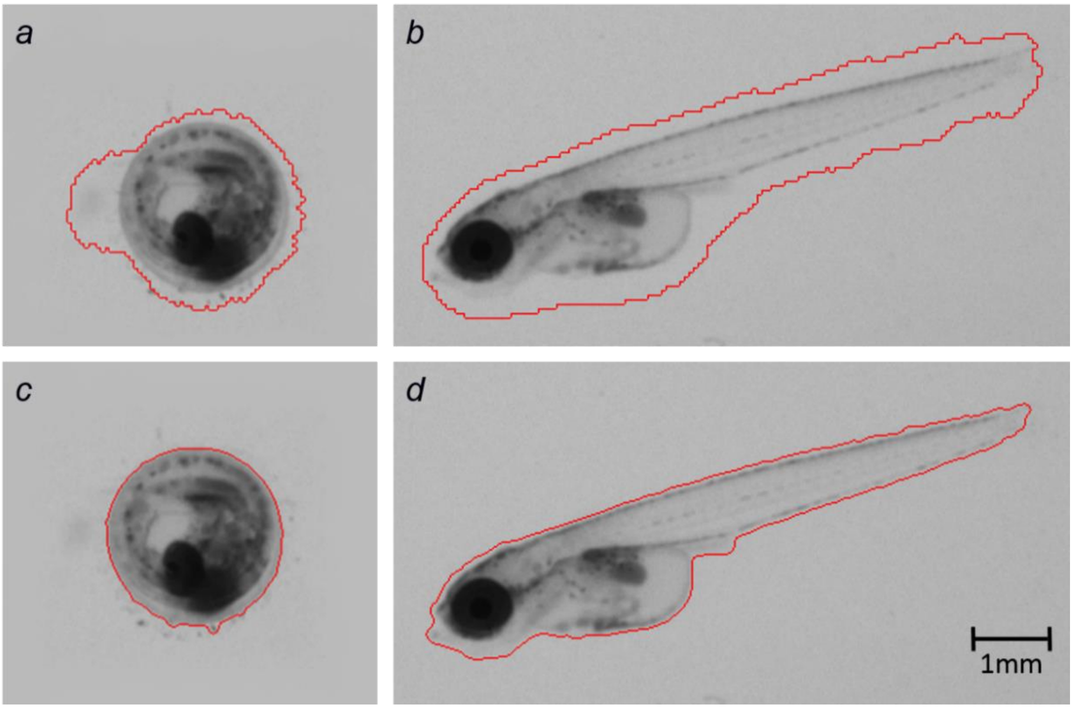


Figure 20. Segmentation of the embryo. a and b: the red line shows the outline of the mask M^1 (before adjustment). c and d: the red line shows the outline of $\mathcal{M}$ (after adjustment).

component is considered to have an eye if an extremely dark spot, representing less than a quarter of its total area, is present. Since impurities are generally homogeneous, this procedure filters out dark impurities, that are uniformly dark, and light impurities and chorions that are evenly bright. However, it can also filter out under-developed eggs. To avoid this problem, we add further classification criteria: a high average intensity or a low variance. We also verify circularity to differentiate between undeveloped eggs, chorions, and bright impurities. Finally, this process enables us to classify components as either "under-developed eggs," "other embryos," or "impurities and chorions." We delete components identified as impurities or chorions. There must only be one embryo per well. Therefore, if several components classified as "under-developed eggs" remain after this step, only the largest is kept. If several components of the other classes remain, we only keep the largest component among those from the "other embryo" class. Indeed, we have experimentally found that it is more difficult to distinguish underdeveloped eggs from chorions than other embryos from impurities. Thus, the probability of making a mistake from the "undeveloped egg" class is higher. The result M^1 is a binary mask (with values in $\{0, 1\}$) containing only one component expected to locate the embryo in the well (see its red contour in Figure 19a,b). If the result is empty, this means that the embryo intersects the edges of the well and the sequence cannot be reliably analyzed. If we find an embryo instead, we crop the sequence and the mask M^1 by defining a bounding box around our segmentation, dilated by Γ_{r_2}. This results in a new video sequence denoted $\mathcal{V}^1$ centred on the embryo (Figure 19c). However, because of contrast variations and the large variability of grey levels between embryos, the mask M^1 is only approximate. In particular, for alevins, it delimits a rough area with the tail included (Figure 19b) and potentially contains some shadows and impurities if they are too close to the embryo. For the purpose of heartbeat detection, we need to exclude the tail from the search field.

2.2.4. Differentiation between eggs and alevins

Because they have different visual properties, it is necessary to identify the embryo type for further processing. The differentiation step is based on the morphological analysis of the embryo contours previously detected. The previous segmentation provides a reliable localization of the embryo, but only a rough approximation of its contours (Figure 20a,b), so these need to be refined. For this, we consider the first frame $\mathcal{F}^1$ of the cropped sequence $\mathcal{V}^1$.

In particular, it is crucial to weed out potential shadows and impurities, which may have been segmented with the embryo, while retaining the tail segmentation for the alevins. We apply the bottom-hat procedure defined in Equation 2.3 of section 2.2.2 to eliminate the background.

Then we experimentally define an adaptive threshold slightly above the average pixel intensity near the border of the cropped frame. For our images, with a 8-bit depth, an increment of 5 was experimentally determined as appropriate: $\theta = average\ (\mathcal{F}^1) + 5$. We obtain a binary image, whose small components are filtered out with an area opening with parameter $\lambda = 5$. We apply the morphological gradient to the resulting image (as defined in Section 1.2.3.1), to obtain image C^1:

$$C^1 = grad_M \left(\gamma_5^{\alpha}((\varphi_{\Gamma_{r_{40}}}(\mathcal{F}^1) - \mathcal{F}^1)_{\geq\theta})\right). \tag{2.9}$$

In order to properly extract the contours of the embryo, without confusing them with residual artefacts that may still be present in the background, we again use a markers-based watershed methodology on the image C^1 as follows:

$$m_{int}^2 = skeleton\left(\gamma_{\Gamma_{r_8}}\left(\varepsilon_{\Gamma_{r_{15}}}(M^1)\right)\right), \tag{2.10}$$

$$\mathcal{M} = watershed(C^1, m_{int}^2, m_{ext}^2). \tag{2.11}$$

The image outline is set as the external marker m_{ext}^2, and the ultimate binary skeleton of the eroded and opened mask M^1 is set as the internal marker m_{int}^2 [101]. The erosion and the morphological opening are respectively performed with a radius-15 and a radius-8 disk, in order to remove potential thin impurities linked to the previous embryo segmentation. The outline of the resulting binary mask $\mathcal{M}$ is shown in Figure 20c,d.

We now use the shape of $\mathcal{M}$ to differentiate between the eggs and the alevins. Eggs are highly circular, so we can use the classical circularity attribute, that depends on the area and the perimeter of the binary shape under study:

$$circ = \frac{4\pi\times area}{perimeter^2}. \tag{2.12}$$

This ratio is at most equal to 1 for a disk and decreases as the elongation becomes more pronounced. However, it is still possible for some alevins to be so tightly wound that their associated binary shape presents a high circularity. Eggs also possess hairs on their chorion that may reduce the circular aspect of their associated mask. To correctly differentiate between both cases, we also consider two other criteria. We have experimentally determined that a healthy well-segmented egg has a radius of around 60 pixels. Therefore, allowing for some margin of error, we apply a morphological opening $\gamma_{\Gamma_{r_{40}}}$ that deletes the alevins'mask, as alevins are much thinner than eggs. If the component under study is filtered out during this step, it is considered to be an alevin. If it is not, we determine the minimum enclosing disk of the mask and we calculate the area difference d between this disk and the mask $\mathcal{M}$. Indeed, since the eggs's hairs are uniformly distributed on the chorion, the difference between the mask area and the area of its minimum enclosing disk is higher for hatched alevins than for eggs. Below an experimentally determined threshold of 3,000 pixels, we consider the component under study to be an egg. Otherwise, we conclude that it is an alevin.

2.3. Image analysis solution

2.3.1. Search of the heart region

2.3.1.1. Segmentation of the alevin's body

In this manuscript, three main segments of the alevin are referred to as the *head*, the *trunk* and the *tail*. We use the term *body* to refer to both the head and the trunk. These segments are illustrated in Figure 21.

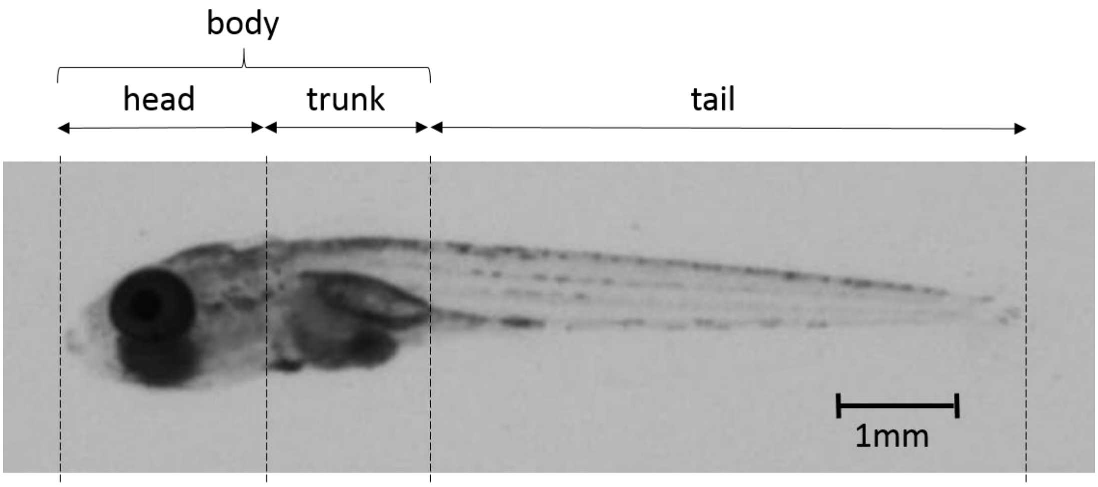


Figure 21. Alevin's head, trunk and tail segments.

For the purpose of heartbeat detection, it is essential to restrict the region of interest to the alevin's body, in order to minimize the probability of false detection due to electronic noise or blood flow in the bright tail regions. Therefore, after the differentiation step, we refine embryo segmentation in the case of alevins. Alevins are darker than the background and their eyes, in particular, are very dark. They are fairly easy to segment as a large connected component associated with the darkest minima in the body region. We apply the same threshold process as in Equation 2.4. Because we have experimentally determined that the minimal area of an alevin's body region is approximately 600 square pixels, we apply a morphological area opening using the criterion $\lambda = 100$ to eliminate small components. Moreover, we limit the body region to M^1, by computing the intersection:

$$M^2 = M^1 \cap \left(\gamma^{\propto}_{100}\left(\left(\mathcal{F}^0\right)_{\leq \theta_{oc}}\right)\right). \tag{2.13}$$

The result M^2 is a new binary mask representing the alevin's body (Figure 22b). We crop the sequence and obtain a new sequence $\mathcal{V}^2$ centered on this area (Figure 22c). These remained unchanged for egg sequences, and M^2 and $\mathcal{V}^2$ are respectively equivalent to $\mathcal{M}$ and $\mathcal{V}^1$.

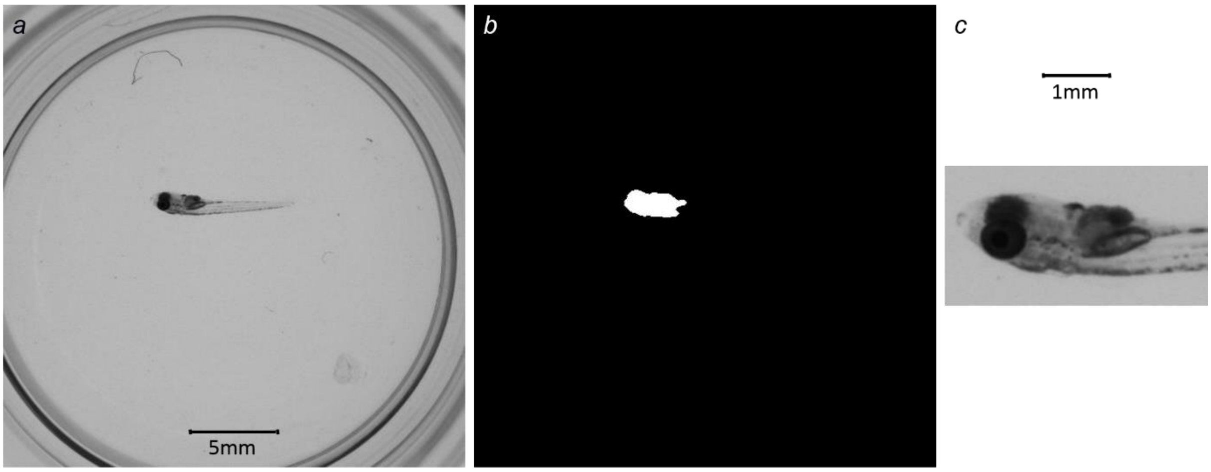


Figure 22. Segmentation of the initial frame to locate the trunk of the alevin. a: initial frame $\mathcal{F}^0$. b: trunk mask M^2. c: first frame of $\mathcal{V}^2$.

2.3.1.2. Registration

Even when anesthetized and subject to vibration isolation, embryos may still move slightly during acquisition. In order to eliminate false positives, all sequences need to be stabilized. For efficiency, we chose a keypoint based methodology, specifically using SIFT [102]. SIFT detects and matches pairs of significant points P_1 and P_2 between pairs of frames. This allows us to solve the equation for rigid transformations:

$$P_1 = P_2 \times R + T. \tag{2.14}$$

Here, $T = (d_x, d_y)$ is the translation vector, and R is the rotation matrix. Since embryos do not deform significantly, it is sufficient to consider this class of transforms. Our model can select between translation-rotations and translation-only transformations. Model selection is a useful feature, because simpler models are usually more robust. Here we distinguish between pure translation (where R is the identity matrix) and translation-rotation by computing the sum of square difference between the two model outputs. If they do not differ significantly, we choose the simpler model. This latter outcome is the more frequent in our experiments. Pure rotation never occurred in our experiments so we do not consider that model. Since impurities may be present in the well, movement is often visible around the embryo. Thus, in order to stabilize the sequence with respect to the embryo and not the other moving components in the well, we ensure that keypoints in the embryo only are selected, restricting key-points to the mask M^2. Taking the first frame of the sequence as reference, the selected model transform is applied to all the following frames. In order that the whole stabilized sequence be of constant width and height, we consider the bounding box of the sequence, and crop it by the maximum displacement in both x and y, which are respectively $max\ (|d_x|)$ and $max\ (|d_y|)$. We call $\mathcal{V}^3$ the stabilized and cropped video sequence of the embryo.

2.3.1.3. Denoising

Depending on the illumination, the sequence may be more or less degraded by noise. We use a bilateral filter in the 2D+t domain to reduce noise [51]. We can interpret the bilateral filter as a neighborhood-dependent convolution (Section 1.2.3.1).

At each pixel (i,j) belonging to a window W, the filtered frame I_D is given by:

$$I_D(i,j,t) = \frac{1}{\sum_{(k,l,m)\in W} \omega(i,j,t,k,l,m)} \sum_{(k,l,m)\in W} I(k,l,m) \times \omega(i,j,t,k,l,m), \qquad (2.15)$$

where:

$$\omega\ (i,j,t,k,l,m) = \exp(\ -\frac{\left((i-k)^2+(j-k)^2+(t-m)^2\right)+}{2^{\sigma_d^2}} - \frac{I(i,j,t)-I(k,l,m)^2}{2^{\sigma_\gamma^2}}). \qquad (2.16)$$

In this formula, I is the original input image. Depending on parameters, this filtering could be too strong and could cause the heartbeat to become undetectable. Experimentally, the best parameters for removing the noise without altering the heartbeat are: $window\ size\ =\ 3 \times 3 \times 3$, $\sigma_r\ =\ 0.5$, $\sigma_d = 0.6$. The outcome of this process is a restored video sequence $\mathcal{V}^4$. Because the bilateral filter is not as effective on the borders of the sequence, it is preferable to remove them. This implies that we lose the first and the last frames, and so $\mathcal{V}^4$ is only 28-frames long.

2.3.1.4. Segmentation of the inner parts of the embryo

To ascertain the presence of a heartbeat in the body region, we look for cyclic motion in this region only. To prevent false motion detection in unrelated areas, we develop a mask M^3 corresponding to the region of interest. This eliminates areas most subject to noise, such that the eyes. We define D^1 as the sequential average image of the video sequence $\mathcal{V}^4$ (Section 1.2.3.2):

$$D^1 = SeqAverage(\mathcal{V}^4). \qquad (2.17)$$

Blood causes the heart and vessel to appear darker, so they are easy to segment as a large connected component associated with the darkest minima, simultaneously maximizing the inter-class variance

$$D^2 = (D^1)_{<\,(\theta_O+\theta_c)}. \qquad (2.18)$$

Here θ_O is obtained using the Otsu criterion. Because the heart and vessels are thin compared with the rest of the body, we need to bias the threshold to encompass a larger region, so we add a constant θ_c to the Ostu threshold, experimentally optimized to 20. The resulting D^2 is a binary mask of the registered body of the alevin. Considering D^2 as a geodesic mask, we now extract the eyes from D^2 as the one or two most prominent minima from its min-tree [103]. We cannot rely on the eyes being separated. Depending on the pose of the embryo, they may be merged. We write:

$$M^3 = \epsilon_{\Gamma_{r_1}}(\gamma_{\Gamma_{r_3}}\left((D^2.D^1)_{\geq(\theta_O-\theta_d)}\right) \tag{2.19}$$

In this equation, θ_O is again the Otsu optimal threshold, which depends on the distribution of grey-levels within D^2. Because we want to bias the threshold nearer to the eyes, which are very dark, we subtract an experimentally optimized constant θ_d, which turns out to be equal to 20 as well, from θ_O. The outline of the resulting mask M^3 in alevins is exemplified in Figure 23a,b. This procedure is used only on alevins in order to restrict the region of interest to detect heartbeats. It is not suitable for eggs, due to the folded aspect of the embryo. For these, we compute M^3 using the same procedure for segmenting the eyes but consider M^2 as the geodesic mask (Figure 23c,d).

2.3.2. Heartbeat detection

2.3.2.1. *Elimination of spurious, non-cyclic motion*

So far we have assumed that heartbeats can be associated with significant variations of grey-levels in the body region of the alevin. For this, we estimate a time-wise, grey-level variance at every location in this region. However, a significant variance during a sequence may be also due to a single, large, spurious motion instead of a regular, periodic heartbeat. To distinguish between these two cases, we split $\mathcal{V}^4$ into four equal length sub-sequences. Each subsequence is 7-frames long, which is enough to record a typical, single heartbeat. The subsequences are called $\mathcal{V}_i^4$, $i \in \{1 \ldots 4\}$. We now consider the sequential variance image $V_i = SeqVariance(\mathcal{V}_i^4), i \in \{1 \ldots 4\}$, and the sequential median $V = SeqMedian\{V_i\}$ (see Section 1.2.3.2 for a definition).

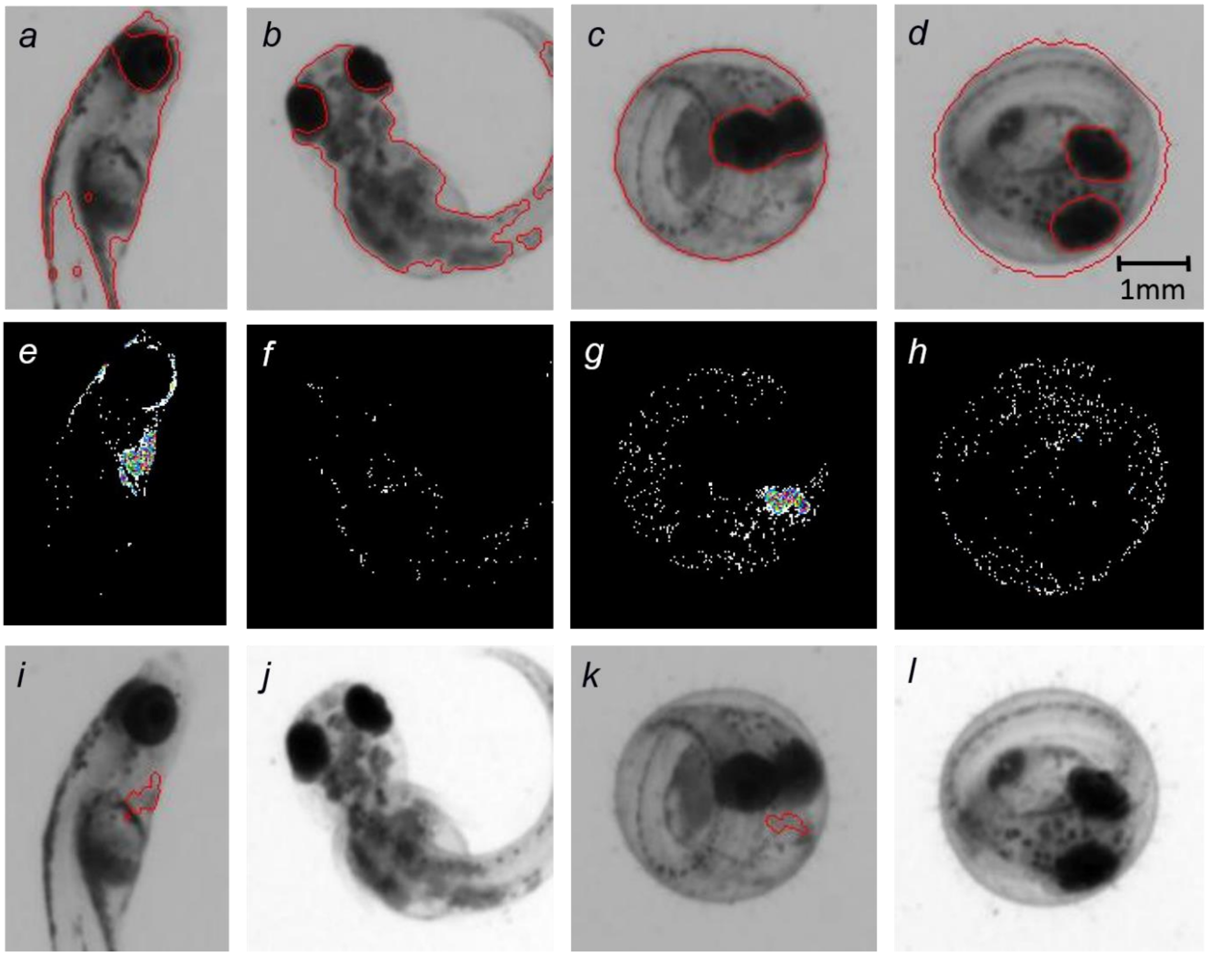


Figure 23. Steps of heartbeat detection method on two alevins and two eggs. a to d: segmentation of the inner parts M^3. e to h: false-color rending of the temporal variance E^1. i to l: segmentation of cyclic motion detection in embryos H^1. The first and the third columns are alive embryos; the second and the fourth are dead embryos.

We see if that a single, large, spurious record high values. The median V of the V_i will still be low. In contrast, if a regular, significant variations occurs in the majority of the V_i, then V will have high values that we assume to be due to periodic motion:

$$\mathcal{V}^4 = \mathcal{V}_1^4 \cup \mathcal{V}_2^4 \cup \mathcal{V}_3^4 \cup \mathcal{V}_4^4, \qquad (2.20)$$

$$\forall i \in [1,4], V_i = SeqVariance(\mathcal{V}_i^4), \qquad (2.21)$$

$$E^1 = M^3.SeqMedian(\{V_{i,}i \in [1.4]\}). \qquad (2.22)$$

The result is shown on Figure 23e to h.

2.3.2.2. Segmentation of cyclic motion areas

We obtain a binary image of E^1 via a small morphological closing, a threshold by the scalar median value μ of all strictly positive variances present in E^1, and a small area opening:

$$\mu = median(E^1), \qquad (2.23)$$

$$H^1 = \gamma_4^{\alpha}(\varphi_{\Gamma_{r_1}}(E^1))_{\geq\mu}). \qquad (2.24)$$

Because the visual properties of eggs are not the same as those of alevins, we notice more residual cyclic motion due to noise in the case of eggs. Therefore, we add an area opening with $\gamma = 8$ only for eggs. If the number of non-zero pixels in H^1 is zero, we consider the embryo to be dead, otherwise, it is considered to be alive (Figure 23i to l).

2.3.3. Detection of cyclic motion associated with the mouth

Some alevins have no detectable heartbeat in the sequence due to significant pigmentation in the body region. However, sometimes cyclic mouth motion induced by natural reflex demonstrates that the alevin is alive. To allow for these specific cases, we recommend the detection of cyclic mouth motion. Once again, we apply our cyclic motion detection and segmentation based on variance, but this time on the inverted mask M^3. If an area of cyclic motion is detected (M_{mouth}), we estimate the distance between each component in this area and the alevin's eyes (M_{eyes}), by superposing both corresponding masks:

$G^1 = M_{eyes} \cup M_{mouth}$, then performing a dilation of G^1 by a radius-14 disk: $G^2 = \delta_{\Gamma_{r_{14}}}(G^1)$. If a single area remains, the area of cyclic motion is close enough to the eyes to be the mouth. In terms of implementation, we keep the component of G^2 that contains the eyes, which we call G^3, and check to see if it also contains a component exhibiting cyclic motion (Figure 24):

$$G^4 = G^3.G^1. \tag{2.25}$$

If the number of components in G^4 is higher than the number of eye components, we consider it a cyclic mouth motion area. An alevin with no heartbeat detected but mouth motion present is considered to be alive.

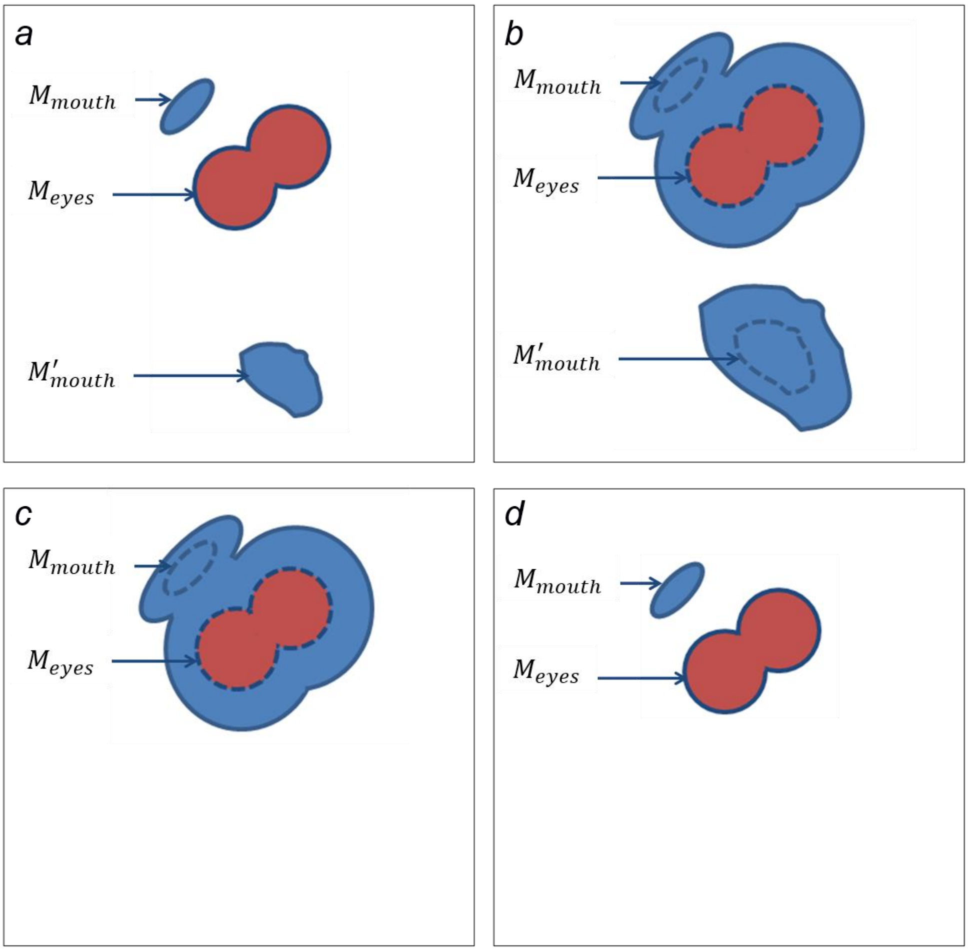


Figure 24. Distance assessment between areas of cyclic motion and the alevin's eyes. a: superposed eyes and areas of cyclic motion masks G^1. b: dilation G^2. c: component of G^3 which contains M_{eyes}. d: the result G^4.

2.4. Assessment of the classification between dead and alive alevins

In this section, we present the results obtained using a total of 3,192 videos, 2,532 of which were actually usable to test for heartbeat detection. We begin describing the experimental setup in part 2.4.1, before moving on to discuss processing. Our results are presented in part 2.4.2 We tackle the problem of remaining limitations in part 2.4.3 and finally discuss about the quality of our validation method in part 2.4.4.

2.4.1. Experimental setup

2.4.1.1. Experimental protocol

On the first day of the experiment, the individual fish eggs are manually placed in a 24-well plate, one egg per well, in an incubation medium that contains or not a pre-determined concentration of the water-soluble chemical under study [104]. After a 9-day exposure, 1.5mL of the incubation medium is removed from each well, and the fish embryos are anesthetized with tricaine. The final concentration of 0.18g/L has been shown not to affect the heart beats frequency within the time frame used for analysis [78]. The plate is then placed under the connector board, and the acquisition is automatically performed under the control of a Visilog Visual Basic script (Figure 15 in Section 1.3.2). For each well, 30 uncompressed video frames at a resolution of 1500 × 1500 pixels are recorded over a duration of 1 second with a monochrome camera. More details about embryo culture, exposure and video acquisition are provided in Section 6.1 of the Appendix.

2.4.1.2. Software and libraries

We used the Python 2.7 environment under Windows 7 (64 bits) in an HP computer with a 3.60 GHz Intel® Core™ i7-4790 CPU and 32 GB of RAM. We used Numpy, Scipy, and Scikit-image [105], Python Imaging Library (PIL), Pink [106], and Open CV for Python [107].

2.4.1.3. Database description and ground truth

For this study, as for all studies presented in this manuscript, the used database 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 the anomaly that is to be automatically detected, *i.e.* cardiac arrest in this part of the manuscript. The pictured alevins have been exposed to a wide variety of chemicals, including none. The nature of the chemical used is not significant for the purpose.

In the context of our test validation, two possible types of ground truth exist: observations under a microscope and those directly on acquired videos. Each present different advantages and drawbacks. The strongest way of assessing the quality of the complete embryo analysis process, including plate preparation, data acquisition, and data treatment, is to compare our results to the observations of embryos under a microscope. On the other hand, the automated method we developed works on video sequences whose information may be much different from the observations made under microscope. Several aspects linked to the experimental protocol or the acquisition method can explain this fact. (i) Video quality is such that some weak heart beats may be undetectable on video even if they are visible under a microscope. (ii) Observations made while using a microscope also depend on operator fatigue and subjectivity. (iii) Because there is a time gap between the observations made under microscope and videos acquisition, an embryo may also die during the interval. (iv) observations made under microscope facilitate scrutiny of the heart since embryos can be moved to a favorable position, whereas in videos the embryo's posture may obscure the heart. For these reasons, it appears that the most relevant way to assess the program's quality is through direct video observation.

For this study, both possible ground truths were considered. Table 1 summarizes the establishment of ground truth datasets. The first expert (named “expert 0”) originally observed the embryos under a microscope before the total set of 3,192 videos were acquired. We refer to these microscope-based observations as “Dataset 0”. The expert identified each case as an alive or dead embryo by checking for the presence of a heartbeat. Another expert (named “expert 1”) analyzed the resulting 3,192 acquired videos (“Dataset 1”). He began manually assessing the usability of the videos, by checking that they were complete, well-recorded, that the well was not empty, and that the embryo was not too close to the well boundary. From this selection process, 655 unusable videos were identified. Then, for the remaining 2,537 usable videos only,

	Dataset 0	Dataset 1	Dataset 2
Screening method	Microscope	Videos	Videos
Dataset size	3,192	3,192	200
Experts in charge of the analysis*	Expert 0	Expert 1	Experts 1, 2, and 3
Classification labels used	"Alive" or "Dead"	"Unusable," "Alive," or "Dead"	"Alive" or "Dead"

** Each expert screened the entire dataset.*

Table 1. Establishment of ground truth datasets. For Dataset 2, a consensus was reached between the three experts and a final set of 200 ground truth data was obtained.

this expert determined if the embryo was alive or dead. However, whereas determining videos usability is easy and thus reliable, detecting a beating heart is sometimes difficult and therefore subject to errors. For this reason, we selected a subset of only 200 usable videos ("Dataset 2") so that the health status of the embryo could be reassessed by two other independent observers ("experts 2 and 3"). In the end, three different observers were involved in analyzing the 200 usable data subset. Because the experts' observations pertaining to Dataset 2 were not always identical, a consensus was then reached between these three observers concerning the videos that they assessed differently. The 200 data resulting from this consensus represent the ground truth we use to validate our automated method, as explained in the following Section 2.4.2.

2.4.2. Results

Our heartbeat detection method returns three possible results: "unusable," "alive," or "dead." This method processes a sequence in approximately 10 seconds, in accordance with our initial constraints. All parameters were hand optimized using a training sample of 100 sequences.

Using Dataset 2 that contains 200 videos, the results of our program are compared to the previously established consensus data ground truth. We consider the program to be erroneous if it detects a dead embryo that was identified as alive according to expert consensus (false positive) or if, on the other hand, it detects an alive embryo that was identified as dead by expert consensus (false negative). We detected 3 errors made by the program for a corresponding error rate of 1.5% (Table 2a). This error rate only corresponds to false positive, meaning the corresponding sensitivity is 100% and the specificity is 98.1%. In toxicity tests, this is a more acceptable type of error since it does not provide a false sense of security with regard to the tested molecule. Moreover, the specificity is maximized, which meet the constraint exposed in the introduction of this manuscript. In Table 2b, we present the error rates that were calculated for each expert as compared to the final consensus based on Dataset 2. Expert 1, who processed the 3,192 videos (Dataset 1), has a similarity rate of 98.5% with respect to the consensus data. Consequently, the experts' observations can be considered sufficiently reliable to analyze the results of the entire program. The results of this analysis (program vs. expert 1) are described below.

Out of the initial Dataset 1 of 3,192 videos, the program correctly flagged 655 as unusable and incorrectly flagged another 5 as unusable due to some error within the program itself. If we consider the entire set, 3,187 videos were correctly flagged, leading to a success rate of 99.85%.

a.

Program results:	Dead	Alive
Ground truth (consensus):		
Dead	38 (19%)	0 (0%)
Alive	3 (1,5%)	159 (79,5%)

b.

	Expert 1	Expert 2	Expert 3	Program
Error rates between expert observations	1.5%	2.5%	1.5%	1.5%

Table 2. Results and error rates calculated on Dataset 3 (200 usable videos). a: distribution of dead and alive embryos in the program results compared to the ground truth data of the consensus. It shows that 1.5% of the Dataset 3 embryos were wrongly identified as dead by the program. b: error rates calculated for each expert and for the program versus consensus data, used as ground truth.

The remaining 2,532 videos were used for mortality test validation. There were 45 errors in this set, for an error rate of 1.77%. Such an error rate is low and can be considered satisfactory. We noticed that 11 of these 45 errors were due to embryos that had died a long time before acquisition and had consequently absorbed the blue marker. These embryos appear very dark on the video and are therefore more affected by noise, which was incorrectly labeled as periodic motion. This is something we can improve in a future version of our software pipeline.

2.4.3. Limitations and further optimizations

In some cases, even embryos that are dead may appear to move. This may be caused by movement in the water, fluttering, shadows, or embryo rotation inside the well (Figure 25). In dark areas, acquisition noise is proportionally more troublesome [108], and may be confused with cyclic motion. Sometimes, the embryo may appear to slide on the water. This happens if we do not correctly compensate for rotation in the sequence stabilization phase. The main remaining cause is ambiguity: in some cases, the heart beats so slowly or weakly that we cannot detect it. In most cases, a human operator would also have difficulties detecting it.

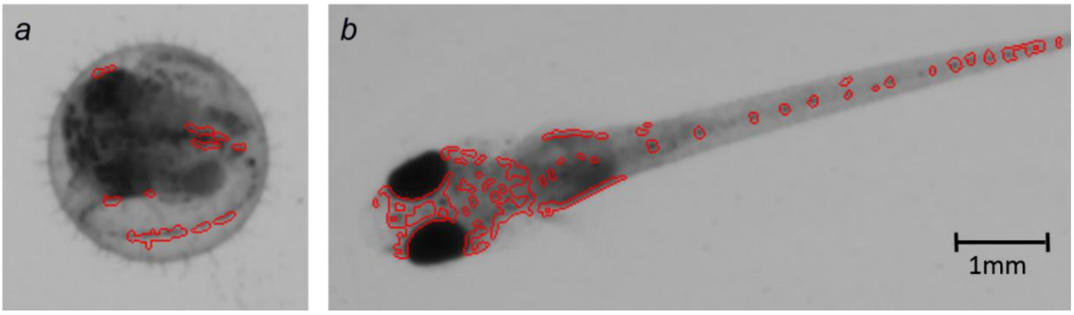


Figure 25. Incorrect segmentations due to fluttering.

2.4.4. Discussion

An image analysis pipeline for detecting a beating heart on 1-second-long videos of fish embryos has been presented and tested on a total set of 3,192 videos acquired over several experimental runs. This is a significant number of videos containing healthy as well as diseased embryos: some with edemas and other malformations. Thus, this set reflects production usage and allows us to validate the robustness of our protocol (Figure 26). Our results on 2,532 usable sequences show an acceptably low error rate, near 1.5% overall. This proves the efficacy and reliability of our image analysis method. However, when considering its integration within the entire system of embryo preparation, image acquisition, and processing, several points remain to be discussed, especially with respect to the validation phase and the establishment of ground truth.

Since we are discussing living organisms, establishing ground truth is not always easy. We rely on multiple visual observations of a subset of video sequences, which were not always consistent: expert observers did not always come to same conclusion. Indeed, we noticed 6 differences between them for 200 assessed videos, a rate of 3%. A second viewing of these videos was consequently performed with all observers present to achieve a consensus. With respect to this consensus, each observer had made between 3 and 5 errors, a rate between 1.5% and 2.5%. We note that our program had also made 3 errors. We conclude that the rate of subjectivity assessment is near 1.5%, which is considered acceptable.

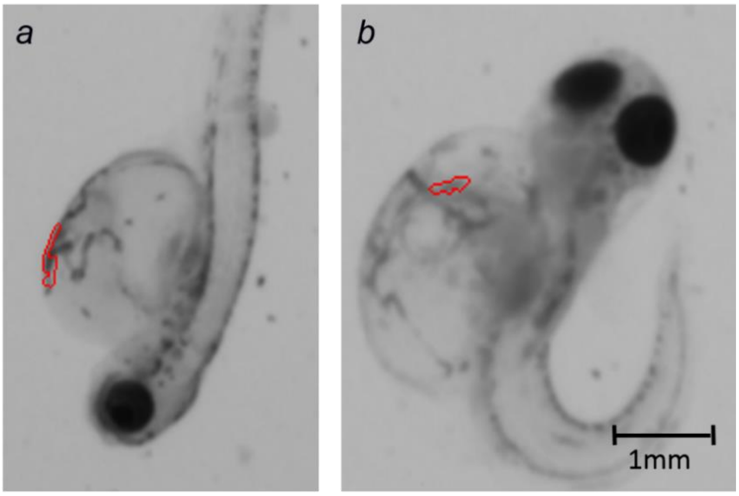


Figure 26. Heart segmentation in the presence of large edemas and axial malformations. The heartbeat is correctly detected.

As explained in Section 2.4.1.3, we still face video quality and accuracy issues with the current acquisition procedure. For example, when comparing the manual health status determination under microscope on Dataset 0 and the videos on Dataset 1, we noticed a discrepancy in 282 cases, for a rate of 11%. Consequently, the question of the overall accuracy of the automated method, including errors due to video acquisition and video treatment, raised. To assess this overall accuracy, a new validation was performed since the development of this mortality assessment test, on newly generated videos. The results were compared to both the reliable mortality assessment visually performed under a microscope, and to video-based observations. On 566 tested video sequences containing mixed eggs and alevins, a success rate of 92% is obtained compared to video-based observations, for a sensitivity of 94.4% and a specificity of 91.4%. By comparing to microscope-based observations, a success rate of 82% was obtained, with 92% of sensitivity and 79% of specificity. The difference rate between these two evaluations almost only concern embryos that were seen alive by looking under a microscope, but that appear dead on the corresponding video, as no beating heart is visible. Most of these cases are eggs. Thus, the program detects more dead embryos as it must do. As a conclusion, low error rates obtained basing on video-based ground truth are only representative of the quality of the program itself. To assess the reliability of the entire procedure, including preparation, acquisition, and treatment, we need to establish ground truth by observing embryos under a microscope.

Many sequences (20.5%) are correctly detected by the program as unusable. Some of them are due to an empty well, and so are not an issue, but the majority are due to embryos being too close to the well boundary. This represents an actual problem for the efficiency of the global procedure. To solve it, we investigate the solution of using wells that have a rounded, rather than flat, bottom. They are compatible with our Hamilton MICROLAB automated filling system and with our acquisition device, and would solve the problem of embryos that are too close to the edges of the well. However, the presence of a centered imprint on the well bottom remains a problem for image treatment as the imprint appears superimposed to the alevin on acquired images. We hope that with some experimental protocol adjustments, we will be able to use them in production. The description of such wells development is described in the Appendix of this manuscript.

After having worked on the mortality assessment of medaka embryos from video sequences in the previous part, we then want to study the malformations of embryos that are detected as alive. For this purpose, the embryo morphology can be analyzed based on images. As introduced in the previous section, both types of embryos are seen the day of acquisition: eggs and alevins, depending on hatching occurred or not. Before hatching, the embryo appears highly folded in its chorion, making the malformations difficult or impossible to detect, even when looking under a microscope. For this reason, the two following parts will focus on the analysis of alevins for the detection of specific malformations.

3. Automated classification of alevins with and without an axial malformation by machine learning

In this new section, an approach based on machine learning is developed to automatically classify alevins according to the presence of spine malformations. We built and validated our learning model on 1459 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, which is similar to the percentage of success of a trained human operator working on 2D images.

The work presented in this section has appeared in the following publication:

- D. Genest, E. Puybareau, M. Léonard, J. Cousty, N. De Crozé, H. Talbot. "High-throughput automated detection of axial malformations in Medaka embryo", in *Computer in Biology and Medicine*, pp 157-168, 2018.

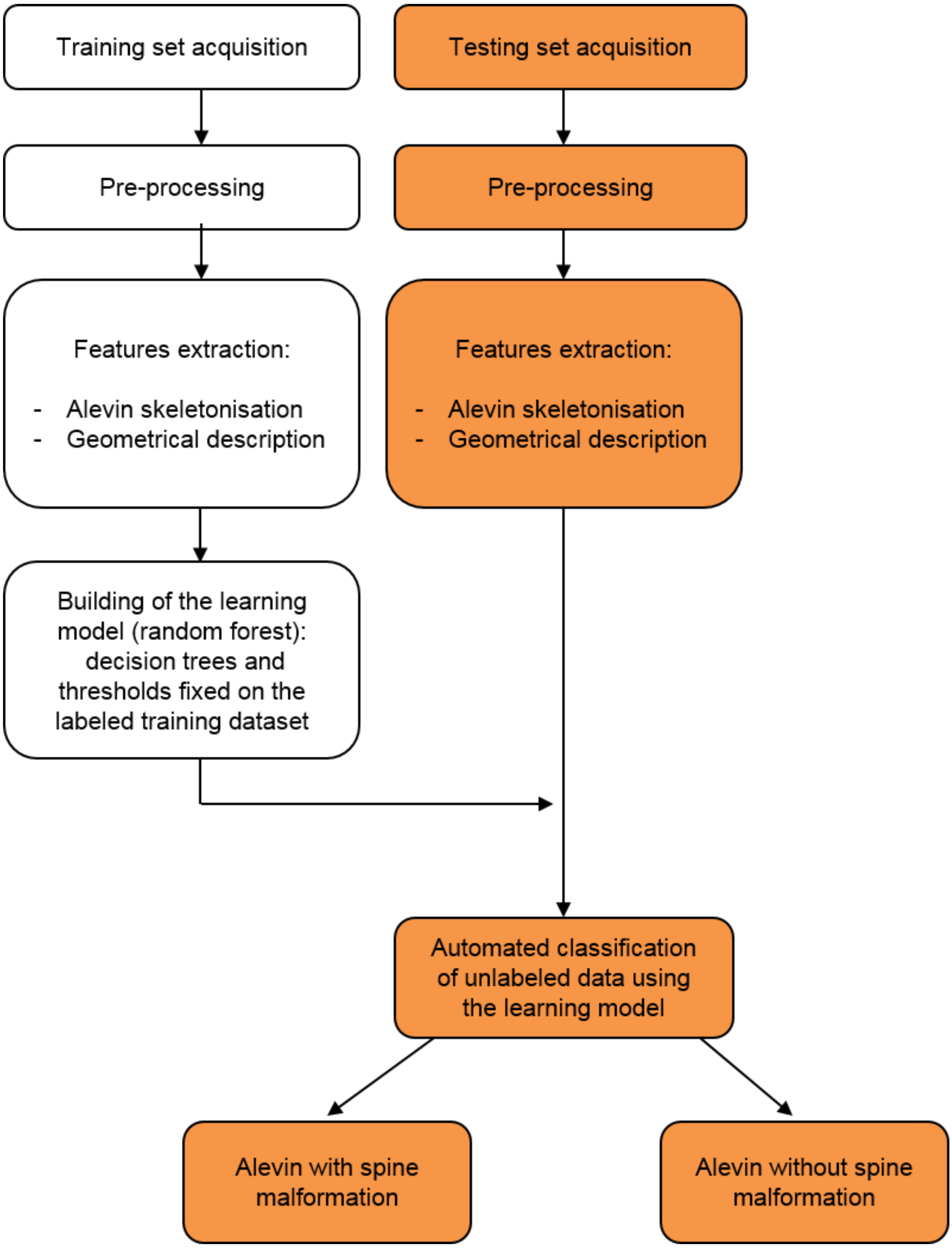

Training set acquisition
Testing set acquisition
Pre-processing
Pre-processing
Features extraction:
- Alevin skeletonisation
- Geometrical description
Features extraction:
- Alevin skeletonisation
- Geometrical description
Building of the learning model (random forest): decision trees and thresholds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

Figure 27. Flowchart of the alevin morphological abnormalities detection assay based on image processing. This detection method is assessed by cross-validation in the presented study.

3.1. Introduction to axial malformation detection

The objective of this part is to propose an automated method for classification of alevins with or without an *axial malformation* (abnormalities on the antero-posterior axis, including spine malformations), one of the most common developmental abnormalities observed in toxicological assays [74, 87]. This classification is based on the analysis of 2D images acquired according to the protocol described in Section 2.4.1.1 and Section 6.1.4 of the Appendix.

In the acquired images, the alevins can appear in any orientation from the lateral view to the dorsal view (Figure 28a to c). Moreover, axial malformations cover an important variety of phenotypes, from the most obvious malformation to slightest defects of the spine curvatures (Figure 28d and e). Indeed, if most of these malformations are characterized by abnormal spine curvature, some alevins also exhibit shortened tails or humps. Some specific cases of strongly bent alevins are referred to as hook-shaped (Figure 28f). This huge variety in alevins phenotypes and the single orientation acquired in 2D images make axial malformation complicated to characterize on 2D images. The first difficulty is thus to identify relevant parameters in order to characterize such a panel of malformations. We show in this part that mathematical morphology operators can provide an accurate description based on binary spine modelling in order to extract numerical values relative to axial malformation characterization. To this end, we consider an approach based on the morphological skeleton [61, 101]. Features such as size, curvature, angles are then deduced from this skeleton and gathered in a features vector in order to feed a random forest classifier [109]. The second difficulty is linked to the information loss when working with 2D images compared to 3D interactive observations made under a microscope. Here, the term "3D interactive" refers to the possibility of manipulating alevins, thus to see it on all positions, and zoom in on it to detect anomalies with a high precision, which is not possible in the single view shown in a 2D image. To validate the proposed set up, we challenge ground truth reliability by quantifying the gap between 3D interactive observations made under a microscope and observations made on 2D images. In addition, in order to quantify human subjectivity, we provide an estimation of the inter-operator subjectivity rate according to image-based observations made by three different observers.

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 labeled data. It begins with a pre-processing step that reduces the acquired data to the

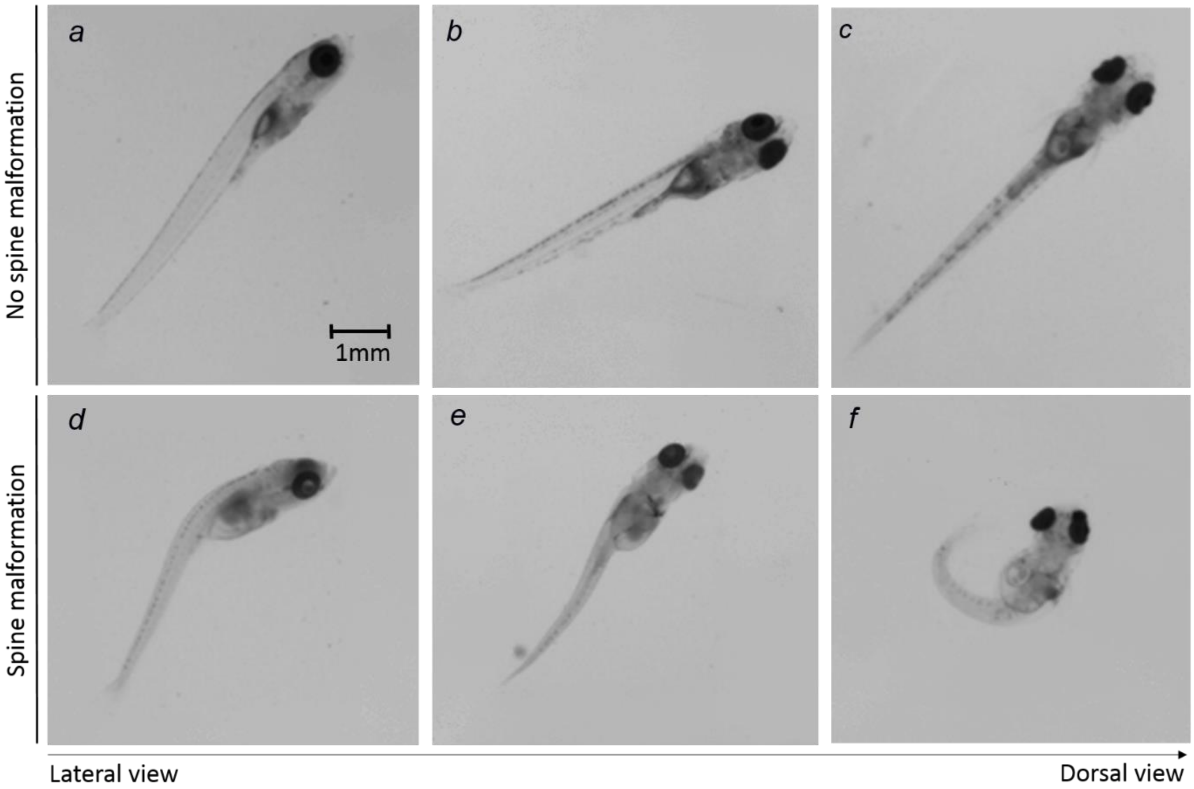


Figure 28. Images of 9dpf medaka alevins as acquired by our set-up. a to c: healthy alevins shown in lateral view in a, three-quarters view in b, and dorsal view in c. d to f: alevins showing different types of spine malformations, d being a major spine malformation (lateral view), e: slight "S-shaped" malformation (three quarters view) and of a hook-shaped alevin (dorsal view).

region of interest M^1 (the process is described in detail from Section 2.2.2 to 2.2.4). In the feature extraction step, the alevin spine is segmented using mathematical morphology operators [54]. 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 labeled data. During the testing phase, features are also extracted from the testing dataset and images are classified according to the trained random forest model. The flowchart of our methodology is summarized in Figure 27.

The method is described in 3.2, including spine segmentation and characterization with geometrical features. The classification method with a random forest model is then assessed in 3.3, including the description of the experimental setup, the classification results, and discussion.

3.2. Feature extraction for alevins 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 [52, 54]. Section 3.2.1 proposes a procedure to approximate the alevin's spine. Then, feature characterization is presented in Section 3.2.2.

3.2.1. Alevin's spine segmentation

In this section, we start from the cropped image and the first segmentation of the whole alevin contour $\mathcal{M}$, both obtained at the end of the pre-processing step presented from Section 2.2.2 to 2.2.4 (Figure 30a). Our aim is then to obtain, from the alevin's mask $\mathcal{M}$, a segmentation which approximates the curve of the alevin's spine. After smoothing the contour of the alevin, this methodology implements morphological skeletonisation (Section 1.2.3.1). More precisely, the spine approximation method uses the curvilinear skeleton principle described in [101] and [61]. An overview of the spine segmentation from the alevin mask $\mathcal{M}$ is given in Figure 29.

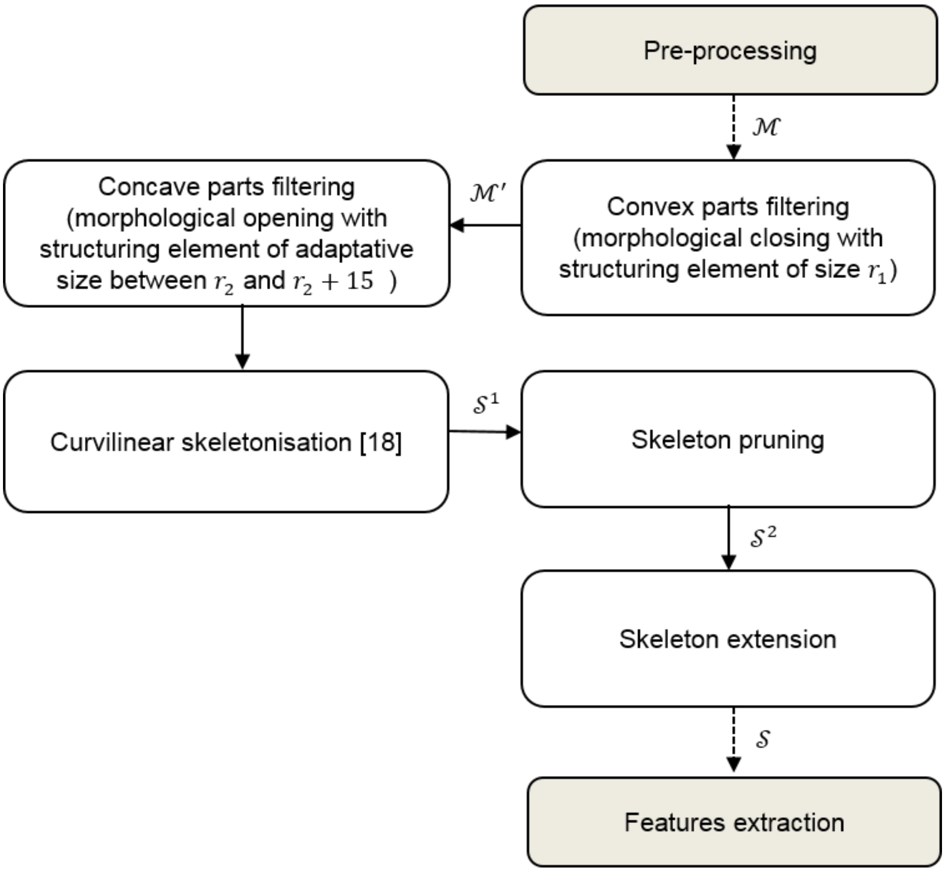

Pre-processing
$\mathcal{M}$
Convex parts filtering (morphological closing with structuring element of size r_1)
$\mathcal{M}'$
Concave parts filtering (morphological opening with structuring element of adaptative size between r_2 and $r_2 + 15$)
Curvilinear skeletonisation [18]
$\mathcal{S}^1$
Skeleton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

Figure 29. Flowchart of alevin's spine approximation.

Firstly, in order to reduce any artefact ramification in the further skeleton, 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 Γ_{r_1} of size r_1 [54]. In the following, we denote by $\mathcal{M}'$, the result of this process applied to $\mathcal{M}$:

$$\mathcal{M}' = \varphi_{\Gamma_{r_1}}(\mathcal{M}). \tag{3.1}$$

On the other hand, concave areas due to alevin abnormalities such as significant edemas 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 morphological opening of X by a disk-shaped structuring element of radius i.

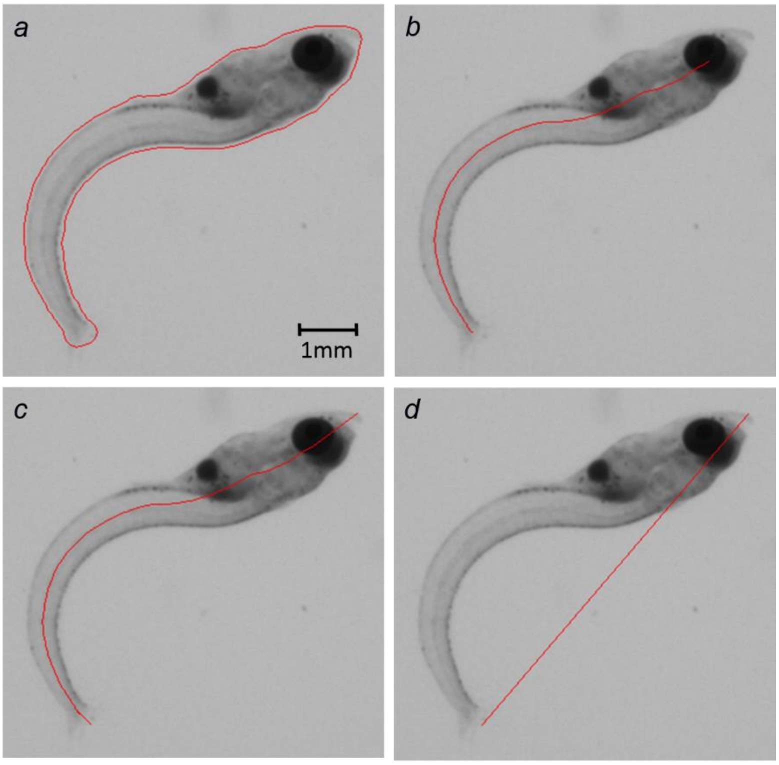


Figure 30. Spine approximation steps on the cropped image of an alevin. The red line represents the contour of the initial mask $\mathcal{M}$ in a, the initial curvilinear skeleton $\mathcal{S}^2$ in b, the extended curvilinear skeleton $\mathcal{S}$ in c and the straight line $\mathcal{L}$ linking both ends.

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:

$$\mathcal{S}^1 = S_{r_2+3.\min(5,\lambda)}(\mathcal{M}'), \tag{3.2}$$

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 α pixels. We write:

$$\mathcal{S}^2 = pruning_\alpha(\mathcal{S}^1), \tag{3.3}$$

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

From its definition, the curvilinear skeleton $\mathcal{S}^2$ (Figure 30b) does not reach the borders of the alevin shape $\mathcal{M}$ (Figure 30a). 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^iall the way to the border of $\mathcal{M}$. The resulting skeleton is denoted by $\mathcal{S}$ in the following (Figure 30c). 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 (Section 3.2.2), and the observed deviation does not highly impact the features measurement further described. Finally, both skeleton extremities are then linked via a line segment $\mathcal{L}$ (Figure 30d). Because a healthy alevin is expected to present a straight spine when it is anesthetized, this segment is used in the following section as a reference to compare the actual alevin's spine to a healthy spine.

3.2.2. Alevin's spine geometrical description

Classifying alevin's 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 3.2.1, 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 size, (ii) the curvature, (iii) the regularity and (iv) the discontinuities of the alevin shape.

3.2.2.1. Size measurement on the alevin masks

A first set of parameters, namely a_{alevin}, l_{alevin}, w_{max}, w_{mean}, r^1_{image} and r^2_{image} described below are related to the size of the alevin. The alevin area a_{alevin} is measured on mask $\mathcal{M}$ in number of pixels. The parameter l_{alevin} refers to the alevin 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 $\mathcal{M}$ [110, 111] 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

[1] This weighted skeleton is called the extinction function [54]

multiplied by two to obtain the maximal and average widths denoted by w_{max} and w_{mean}, respectively. We compute the ratios r_{image}^{1} and r_{image}^{2} between the alevin length and width as follow:

$$r_{\text{image}}^{1} = \frac{w_{\text{mean}}}{l_{\text{alevin}}}\text{; and } r_{\text{image}}^{2} = \frac{w_{\text{max}}}{l_{\text{alevin}}}. \tag{3.4}$$

3.2.2.2. Curvature assessment from the graphical representation of the alevin's spine

The aim of this section 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_{mean}, r_{graph}^{1}, r_{graph}^{2}, and r_{graph}^{3}. We build an image representation of the alevin's spine in order to simplify its analysis in a direct 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 $\vec{T}$ 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 31b.

Depending on the curve shape, it is not always possible to represent the detected alevin's spine as an explicit function. In particular, when multiple points of the curve, representing the alevin's spine in the presenting orthonormal frame, have the same abscissa, the spine is considered to have a hook. This case is described in Section 3.1. and Figure 28f. 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 ith 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.

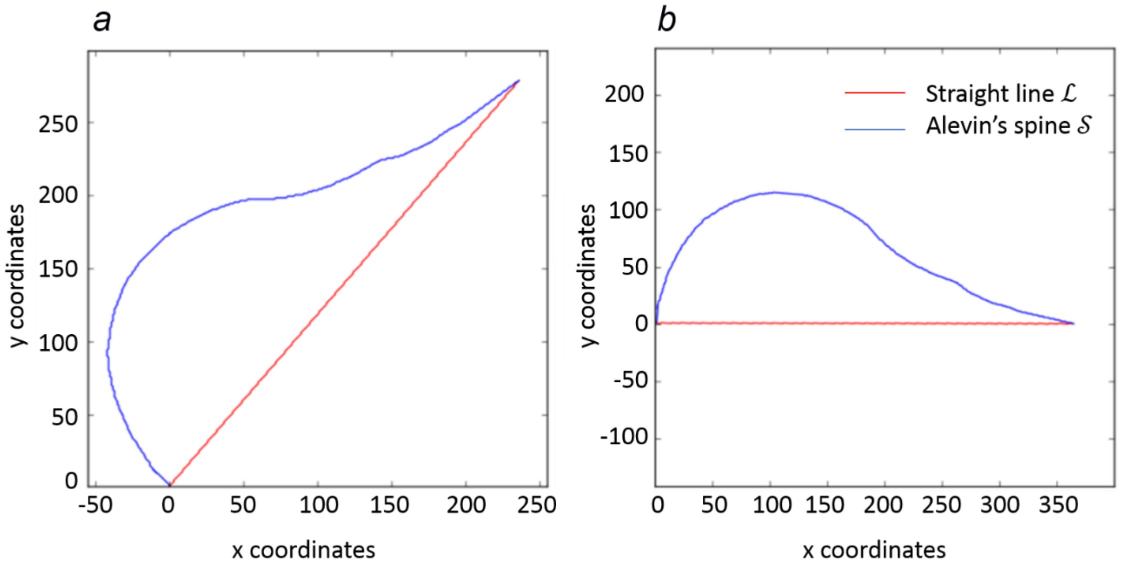


Figure 31. Graphical representation of the curvilinear skeleton $\mathcal{S}$ in a direct orthonormal frame. a: spine curve represented after translation $\vec{T}$. b: spine curve represented after translation $\vec{T}$ and rotation R.

The area under the curve (AUC) of the function $|f|$ is computed using the trapezoidal rule [112], where $|f|$ is the absolute value of $f(x)$ for every points x of the domain:

$$AUC = \sum_{i=1}^{n} \frac{(|f(x_{i-1})|+|f(x_i)|)}{2} \times (x_i - x_{i-1}). \tag{3.5}$$

The use of the absolute value allows analyzing 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:

$$d_{\max} = \max(f(x_i))\ for\ i\ \in [0, n]\ ;\ \text{and} \tag{3.6}$$

$$d_{\text{mean}} = \frac{1}{n} \sum_{i=0}^{n} f(x_i). \tag{3.7}$$

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

$$r_{\text{graph}}^1 = \frac{d_{\max}}{l_{\text{alevin}}}\ ;\ r_{\text{graph}}^2 = \frac{d_{\max}}{d_{\text{mean}}}\ ;\ \text{and}\ r_{\text{graph}}^3 = \frac{\text{AUC}}{l_{\text{alevin}}}. \tag{3.8}$$

3.2.2.3. *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 anesthetized 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 (defined in previous Section) by a parabola. Hence, we define the parabolic function f_p defined by:

$$f_p(x) = a_1x^2 + b_1x + c_1, \tag{3.9}$$

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:

$$r_p^2 = 1 - \frac{\sum_{i=0}^{n}(f(x_i)-f_p(x_i))^2}{\sum_{i=0}^{n}(f(x_i) - m)^2}, \tag{3.10}$$

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:

$$r_c^2 = 1 - \frac{\sum_{i=0}^{n}(f(x_i)-f_c(x_i))^2}{\sum_{i=0}^{n}(f(x_i) - m)^2}. \tag{3.11}$$

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

3.2.2.4. *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's 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 σ. 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 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 centered point. The broken line that links the selected extrema is finally considered. An example of this process is presented in Figure 32. The number of angles n_{angles} detected on the broken line created, the minimal angle θ_{min}, and the maximal angle θ_{max} are saved as features.

We summarize the parameters characterizing the alevin's spine and used during classification in Table 3.

Alevin's size descriptors	a_{alevin} ; l_{alevin} ; w_{mean} ; w_{max} ; r^1_{image} , r^2_{image}
Curvature descriptors	AUC ; d_{max}, d_{mean} ; r^1_{graph}, r^2_{graph}, r^3_{graph}
Curve regularity descriptors	r_p^2 ; r_c^2
Curve break descriptors	n_{angles} ; θ_{min} ; θ_{max}

Table 3. List of features extracted from alevin segmentations and used during axial classification.

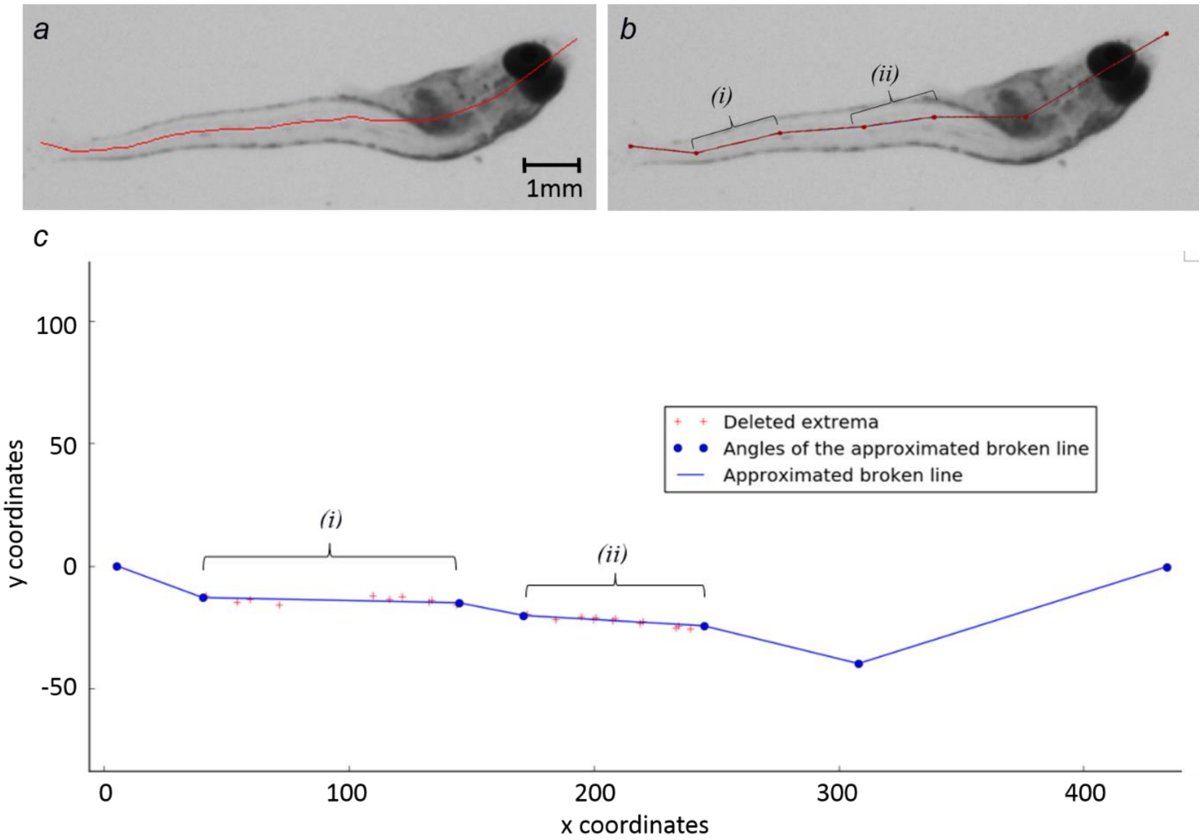


Figure 32. Alevin's spine approximation by a piecewise affine function. The red line shows the spine segmentation $\mathcal{S}$ in a, the approximated spine in b, superimposed on the cropped image. The approximated spine is represented in a direct orthonormal frame in c. In b and c: the areas (i) and (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 $\theta_{max} = 172°$.

3.3. Assessment of the learning classification of alevins with and without a spine malformation

The axial malformations detection method is assessed in this section. Section 3.3.1 presents the experimental set-up. Results are then presented in Section 3.3.2. before discussing them in Section 3.3.3.

3.3.1. Experimental set-up

The experimental set-up includes the dataset and ground truth establishment, the relevant tested methods and the performance measures.

3.3.1.1. Experimental protocol

The experimental protocol is the same as the one described in Section 2.4.1.1 and Section 6.1 of the Appendix. For each well, we record one photograph at a resolution of 1500×1500 pixels.

3.3.1.2. Software and libraries

We use the same Python 2.7 environment as described in Section 2.4.1.2. We used Numpy, Scipy and Pink libraries [106] for segmentation and features extraction, and Scikit-learn [113] for machine learning-based classification.

3.3.1.3. Database description

As seen in Section 3.2.2.2, feature characterization of our abnormality detection test depends on the alevin 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

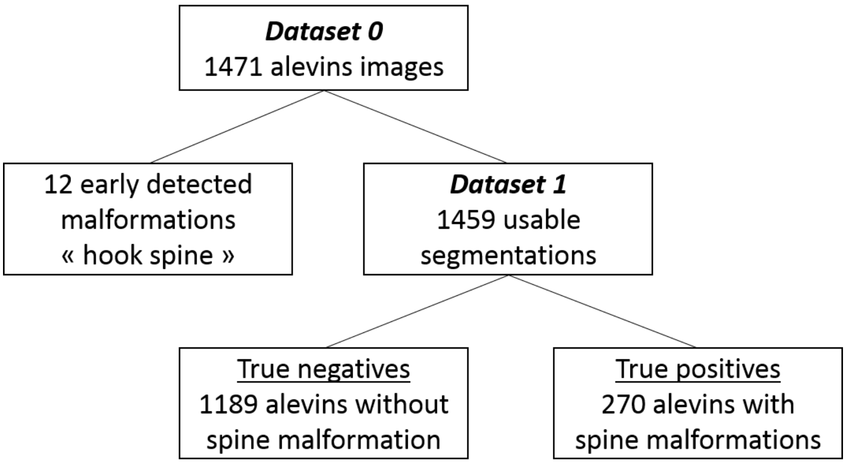


Figure 33. Datasets establishment for the assessment of axial malformation detection.

a hook (Figure 28). To deal with these cases, alevins identified as such are directly labeled 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 33.

3.3.1.4. 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 malformation. 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 a 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 33, the dataset of 1,459 images contains 270 images of alevins with a spine malformation and 1,189 images of alevins without.

3.3.1.5. *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 3.3.1.1 and 3.3.1.4. More precisely, we describe the setting of parameters presented in Section 3.2.1 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 malformations, 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 the Section 3.3.2.

The automated classifier parameters are set up as follow. All parameters described in Section 3.2.2 are experimentally determined in order to optimize segmentation results. Segmentation and geometric parameters are listed in Table 4 To set up the classifier parameters as described Section 1.2.2, 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 1.2.2. In our program, we use the implemented algorithm GridSearch from the scikit-learn library [113].

By testing different values for the weights w_{L_-} and w_{L_+} (see Equation 1.8) 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

Parameter name	Description	Value
r_1	Radius of Γ_{r_1} , the disk structuring element of the morphological closing $\varphi_{\Gamma_{r_1}}$ (Equation 3.1)	10
r_2	Minimal opening radius used for skeletonisation $S_{r_2+3.\min(5,\lambda)}$ (Equation 3.2)	14
α	Minimal branch length used for skeleton pruning (Equation 3.3)	25
σ	Size of the convolution scaled window used for skeleton curve smoothing (Section 3.2.2.4.)	11
d_1	Minimal vertical distance that must separate two successive extrema to maintain them during spine approximation by a piecewise affine function (Section 3.2.2.4.)	4
d_2	Minimal horizontal distance required by a steady portion to be considered during spine approximation by a piecewise affine function (Section 3.2.2.4.)	10

Table 4. Parameters determination for alevin’s spine segmentation and geometrical description of classification features.

depends on both its number of data $n_{L\_}$ and the weight of each data $w_{L\_}$ as it is described in Equation 1.8. According to the ground truth described in Section 3.3.1.4, 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 1.8.

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 labeled data. Both are used as input for the training algorithm and the model is fitted as explained in Section 1.2.2.

3.3.1.6. 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 [114]. 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

Results: Ground truth:	No axial malformation	Axial malformation
No axial malformation	TN	FP
Axial malformation	FN	TP

Table 5. 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.

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 3.3.1.1 and 3.3.1.4.

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 the following section in the form of confusion matrices. A confusion matrix [115] 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 5 for standard representation of a confusion matrix.

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

$$\text{specificity} = \text{true negative rate} = 100 \times (\frac{TN}{TN+FP}), \tag{3.12}$$

$$\text{sensitivity} = \text{true positive rate} = 100 \times (\frac{TP}{TP+FN}), \tag{3.13}$$

$$\text{FPR} = \text{false positive rate} = 100 \times (\frac{FP}{TN+FP}), \tag{3.14}$$

$$\text{FNR} = \text{false negative rate} = 100 \times (\frac{FN}{TP+FN}). \tag{3.15}$$

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 1.2.2: $\text{accuracy percentage}(y, \hat{y}) = \text{accuracy}(y, \hat{y}) \times 100$. 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.

3.3.2. Classification results

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.

3.3.2.1. 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 6a,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 6d.

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.

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 false positives of AC results are illustrated in Figure 34.

a

Classifiers results:	AC		HC	
Ground truth:	No axial malformation	Axial malformation	No axial malformation	Axial malformation
No axial malformation	1142	47	1163	26
Axial malformation	161	109	142	128

b

AC Results HC results	No axial malformation	Axial malformation
No axial malformation	1240	65
Axial malformation	63	91

c

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

Table 6. 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.

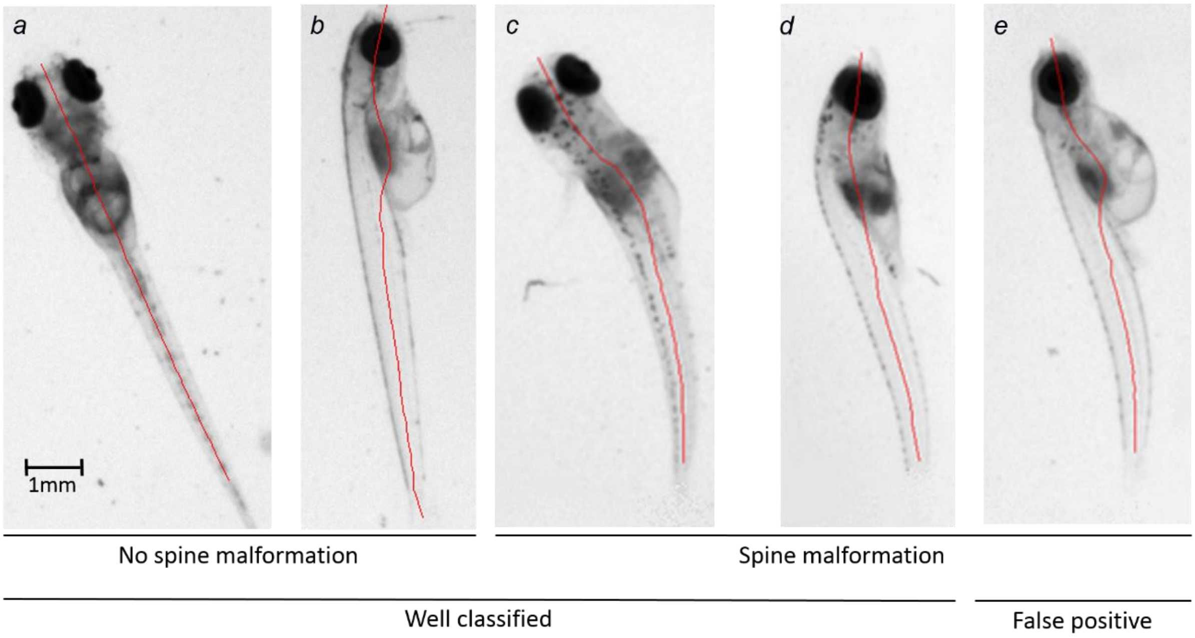


Figure 34. Results of alevin's spine classification. The red line represents the result of the spine segmentation $\boldsymbol{S}$. The method leads to proper classification (a and b: no spine malformation; c and d: spine malformation) or to a false positive (e: false detection of a spine malformation). a and c are presented in dorsal view while b, d and e are presented in lateral views.

3.3.2.2. 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 σ_X of 2.6. Such a low variability is acceptable.

For scaling up, close attention is paid to analyzing 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

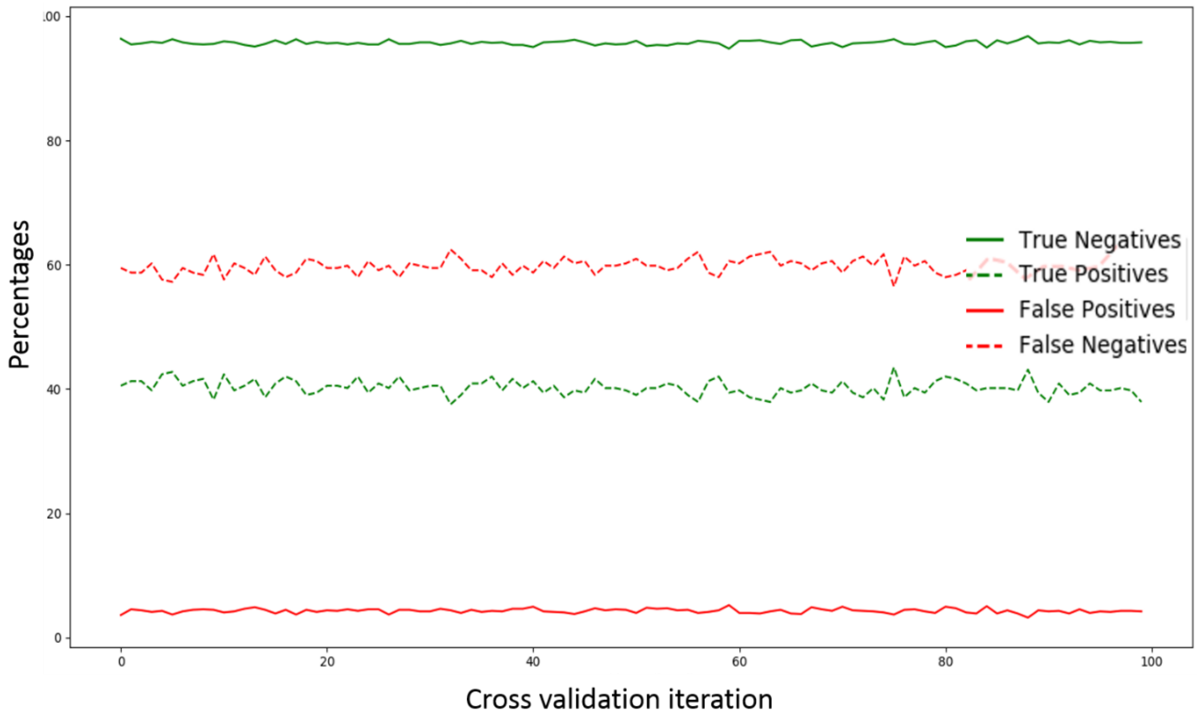


Figure 35. Evolution of the program results over 100 successive 10-fold cross-validations. At each new cross-validation, we calculate the rates of true negatives (the specificity), true positives (sensitivity), false negative and false positive on the whole database of 1459 images. Since we favor specificity, the false positive rate is minimized.

positive, false positive and false negative ratios are calculated according to Equations 3.12 to 3.15. As shown in Figure 35, 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.

3.3.2.3. Quality control of early data sorting

As previously explained in Section 3.3.1.1, 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%).

3.3.2.4. Inter-operator subjectivity

This last section of our study concerns inter-operator 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-operator 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 analyzed later. For this reason, our inter-operator 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 7b 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 a

Results:	Operator 2		Operator 3		Program	
	No axial malfor-mation	Axial malfor-mation	No axial malfor-mation	Axial malfor-mation	No axial malfor-mation	Axial malfor-mation
Ground truth:						
No axial malformation	154	23	167	10	169	7
Axial malformation	0	23	1	22	9	15

b

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

Table 7. Results and error rates obtained for each operator 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 operators 2 and 3 compared to the microscope-based ground truths. We report in b the percentage error calculated for each operator and for the automated classifier on this 200 data sample.

5.5% and 11.5%. We also note that the results distribution in Table 7a shows that errors made by the program are more balanced between false positives and false negatives. These results can be considered to be acceptable.

3.3.2.5. *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. Features calculation takes about a few seconds for each image (up to 5 seconds, including pre-processing). The classifier training step can be repeated as much as necessary on the calculated features in about one second. Our program then classifies an image in only about 1 second.

3.3.3. Discussion

This work aimed to develop an automated image processing-based assay for the detection of spine malformations in medaka alevins. As for every study presented in this manuscript, the emphasis was put on the overall accuracy of the test. As shown in Section 3.3.2.1 , 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, edemas 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.

As introduced in Section 1.3.1, several methodologies have been published in the context of alevin spinal cord analysis using image processing. Most were conducted on zebrafish embryos. In [74], the authors assessed the development of specific neuron population by extracting a quantitative information from fluorescent proteins labeled spinal cord neurons in transgenic zebrafish. An automatic system for the detection of abnormal curvature zebrafish tail is described in [87]. However, the study is limited to the classification of obvious abnormalities in tail curvature (up or down). A method is proposed in [41] 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. 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.

In this study, 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.

4. Automated classification of alevins with and without a swim bladder based on atlas and machine learning classification

The objective of this section is to present the method developed in order to automatically classify images of medaka alevins according to the presence or the absence of a swim bladder. The main challenge consists of developing a method which is accurate, regardless to the alevin orientation. The methodology relies on an adaptive features extraction step with a 2D atlas of a healthy alevin, and machine learning-based classification. An average precision rate of 95% is obtained in the total dataset of 380 images following 5-fold cross-validation.

The work described in this section has appeared in the following publication:

- (submitted) D. Genest, M. Léonard, J. Cousty, N. De Crozé, H. Talbot. “Atlas-based automated detection of swim bladder in Medaka embryo”, in *International Symposium on Mathematical Morphology*, Saarbrücken, Germany, 2019 (oral presentation).

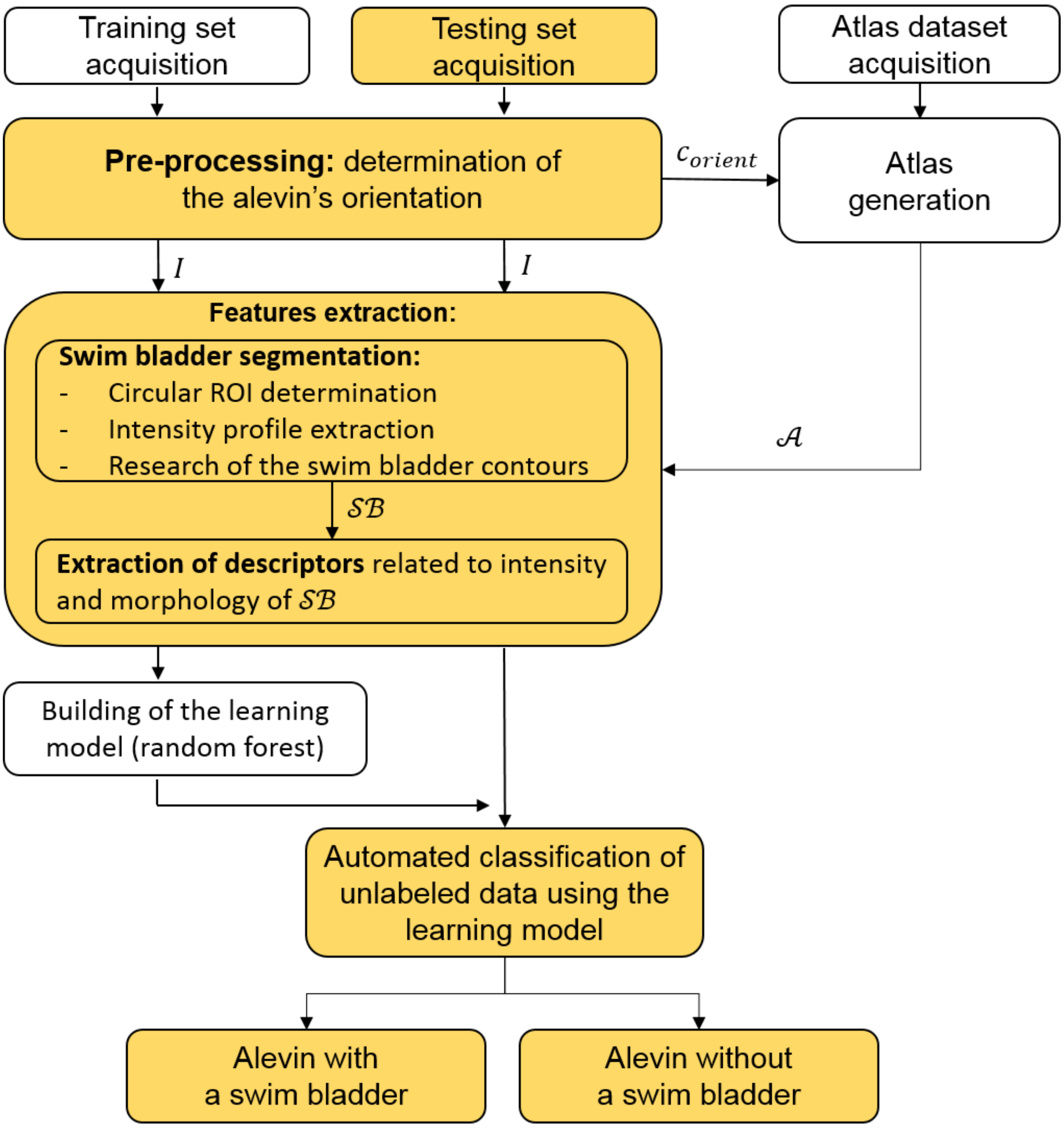

Training set acquisition
Testing set acquisition
Atlas dataset acquisition
Pre-processing: determination of the alevin's orientation
c_{orient}
Atlas generation
I
I
Features extraction:
Swim bladder segmentation:
- Circular ROI determination
- Intensity profile extraction
- Research of the swim bladder contours
$\mathcal{A}$
$\mathcal{SB}$
Extraction of descriptors related to intensity and morphology of $\mathcal{SB}$
Building of the learning model (random forest)
Automated classification of unlabeled data using the learning model
Alevin with a swim bladder
Alevin without a swim bladder

Figure 36. Flowchart of the swim bladder detection assay.

4.1. Introduction to swim bladder detection

Among abnormalities that can be visible in medaka alevins at 9dpf (days post fertilization), the absence of swim bladder, an internal gas-filled organ that allows the embryo to control its buoyancy is one of the most sensitive marker of developmental toxicity [23, 116]. In particular, it is known that blood circulation is a key factor in normal development of the swim bladder. The absence of an inflated swim bladder could be a marker of a heart failure [117]. In this section, we focus on the automated detection of an inflated swim bladder on 2D images of medaka fish embryos at 9dpf. As uninflated swim bladder is simply not visible on images, we will further refer to our method as a swim bladder detection method. Thus, alevins will be classified into those with a detected swim bladder and those without.

The proper detection of the swim bladder depends on the orientation of fish embryos. This implies to manually place the fish embryo before the image acquisition which is tedious and time consuming. Here, a methodology based on morphological operators is proposed to automatically detect the swim bladder on 2D images of fish embryos regardless of the fish embryo position. The first challenge of this study consists in automatically identifying the orientation of the alevin in the tested image. Then, the second challenge consists in developing an adaptive swim bladder analysis method, according to the orientation previously identified. After a pre-processing step (Section 2.2), the methodology consists of (i) the automated determination of the embryo orientation, (ii) the generation of an atlas representative of a healthy embryo in the detected orientation, (iii) the swim bladder segmentation using this atlas, (iv) the descriptors calculation and (v) the embryos classification according to these descriptors, between alevins with and without a swim bladder.

As for the detection of axial malformations presented in the previous part of this manuscript, we work on images that were cropped after applying the pre-processing step described in Section 2.2.2 to 2.2.4, that includes well borders extraction, embryo localization in this previously delimited area, and differentiation between eggs and alevins. As the swim bladder is not visible on eggs, even when looking them under a microscope, the further treatment is only made on alevins. On the cropped image of the alevin, a new pre-processing step is applied that is composed of a compartmentation step, during which the alevin is divided into three parts representative of the three segments presented in Figure 21 in Section 2.3.1: the head, the trunk and the tail, and of an orientation identification step based on features extraction and linear regression. We then identify a region of interest (ROI) in which the swim bladder is searched. For this purpose, an atlas of a healthy embryo is built for each studied orientation and used to extract the center of the circular ROI [118, 119, 120]. Because a swim bladder is not always visible on studied images, the following step of the method consists of extracting what we call the most probable contour of a swim bladder, relying on a geodesic active contours algorithm applied on the previously identified ROI. As visible in Figure 37, the swim bladder is characterized by a high contrast between dark contours and light inner part. On the contrary, embryos without a swim bladder present a homogeneous body in the location where the swim bladder should be present. For this reason, the detection method relies on the extraction of the polar intensity profile of the circular ROI (the intensity profile of each radius is extracted and concatenated), its representation by a direct weighted graph, and the determination of a circular shortest path, meaning a path of minimum energy cost of the intensity profile [121, 122]. This methodology expects to segment the swim bladder if present. If not, the segmented shape corresponds to a random part of the embryo body. Descriptors are subsequently extracted from this segmentation in order to conclude if the segmented shape is a swim bladder or not. An automated random forest classifier is finally trained on these descriptors in order to classify embryos with respect to the presence or absence of a swim bladder. The successive steps of the method are represented in Figure 36.

Section 4.2 describes the pre-processing step including alevin compartmentation and orientation identification. Section 4.3 introduces the swim bladder localization step. This includes the atlas generation, the identification, using this atlas, of a region of interest (ROI) for the search of the swim bladder and the swim bladder segmentation, and the description of how the segmented shape is characterized with intensity and morphological descriptors extraction. Finally, Section 4.4 presents the process of embryos classification by a random forest classifier according to the presence or absence of a swim bladder [109].

4.2. Pre-processing

From images initially acquired (see Sections 1.3.2), the pre-processing described in Sections 2.2.2 to 2.2.4 is applied. This pre-processing allows to obtain the cropped image I of

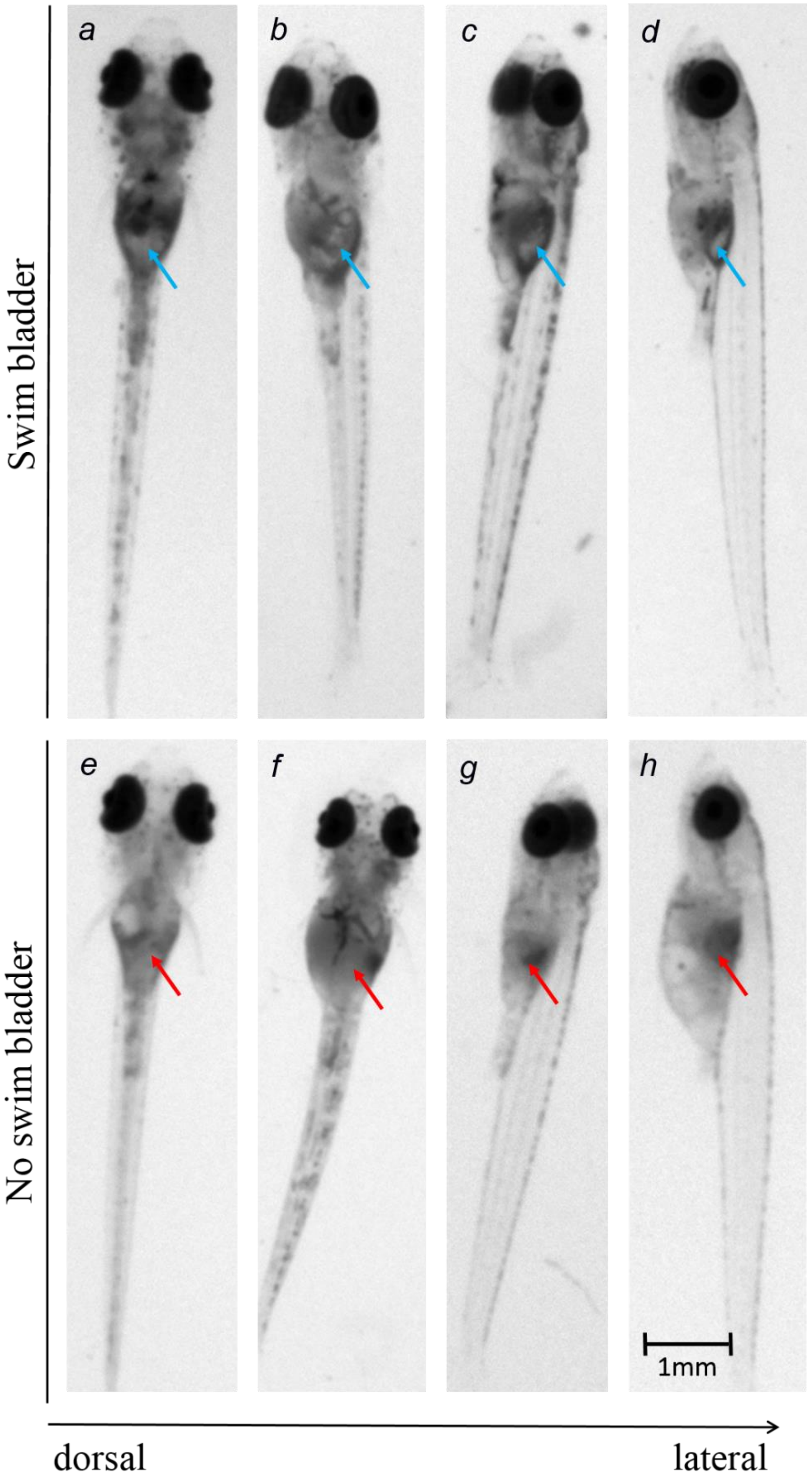


Figure 37. Medaka alevins with or without swim bladder and seen in different orientations from the dorsal view (left) to the lateral view (right). The blue arrow indicates the swim bladder location for embryos with a swim bladder in a to d, and the red arrow indicates the location where a swim bladder should be present for embryos without a swim bladder in e to h.

the studied alevin and the mask of the alevin contour $\mathcal{M}$ (Figure 38c), that will undergo the pre-processing treatment described in this section. The aim of this section is to automatically determine the orientation of the considered embryo, in order to adapt the method of the swim bladder detection. To do so, a compartmentation step is described in Section 4.2.1, that allows to divide the alevin into three different region of interest. These regions are then used in Section 4.2.2 to determine the alevin orientation, by extracting descriptors related to the eyes and to the tail in 4.2.2.1, and applying a linear regression on these descriptors in 4.2.2.2.

4.2.1. Alevin compartmentation

A first basic compartmentation was already used in Section 2.3.1.1 to extract the alevin's body (alevin's trunk and head). This process is here refined and completed in order to divide the binary mask of the alevin contour $\mathcal{M}$ into three segments: the head, the trunk and the tail (as introduced in Section 2.3.1). The process is described as the succession of three steps: markers extraction (in 4.2.1.1), markers superposition (in 4.2.1.2), and partition refinement (in 4.2.1.3).

4.2.1.1. Markers extraction

The aim of this subsection is to define markers that will be used for the segmentation of the three alevin's parts. We search for the marker of the alevin's head m^{head}, and the marker of the alevin's body m^{body}. For this purpose, we use an adaptative Otsu threshold described in Section 1.2.3 to segment the darkest areas of the alevin from the original image I [50]. These darkest areas correspond to the alevin's body. We only keep the largest connected component of the resulting mask and fill any of its holes. The result is dilated geodesically by a 10-radius disk in the limit of the alevin mask $\mathcal{M}$ (see [54] for a definition). We name the resulting mask m^{body}, illustrated in Figure 38d. In a similar way, alevin's eyes are segmented and dilated geodesically by a 20-radius disk, in the limit of the alevin mask $\mathcal{M}$, leading to the mask m^{head} (Figure 38e). We must underline that the radii of the structuring elements used to dilate the markers m^{body} and m^{head} are experimentally determined so that they are sufficient to separate wrongly labeled pixels from the well labeled connected components.

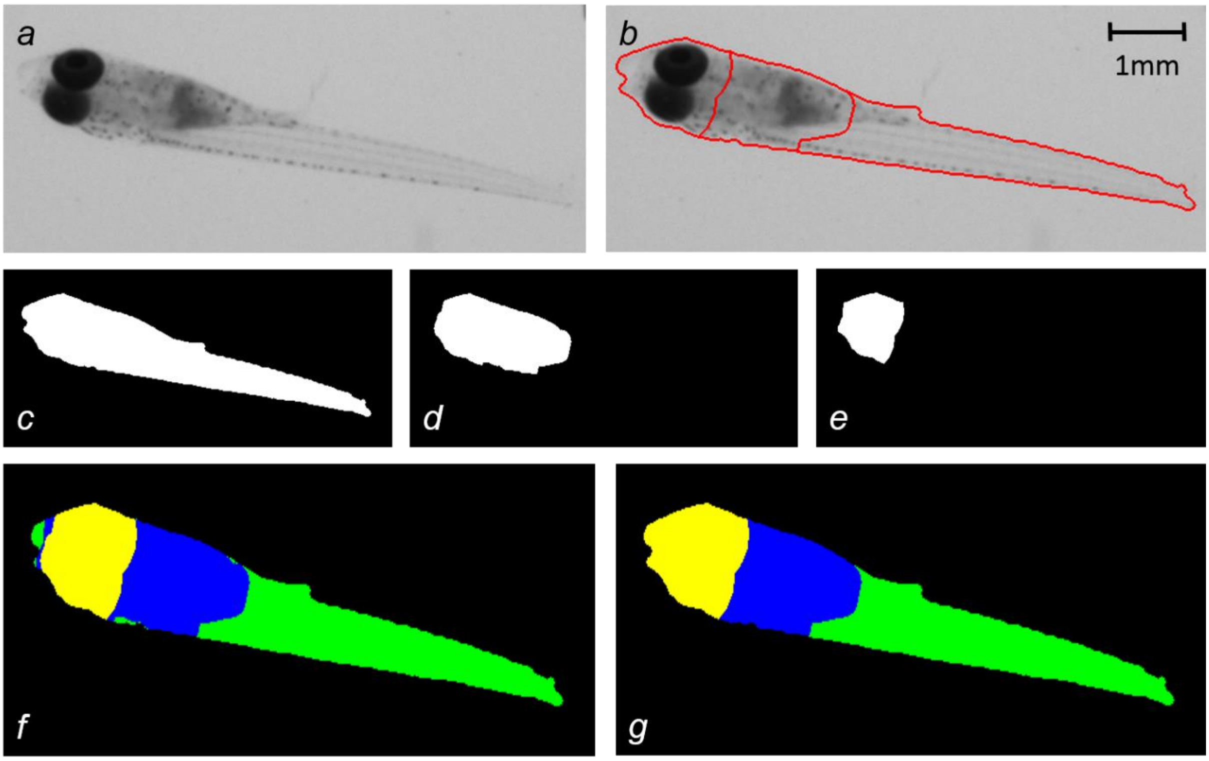


Figure 38. Alevin compartmentation. a: initial image I. b: final result of the compartmentation process $\mathcal{M}^{comp}$ appears in red, superimposed on the initial image I. c to g: compartmentation steps. c: whole alevin mask $\mathcal{M}$. d: body marker m^{body} . e: head marker m^{head}. f: result m^{sup} of the superposition and labeling of $\mathcal{M}$, m^{body} and m^{head}. g: final result after completion $\mathcal{M}^{comp}$, whose the yellow part corresponds to the head compartment $\mathcal{M}^{head}$, the blue part to the trunk compartment $\mathcal{M}^{trunk}$ and the green part to the tail compartment $\mathcal{M}^{tail}$.

The masks $\mathcal{M}$, m^{body} and m^{head} are then used as markers of the whole alevin, the alevin's body and the alevin's head respectively as described in the following section.

4.2.1.2. *Markers superposition*

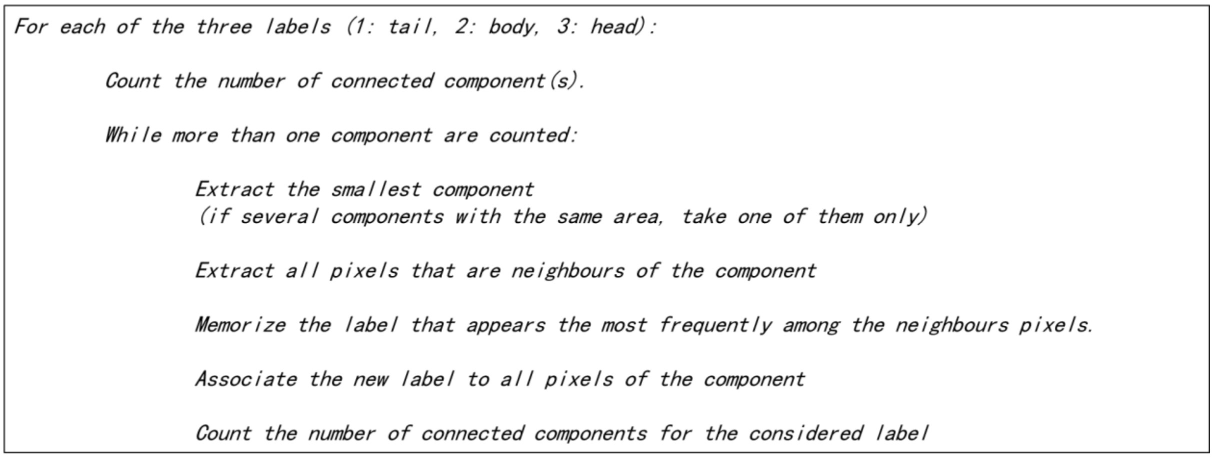
The masks $\mathcal{M}$, m^{body} and m^{head} are labeled 1, 2 and 3 respectively. The background of each image is labeled 0. The supremum, *i.e.* the pointwise maximum, of the three labeled images is calculated:

$$m^{sup} = \vee(\mathcal{M}, m^{body}, m^{head}) \tag{4.1}$$

The resulting labeled image m^{sup} corresponds to a first approximation of the three alevin's parts (head, trunk and tail). It is represented in Figure 38f.

4.2.1.3. *Partition refinement*

After markers superposition, each label does not correspond to a single connected component, as shown in Figure 38f. Some pixels in the areas of the alevin's head or trunk are still wrongly labeled. An algorithm is used to merge each wrongly labeled pixel to the connected component it actually belongs to. It is based on the assumption that wrongly labeled pixels always represent the smallest components with the considered label (Figure 39). At the end, only one connected component remains for each labeled component. We name $\mathcal{M}^{comp}$ the resulting labeled image which is composed of the head compartment $\mathcal{M}^{head}$, the trunk compartment $\mathcal{M}^{trunk}$ and the tail compartment $\mathcal{M}^{tail}$ (Figure 38b and g).



```
For each of the three labels (1: tail, 2: body, 3: head):
        Count the number of connected component(s).
        While more than one component are counted:
                Extract the smallest component
                (if several components with the same area, take one of them only)
                Extract all pixels that are neighbours of the component
                Memorize the label that appears the most frequently among the neighbours pixels.
                Associate the new label to all pixels of the component
                Count the number of connected components for the considered label
```

Figure 39. Partition refinement algorithm for automated alevin compartmentation.

4.2.2. Determination of the alevin orientation

The aim of this step is to automatically determine in which orientation the alevin appears in the studied image. Because, alevins can appear in every possible orientation between the most distant dorsal to the most distant lateral position, it corresponds to an infinity of orientations. For this reason, we aim to obtain an orientation related index. For this purpose, the orientation determination method relies on the extraction of descriptors related to the alevin orientation and linear regression.

4.2.2.1. *Extraction of orientation related descriptors*

When the orientation of an alevin changes, it is especially visible by looking on the eyes or on its tail. As shown in Figure 37, alevin's eyes are separated as the alevin is seen from an orientation close to the dorsal view whereas they overlap when it is seen in lateral view. Moreover, the tail pigmentation is not homogeneous. Alevins present a dark pigmented line along their dorsal line. This line is centered if the alevin is seen in dorsal orientation, whereas it is shifted to one side of the alevin when it is seen in lateral orientation. For these reasons, the method for alevins orientation determination is focused on the extraction of descriptors related to eyes morphology and tail intensity.

Eyes related descriptors

From the image I of the alevin, a median filter is firstly applied in order to remove the background noise. An adaptative Otsu threshold is then applied in order to perform the clustering-based image thresholding between dark intensities of the alevin's body, including eyes, and light intensities of the background and alevin's tail [50]. The obtained value θ is reduced from 100, up to the limit of 40. This calculation was experimentally determined in order to correctly distinguish alevins's eyes to the other parts of the alevin. We name the resulting constrained value θ_{co2}:

$$\theta_{co2} = maximum(\theta - 100, 40), \tag{4.2}$$

$$\mathcal{M}^1 = (I)_{\geq \theta_{co2}}. \tag{4.3}$$

A morphological closing $\varphi_{\Pi_{r_4}}$ is used with a polygonal structural element r_4 of radius 4 and we take the inverse of the resulting binary image:

$$\mathcal{M}^2 = inverse(\varphi_{\Pi_{r_4}}(\mathcal{M}^1)). \tag{4.4}$$

The resulting image $\mathcal{M}^2$ is expected to contain either one connected component if the alevin is seen close to lateral view and that one eye is occulted by the second, or two connected components if not. Thus, if more than two 8-connexity connected components are present in $\mathcal{M}^2$, we measure the difference between the areas of the two largest connected components. In case of a difference inferior to 500 pixels, experimentally determined, we consider both components are too different to correspond to properly segmented eyes, and we only keep the biggest one as the alevin's eyes. In case of a difference superior to 500 pixels, the two largest connected components are kept, each of them standing for an alevin's eye. The result is denoted $\mathcal{M}^3$ hereafter.

Two different descriptors are derived from $\mathcal{M}^3$: the size of the gap between the detected eyes gap and a binary orientation indicator i_{orient} which is equal to 1 if the eyes totally overlap and 0 else. We determined their values as follow. If two connected components are detected, the gap value is the shortest distance between the contours of the detected components. It is calculated by linking both connected components centroids by a line segment and by measuring the length of the subsegment bounded by the intersection points with the connected components contours, as shown in Figure 40. If only one connected component was detected, then the descriptor gap is set to 0. Concerning the orientation related indicator i_{orient}, we set it to 0 if two connected components have been detected on $\mathcal{M}^3$. However, if only one connected component was previously detected, it is analyzed in the aim to distinguish cases of alevins seen in lateral view from others seen in an intermediary orientation between dorsal and lateral. When seen in perfect lateral orientation, the visible connected component appears circular since the two circular eyes should theoretically overlap completely. In contrast, in an intermediary orientation, the connected component appears in the form of a "8-shape" component (*i.e.* with two lobes). For this reason, we expand the contour of the detected eye in $\mathcal{M}^3$ by taking its convex hull, meaning the smallest convex set containing this component [100]. If $\mathcal{M}^3$ is convex, then i_{orient} is set to 1, meaning that the alevin is considered in a lateral orientation, whereas if $\mathcal{M}^3$ is not convex, then i_{orient} is set to 0, meaning that the alevin is considered in a non-lateral orientation. These descriptors will then be combined to other descriptors related to alevin's tail to characterize the alevin orientation.

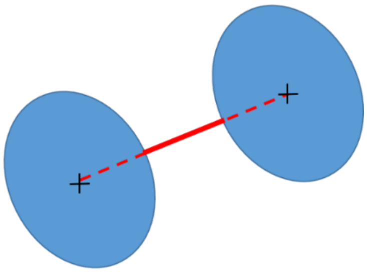


Figure 40. Measurement of the gap between alevin's eyes. Both eyes are represented in blue, with black crosses indicating their centroids. The red line shows the distance between both centroids. The gap between eyes corresponds to the distance shown by the whole part of the red line.

Tail intensity related descriptors

The pigmentation in the alevin's tail varies from the dorsal line to the belly line (along the dorsoventral axis). When seen in a dorsal orientation, the dorsoventral axis is not visible, and the pigmentation appears centered on the left-right axis (see alevin axis in Figure 2 in the introduction). Thus, our aim is to characterize the distribution of the pixel intensity along the axis that transversally crosses the alevin's tail. The methodology relies on the representation of the region of interest that delimits the pixels of the alevin's tail $\mathcal{M}^{tail}$ (obtained in Section 4.2.1), in a dual frame. This dual representation is denoted $\mathcal{M}_d^{tail}$ and results from the concatenation of all cross sections extracted from the tail along the anteroposterior axis of the alevin. Features will then be extracted from this dual representation.

We firstly want to extract the mask $\mathcal{M}^{side}$ that will be used to orient each tail cross section. For this purpose, the mask of the alevin contour $\mathcal{M}^4$, which is the contour of $\mathcal{M}$ (Section 2.2.4), and the alevin skeleton $\mathcal{S}$ (see Section 3.2.1) are used. The skeleton $\mathcal{S}$ is considered representative of the anteroposterior axis of the alevin. It crosses the mask $\mathcal{M}^4$ in two points corresponding to the two extremities of the skeleton. Thus, the skeleton $\mathcal{S}$ is used to cluster and label the pixels of $\mathcal{M}^4$ into two different components corresponding to both side of the anteroposterior axis. Among the two connected components of $\mathcal{M}^4$, only the one with the highest area $\mathcal{M}^{side}$ is retained. Such binary image will allow to obtain directed cross sections of the alevin's tail.

We now want to build the image $\mathcal{M}_d^{tail}$ which is the dual representation of the tail image $\mathcal{M}^{tail}$. Considering the region of interest $\mathcal{M}^{tail}$, we define its associated representation in the dual frame $\mathcal{M}_d^{tail}$ as the concatenation of all cross sections, *i.e.*, organized and oriented lines that are perpendicular to the skeleton $\mathcal{S}$ and limited to the contour of the tail $\mathcal{M}^{tail}$. Each cross section s_x of the image $\mathcal{M}^{tail}$, defined by a distance x in pixels from one of the intersection points between $\mathcal{M}^{tail}$ and $\mathcal{S}$, is then precisely the x^{th} column of the dual image representation $\mathcal{M}_d^{tail}$. It is oriented by finding its extremity that belongs to $\mathcal{M}^{side}$. The intersection point between the skeleton $\mathcal{S}$ and the cross section s_x is denoted by O_x . Hence, to each pixel $p = (x_p, y_p)$ in $\mathcal{M}_d^{tail}$, we associate the intensity $c(p)$ of the pixel p' of the primal image $\mathcal{M}^{tail}$ such that p' is on the cross section s_{x_p}, and $\left\|\overrightarrow{pO_{x_p}}\right\| = y_p$ (Figure 41). During $\mathcal{M}_d^{tail}$ normalization, the length of each section s_x of $\mathcal{M}_d^{tail}$ is extended, taking the maximal

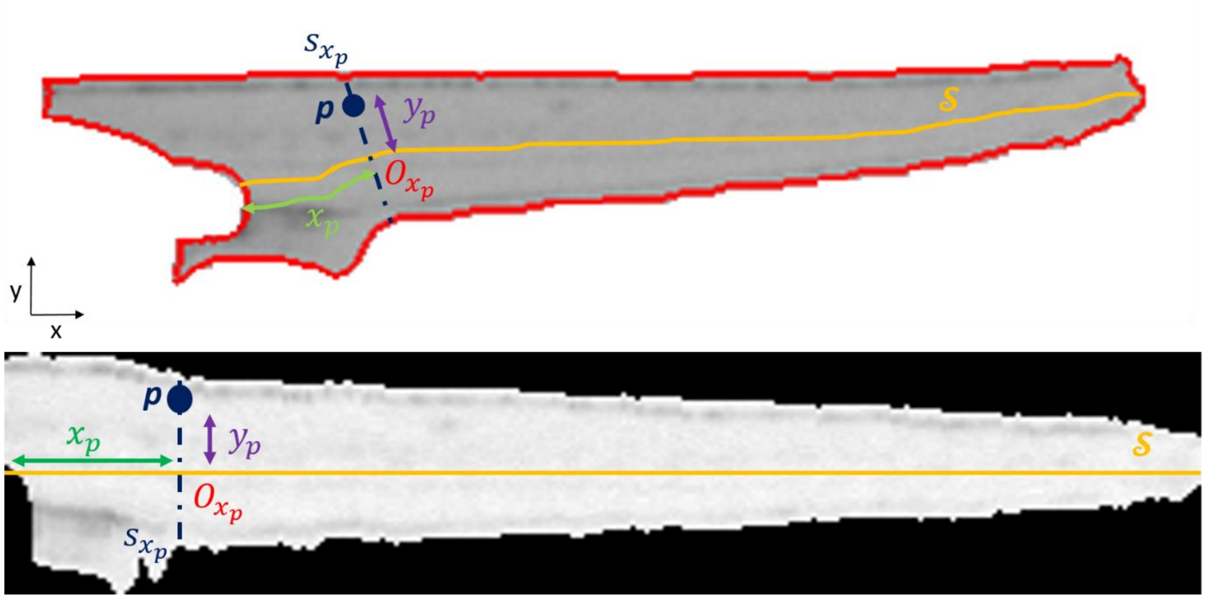


Figure 41. Representation of the region of interest of the alevin's tail in the primal frame $\mathcal{M}^{tail}$ and its associated dual representation $\mathcal{M}_d^{tail}$. The yellow line represents the skeleton $\mathcal{S}$.

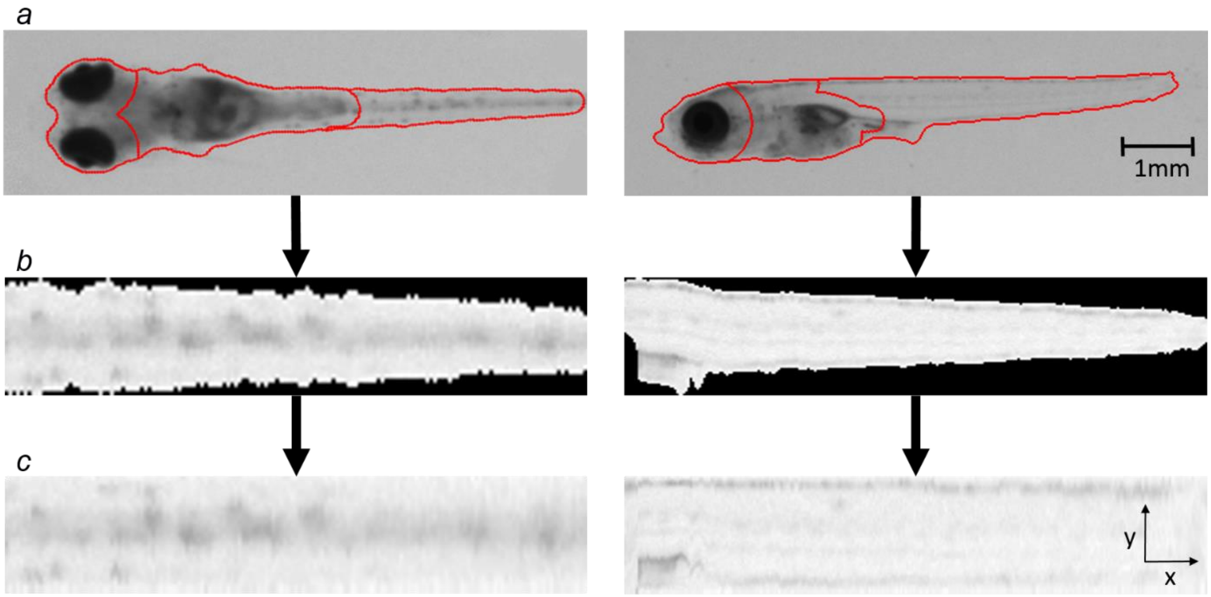


Figure 42. Generation of the dual representation of the alevin's tail $\mathcal{M}_d^{tail}$ for an alevin seen in dorsal view (left) and for an alevin seen in lateral view (right). a: result of the compartmentation $\mathcal{M}^{comp}$ is shown in red, superimposed on the initial cropped image I of the alevin. The right component of $\mathcal{M}^{comp}$ delimits the tail's contour $\mathcal{M}^{tail}$. b: dual representation of the tail $\mathcal{M}_d^{tail}$ obtained after concatenation of all the cross sections extracted along the alevin's tail, and before normalization. c: dual representation $\mathcal{M}_d^{tail}$ obtained after normalization.

column length as reference. The value of newly inserted pixel is calculated by linear interpolation. To each pixel $p = (x, y_p)$ of a section s_x in $\mathcal{M}_d^{tail}$, we associate the intensity $c_{norm}(p)$ such that:

$$c_{norm}(p) = \frac{y_b - y_p}{y_b - y_a} c(a) + \frac{y_p - y_a}{y_b - y_a} c(b), \tag{4.5}$$

with $a = (x, y_a)$ and $b = (x, y_b)$ are pixels of the section s_x in $\mathcal{M}_d^{tail}$ such that $y_a < y_p < y_b$ and $c(a)$ and $c(b)$ are known values. We call this process normalization of the image $\mathcal{M}_d^{tail}$. The process of $\mathcal{M}_d^{tail}$ generation and normalization is illustrated in Figure 42.

After $\mathcal{M}_d^{tail}$ normalization, the average intensity profile is computed along the x axis as shown in Figure 43, and descriptors are extracted. Concerning the tail morphology, the maximal width of the tail $width_{tail}$ is selected and corresponds to the number of rows in $\mathcal{M}_d^{tail}$. Concerning the intensity variation along the tail cross section, the averaged intensity profiles shown in Figure 43 reveal that the location of the global minimum tends to move to the tail edges as the alevin is visible from a more lateral view. Moreover, considering the global minimum peak as a Gaussian, we note that the standard deviation also reduces when the alevin is seen in lateral view. For these reasons, we extract the position, the mean and the standard deviation of the global minimum from the averaged intensity profile. Combined to $width_{tail}$, these measurements are used as features of the alevin's tail.

Combined with eyes-related descriptors previously presented, these features will allow to predict the orientation of a concerned alevin.

4.2.2.2. *Linear regression for alevins orientations classification*

In the previous section, features were extracted from the eyes and from the tail of the alevin, that allow to characterize the alevin orientation. From these features, we now want to predict the alevin orientation, by performing a linear regression.

Ground truth

Establish ground truth in order to reliably represent all orientation samples is a difficult task as it exists an infinity of possible orientations that the alevin can have from the extreme dorsal to

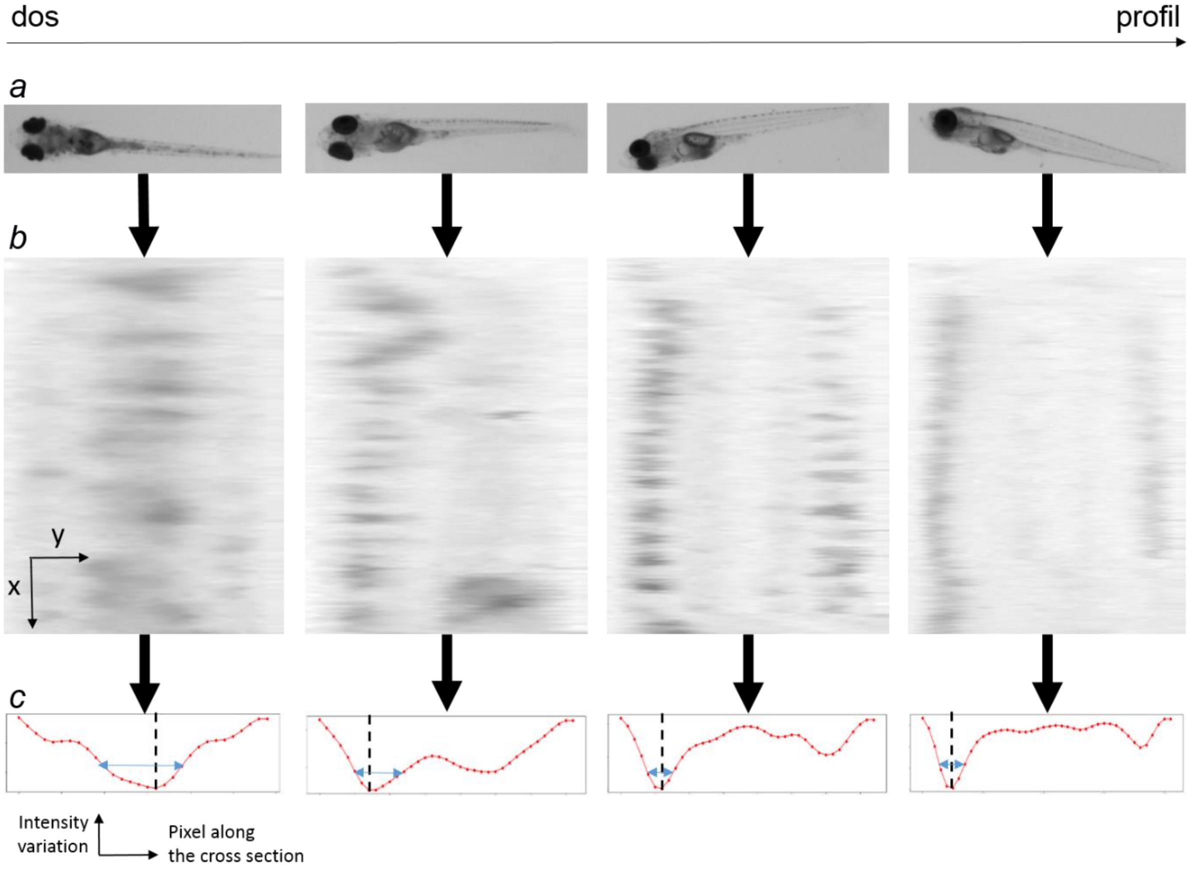


Figure 43. Extraction of alevin orientation related descriptors for alevins of the dorsal class O_D, of the almost dorsal class O_{AD}, of the almost lateral class O_{AL} and of the lateral class O_L presented from the left to the right. a: initial images I. b: dual representations of the tail $\mathcal{M}_d^{tail}$. c: graphical representations on a direct orthonormal frame of the averaged intensity variation along the tail cross section (averaged intensity profile) and descriptors extraction. The black dotted line indicates the location of the global minimum along the cross section and the blue double arrow shows its standard deviation.

the extreme lateral orientation. It is not feasible to manually associate a continuous value related to the alevin orientation on observed images. For this reason, we decide to manually classify images into four easily definable classes. A first class, named *dorsal orientation class* O_D, include all images where alevin's eyes are clearly separated, where the whole alevin appears symmetric, with a centered dark pigmented line on the alevin's tail. A second class refers to alevins seen in an intermediary orientation close to dorsal view. Alevins of this class appear with separated eyes but their tail does not appear symmetric, with a most peripheral pigmented line. We refer to this second class as the *almost dorsal orientation class* O_{AD}. A third class includes alevins seen in an intermediary orientation close to lateral view. It corresponds to alevins whose eyes appear superimposed and thus are not separated anymore. Such alevins are not symmetric. In particular, the tail appears wider and lighter than in previous class, with a peripheral pigmented line almost merged with the tail contour. This third class is called O_{AL} for *almost lateral orientation class*. The final *lateral orientation class* O_L gathers all alevins seen in lateral view. These alevins present a wide and light tail, with a pigmented line totally merged with the tail contour, and their eyes are totally superimposed such that only one circular eye is actually visible. An example of this manually generated distribution is represented in Figure 37 whose the first column present alevins from O_D, the second column present alevins from O_{AD}, the third show alevins from O_{AL} and the last column contains alevins from O_L.

Dataset description

A dataset of 293 images was constituted. According to the predefined ground truth, this includes 197 alevins that belong to O_D (where 151 are healthy and 46 are malformed), 16 alevins that belong to O_{AD} (12 healthy and 4 malformed), 67 that belong to O_{AL} (35 healthy and 32 malformed), and 13 that belong to O_L (7 healthy and 6 malformed). Malformed alevins included in this dataset present malformations such as spine malformations, absence of a swim bladder or presence of edemas. As the output of a regression is continuous instead of categorical, each data is associated to a real value that depends on the orientation class it belongs to (Section 1.2.1).Real values comprised in the interval [0,1] are associated to the 293 data depending on the class they belong to. Images of O_D are associated to a value of 0.25, images of O_{AD} are associated to a value of 0.5, images of O_{AL} are associated to a value of 0.75, and images of O_L are associated to 1.

Linear regression results

A linear regression is then performed on this labeled dataset using the previously described features.

The results of the linear regression applied on the dataset previously described are presented in Figure 44. In the further analysis, the predicted value is referred to as the orientation coefficient c_{orient}. In this figure, we observe that the values of c_{orient} are generally distributed into four staggered intervals. We observe that all data of O_D are associated with a value of c_{orient} less than 0.4. Data of O_{AD} are distributed between 0.2 and 0.6, all data of O_{AL} are distributed between 0.6 and 1 and all data of O_L are distributed between 0.8 and 1. Finally, these four intervals appear superimposed at the transition between two successive orientation classes only, making these errors less significant. Applying the following thresholds on the predicted value c_{orient}, an orientation-based classification can be made according to the linear regression results. We classify in predicted O_D the images with $c_{orient} \leq 0.4$. In the same way, images with $0.4 < c_{orient} \leq 0.6$ are classified into predicted O_{AD}, images with $0.6 < c_{orient} \leq 0.8$ are classified into predicted O_{AL}, and remaining images with $0.8 < c_{orient} \leq 1$ are classifies into predicted O_L. According to this classification, a second observation was made only on the wrongly classified data and revealed their ambiguity, as the sort criteria previously described do not allow to perfectly and reliably classify these data in a manual way into the four classes. In particular, some of them present axial malformation or torsion that prevent from determining a precise orientation. Actually, regarding these ambiguous cases, classification results remain consistent. Examples of such ambiguous data are represented in Figure 45.

In the following treatment, this linear regression model is retained as a model for alevins orientation determination. The model returns an index, called the orientation coefficient c_{orient}, included in the interval $[0,1]$. In further treatment, we decide to use this coefficient to classify alevins in three orientation-related classes defined as follow: the dorsal orientation class is defined by $0 \leq c_{orient} \leq 0.4$, the three quarters orientation class is defined by $0.4 < c_{orient} \leq 0.6$, and the lateral orientation class by $0.6 < c_{orient} \leq 1$. This classification will allow to adjust the further treatment applied on alevins to detect the swim bladder. Indeed, images of each class will undergo a specific swim bladder localization method, described in the following section.

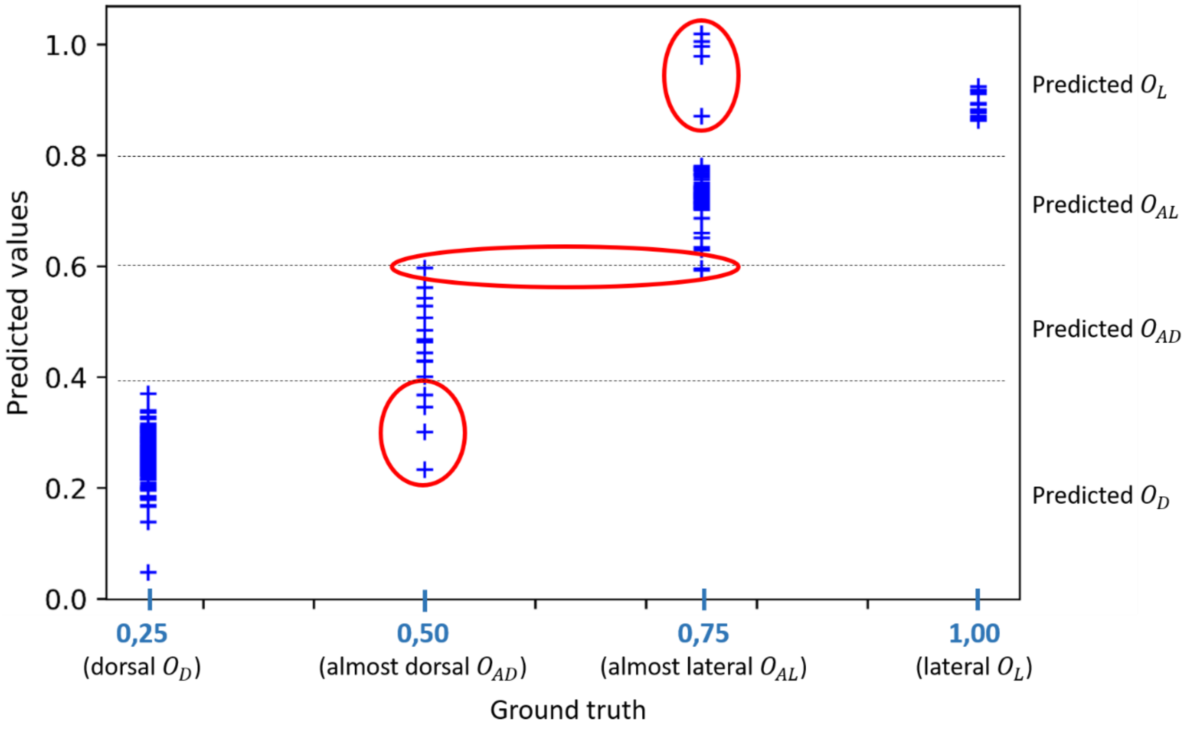


Figure 44. Results of the linear regression on alevins orientation. The four ground truth categories are labeled with real values from 0.25 to 1 to perform the linear regression. Four intervals are obtained for predicted values. The red circles point out the ambiguous data where these intervals are superimposed, that corresponds to wrongly classified data according to the ground truth. For ambiguous data, ground truth and associated predicted value always concern adjacent classes.

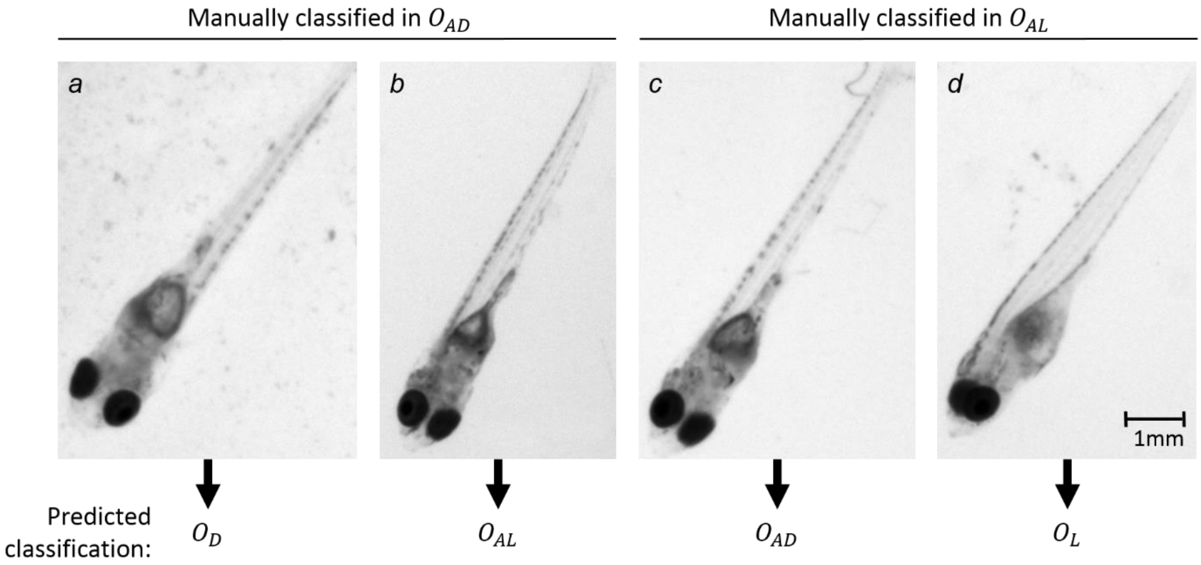

Manually classified in O_{AD}
Manually classified in O_{AL}
a
b
c
d
1mm
Predicted classification:
O_D
O_{AL}
O_{AD}
O_L

Figure 45. Results of the orientation classification on images of alevins presenting ambiguous orientations. a and b: alevins manually classified as almost dorsal and automatically classified as dorsal and almost lateral respectively. c and d: alevins manually classified as almost lateral and automatically classified as almost dorsal and lateral respectively.

4.3. Features extraction for swim bladder characterization

This Section describes how we perform the swim bladder detection and characterization. Because the swim bladder is not always present in studied images and the purpose is to distinguish cases of alevins with a swim bladder from cases of alevins without, the method relies on the extraction of the most probable contour of the swim bladder, and is adapted to the orientation coefficient previously calculated in Section 4.2.2.2. To this aim, we begin by identifying the region of interest (ROI) in which the swim bladder will be localized. To this end, we use an atlas of a healthy alevin. Then the most probable contour of the swim bladder is extracted before analyzing it in order to identify if this contour is the one of a swim bladder or not. For practical reasons, the swim bladder segmentation method is firstly described in this section with respect to the embryo orientation that appears the most frequently in our database, *i.e.* the dorsal view. The adjustments made to adapt the methodology to other the orientations coefficient is further described in Section 4.3.2.3.

4.3.1. Swim bladder atlas generation

The objective here is to build a median image I_{med} representative of a typical healthy embryo with respect to the diversity of embryos that exists in experimental conditions, and a probability function p_{sb} defined on the ensemble of I_{med} pixels coordinates and that represents the likelihood of each pixel of the represented image to belong to the swim bladder. In the following, such pair (I_{med}, p_{sb}) is called an atlas of the swim bladder for medaka embryo images and will be denoted by $\mathcal{A}$.

In order to build the atlas, n images of healthy embryos are selected and their swim bladder are manually segmented. Among these images, one is randomly chosen as being the fixed reference image I_F. Then, the $n-1$ remaining images, called the moving images hereafter, are aligned on I_F by applying an affine image registration algorithm. It consists of finding, for each moving image, an affine transformation that minimizes a similarity measure between the fixed image and the transformed moving image [123, 124, 125]. The mutual information is used as a similarity measure in this process. It aims at maximizing the measure of the mutual dependence

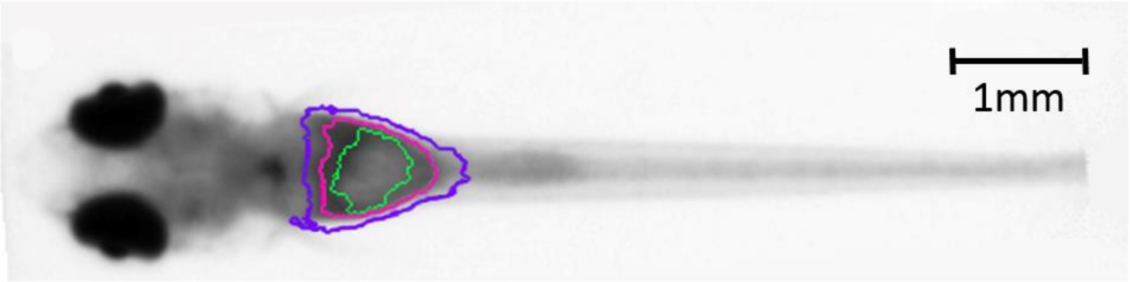


Figure 46. Atlas $\mathcal{A} = (I_{med}, p_{sb})$ obtained for fish embryos seen in dorsal view. The three red lines show the isocontours that delimit the areas where the pixels have a probability equal to 1, to 0.5 and to 0,05 to belong to the swim bladder.

between the pixel intensity distributions of the fixed and of the moving images, defined in [126] by:

$$MI\left(I_{F,}I_{M}\right) = \sum_{m} \sum_{f} p(f,m) \, log_2 \left(\frac{p(f,m)}{p_F(f)p_M(m)}\right) \tag{4.6}$$

where m and f are the intensities of the fixed and moving image respectively, p is the discrete joint probability, and p_F and p_M are the marginal discrete probabilities of the fixed image I_F and the moving image I_M [127]. The multiresolution affine registration algorithm from the Elastix toolbox is used to perform such process [128] . The median I_{med} of the n registered images is calculated. For each moving image I_M, the optimal transformation μ is also applied to the corresponding manual segmentation of the swim bladder. The average image I_{av} of the n registered swim bladder segmentations is calculated. We define the probability function p_{sb} as the mapping which maps to each pixel of coordinates (x, y) of a new image I with the same dimensions as I_{med}, the value $I_{av}(x, y)$ (Figure 46). The atlas $\mathcal{A} = (I_{med}, p_{sb})$ will then be used in order to identify, in a new image I, the ROI where to search the swim bladder.

4.3.2. Swim bladder localization

After registering the atlas on a new image I, we observe that the registered atlas does not segment the contour of the swim bladder with precision, as shown in Figure 47. As our methodology relies on the characterization of the swim bladder most probable contours to distinguish embryos with and without a swim bladder, we need to obtain a more accurate delineation of the swim bladder, when it is present.

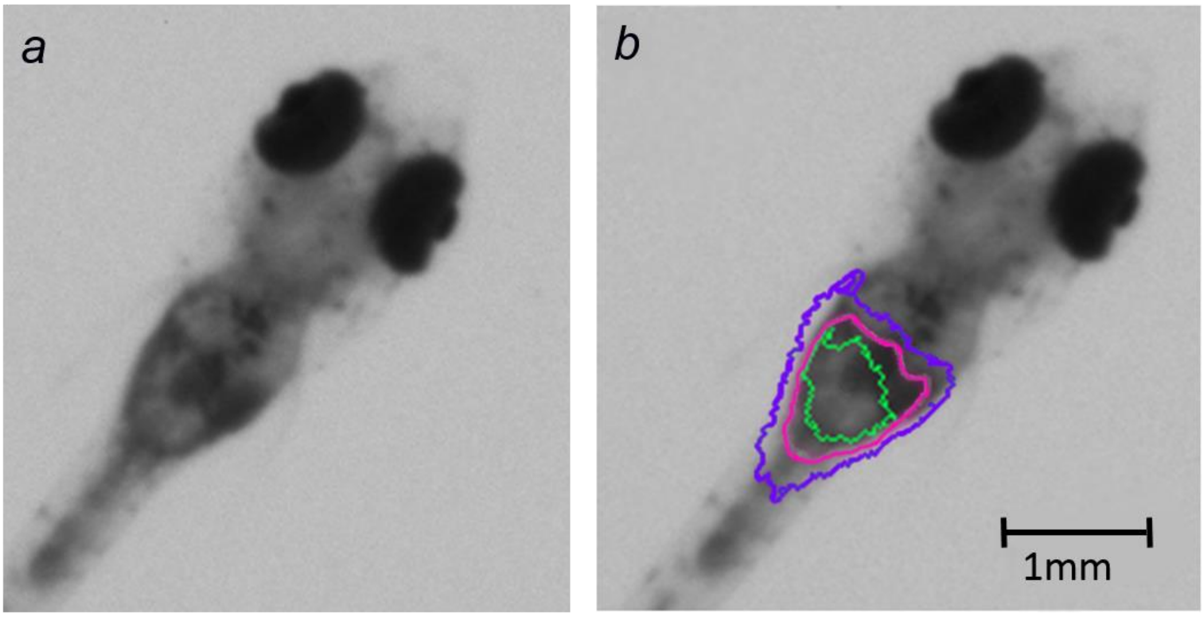


Figure 47. Projection of the atlas $\mathcal{A}$ on an embryo image. The three red lines are the isocontours that delimit the areas where pixels have a probability equal to 1, to 0.5 and to 0.05 to belong to the swim bladder according to the probability function p_{sb} of the atlas.

4.3.2.1. ROI localization

The atlas $\mathcal{A} = (I_{med}, p_{sb})$ is used on an embryo image I in order to identify the ROI in which the swim bladder will be searched. To this aim, we search for the transformation μ' that optimally registers I_{med} to the analyzed image I. We apply the same affine registration process as described in the previous section. This transformation μ' is then applied to the probability function p_{sb}, leading to a transformed probability function p'_{sb}. We consider the isocontour that delimits the area where the pixels have a probability equal to 1 to belong to the swim bladder, if a swim bladder is present (Figure 46). The barycenter of this area is extracted and considered as the center C of the ROI. The ROI $\mathcal{C}$ is then defined as the circle of center C and of diameter 40 pixels, experimentally determined (Figure 48).

4.3.2.2. Extraction of the swim bladder most probable contour

As visible in Figure 37, the swim bladder is characterized by a high contrast between dark contours and light inner part that cannot be too small. On the contrary, embryos without a swim bladder present a homogeneous body in the location where the swim bladder should be present.

For this reason, the swim bladder detection method relies on the determination of a circular shortest path extracted from the image $\mathcal{C}$ represented in a dual polar frame defined as follow. Considering $\mathcal{C}$ of center $C = (\mathrm{x}_C, \mathrm{y}_C)$ and of radius r in the primal frame of the image I, we define its associated representation in a dual polar frame $\mathcal{C}_d$, as the image that is the concatenation of all the ROI radial sections, starting from a radial section s_1 that is perpendicular to the embryo skeleton obtained during pre-processing.

Each radial section s_θ of $\mathcal{C}$, defined by an angle θ in degrees from the initial section, is then precisely the θ^{th} column of the dual image representation $\mathcal{C}_d$ (Figure 48). Hence, to each pixel $p = (\theta_p, r_p)$ in $\mathcal{C}_d$, we associate the intensity $c(p)$ of the pixel $p' = (r_p \cos\theta_p, r_p \sin\theta_p)$ in the primal image $\mathcal{C}$.

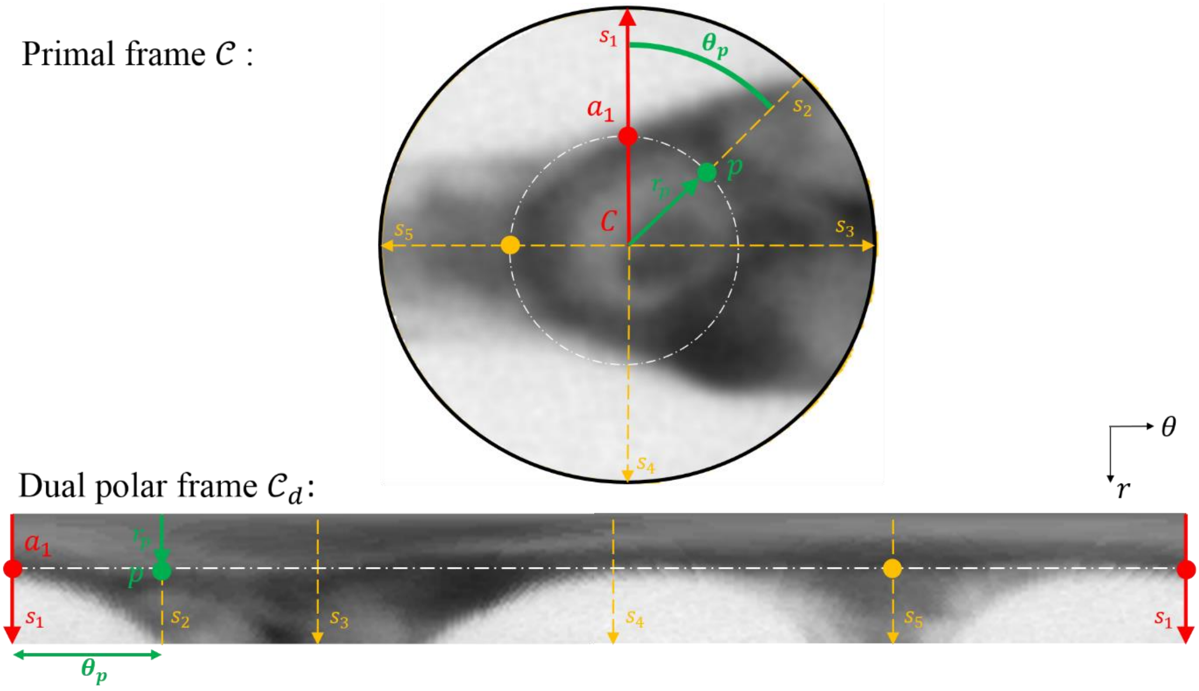


Figure 48. Representation of the ROI $\mathcal{C}$ in the primal frame and of its associated image $\mathcal{C}_d$ in the dual frame. The angle θ_p varies from 0 (section s_1) to 360, and all radial sections of $\mathcal{C}$ are concatenated to create the dual representation $\mathcal{C}_d$.

We then consider circular shortest paths in the image $\mathcal{C}_d$, *i.e.*, paths corresponding to contours of minimum energy in the primal image [121, 122]. For this purpose, the image $\mathcal{C}_d$ is equipped with a directed graph such that a pair (a, b) of two pixels of $\mathcal{C}_\mathrm{d}$ is a directed arc if $a_1 = b_1 - 1$ and $|a_2 - b_2| \leq 1$, where $a = (a_1, a_2)$ and $b = (b_1, b_2)$. In this graph, a *circular path* is a sequence $(p_0, \ldots, p_l)$ of pixels of $\mathcal{C}_\mathrm{d}$ such that:

- for any *i* in $\{1, \ldots, l\}$, the pair (p_{i-1}, p_i) is an arc;
- the first coordinate of p_0 is equal to 0;
- the first coordinate of p_l is equal to 360 (*i.e.* the maximal possible value); and
- the second coordinate of p_0 and of p_l are the same.

The *energy cost* $EC(\pi)$ of a circular path $\pi = (p_0, \ldots, p_l)$ is defined as the sum of the intensities of the pixels in the path: $EC(\pi) = \sum_{i \in \{0,\ldots,l\}} c(p_i)$. A circular path $\pi = (p_0, \ldots, p_l)$ is called *optimal* whenever the energy cost of π is less than or equal to the energy cost of any circular path from p_0 to p_l. Such circular optimal path can be found with any graph shortest path algorithm such as the one of Dijkstra [121].

In order to obtain the most probable contour of the swim bladder, we start by selecting the most peripheral local minimum of the first radial section r_1, called a_1. We also define $rmin$ as the minimal radius of $\mathcal{C}$ below which the shortest path must not be searched. It is experimentally set to 10 pixels. We then consider a circular shortest path π starting at a_1.This circular shortest path found in the image $\mathcal{C}_d$, corresponds to a closed contour $\mathcal{SB}$ in the image $\mathcal{C}$ which surrounds the centre C and which is of minimal energy. Such optimal contour $\mathcal{SB}$ is hereafter referred to as the most probable contour of the swim bladder (Figure 49). This methodology expects to segment the swim bladder if present, or a meaningless part of the embryo body otherwise.

4.3.2.3. Adaptation of the swim bladder localization method to the alevin orientation

Depending on the predetermined alevin orientation that depends on the previously described orientation coefficient c_{orient} (Section 4.2.2), the swim bladder localization parameters are adapted according to three different methods: the dorsal, the three quarters and the lateral methods.

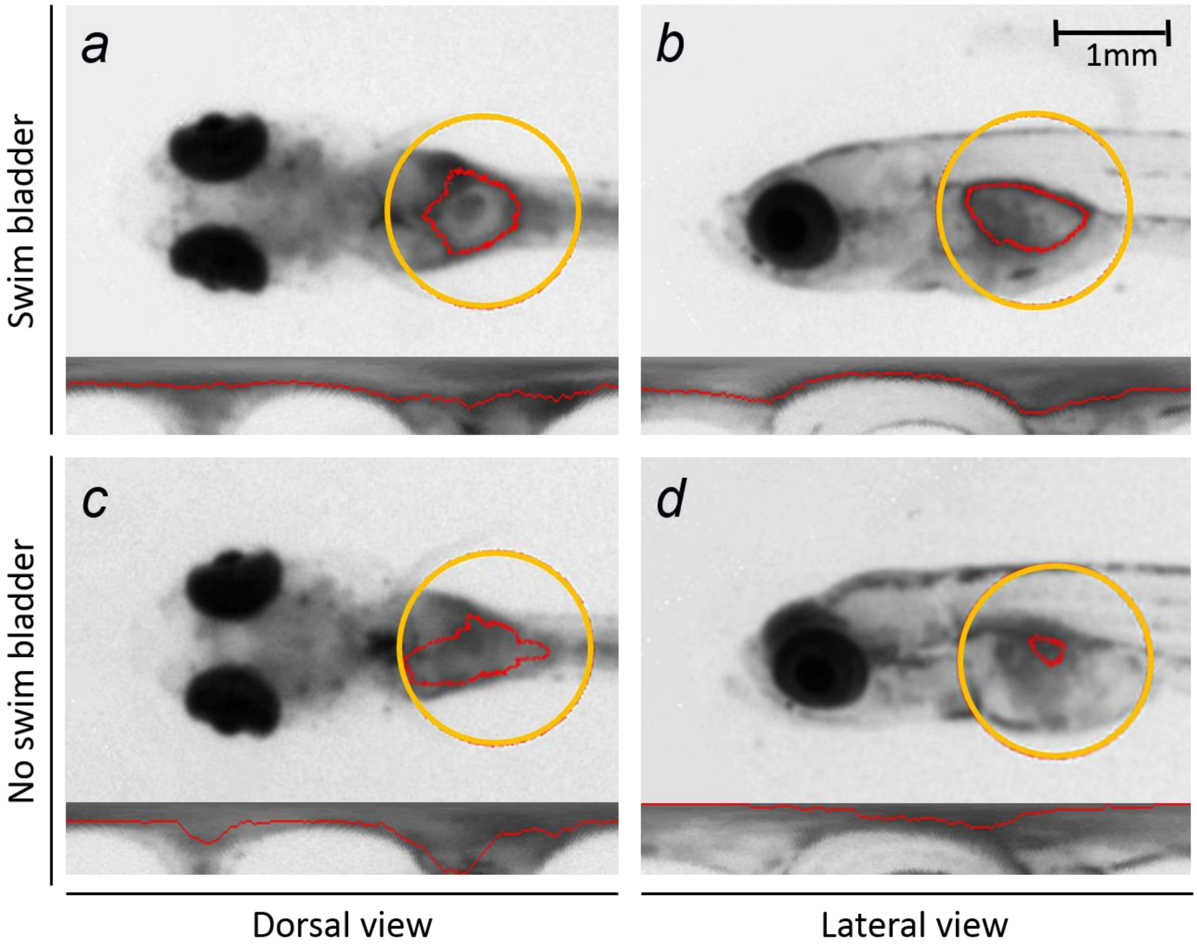


Figure 49. Swim bladder segmentation results on the primal frame of the image and associated shortest path in the dual polar frame. The yellow circle delimits the ROI $\mathcal{C}$ in which the red line shows the contour of the segmented shape $\mathcal{SB}$ in case of embryos with a swim bladder (a and b) and embryos without a swim bladder (c and d). a and c: embryos seen in dorsal view. b and d: embryos seen in lateral view.

Atlas generation

Following the atlas generation process described in Section 4.3.1, different atlases are built. The first one, denoted by $\mathcal{A}_D = (I_{med_D}, p_{sb_D})$, corresponds to the dorsal view of the alevin and is generated from images of the dorsal orientation class as defined in in Section 4.2.2. The second one, denoted by $\mathcal{A}_L = (I_{med_L}, p_{sb_L})$, corresponds to the lateral view of the alevin and is generated from images of the lateral orientation class. Concerning alevins images of the intermediary three quarters orientation class, the too important variability between individuals does not allow to generate a representative atlas. We solve this problem by creating an adaptive atlas $\mathcal{A}_{\propto} = (I_{med_{\propto}}, p_{sb_{\propto}})$ that depends on the angle $\propto$ defined as $\propto \; = c_{orient} \times \frac{\pi}{2}$. Such atlas is created by rotating the previously described I_{med_D} from the angle $\propto$ around the horizontal axis linking both extremities of the alevin's head and tail, then by taking its projection on the initial plan of the atlas $\mathcal{A}_D$, as illustrated in Figure 50. The same rotation is applied to the definition ensemble of the probability function p_{sb_D}. An illustration of $\mathcal{A}_D$, of $\mathcal{A}_L$ and of an example of $\mathcal{A}_\propto$ for $\propto = 45°$ is given in Figure 51.

ROI localization

During the ROI extraction, the atlas must be registered on a new image I. If embryos seen in dorsal or three quarters views are relatively symmetric, it is not the case for embryos seen in lateral view that can be seen in either the right side or the left side. In order to select the correct side of the atlas in case of a lateral embryo analysis, two registrations are applied: we search for the transformation mapping μ_1' between the image I and the atlas $\mathcal{A}_L$, and for the transformation mapping μ_2' between I and the symmetric of the atlas $\mathcal{A}_L$. For both results, the Dice Similarity Coefficient (DSC) is computed:

$$\mathrm{DSC}(\mathrm{X},\mathrm{Y}) = \frac{2 \times |\mathrm{X} \cap \mathrm{Y}|}{|\mathrm{X}| + |\mathrm{Y}|} \tag{4.7}$$

where X represents the binary mask of the embryo of either the atlas or of the symmetric of the atlas, and Y represents the binary mask of the embryo in the image I. The transformation that maximizes the DSC is retained as the transformation μ', used to delimit the region R in I where the probability to find the swim bladder, if a swim bladder is present, is equal to 1. The center C of the ROI $\mathcal{C}$ is extracted from R in a different way depending on the orientation. For embryos seen in dorsal or three quarters views that do not present a huge variability in the swim bladder area and shape, we take the barycenter of R. However, embryos seen in lateral view can present a variability in their swim bladder shapes from the most flattened to the most inflated one. Taking the barycenter can lead to a false detection of C on the dark swim bladder contour instead of in the light swim bladder inner part. To avoid that, the lightest point of the delimited region is taken as the center C of the ROI.

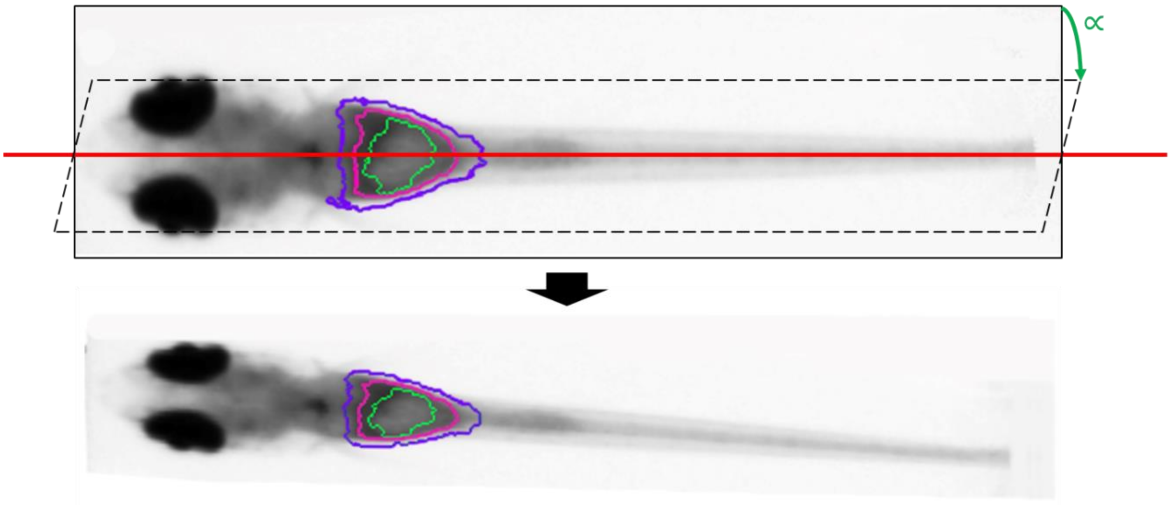


Figure 50. Generation of the atlas $\mathcal{A}_{\propto}$ by rotation of the atlas $\mathcal{A}_D$ from angle $\propto = 0.78$rad (45°) The red straight line is the rotation axis linking both extremities of the alevin. The three thin red lines are the isocontours that delimit the area where the pixel have a probability equal to 1, 0.5 and 0.05 to belong to the swim bladder.

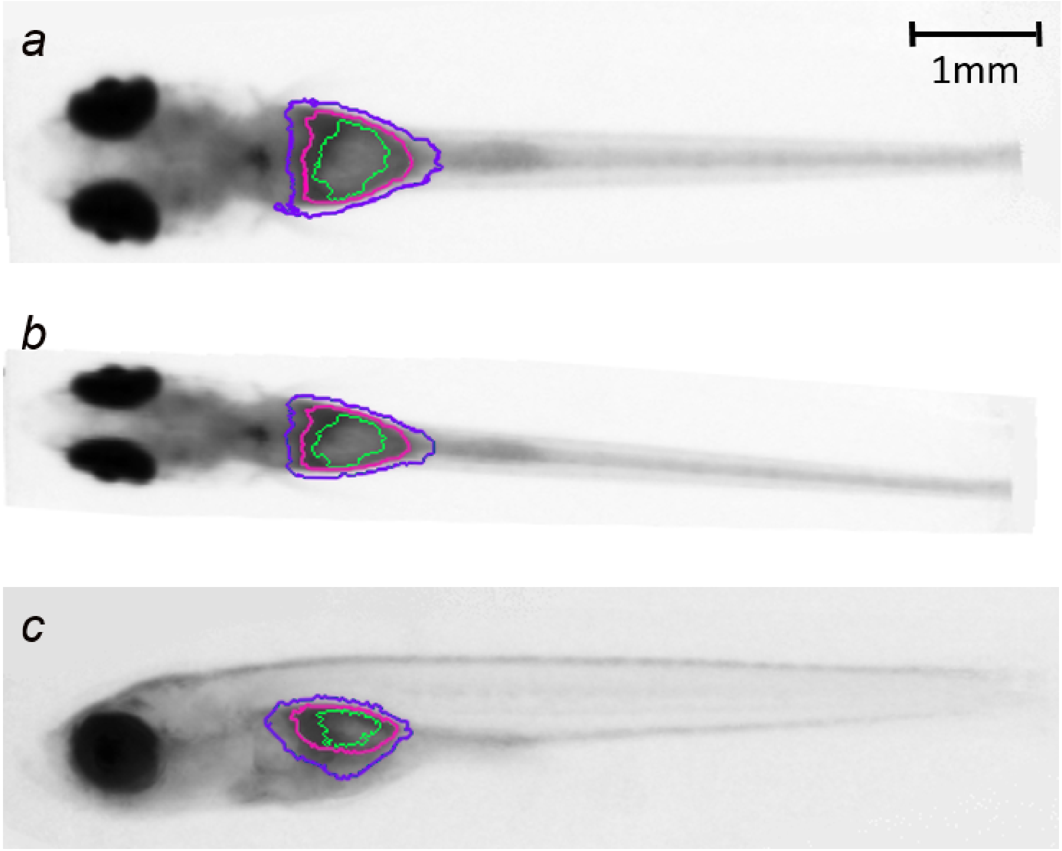


Figure 51. Results of atlas generation. a: dorsal atlas $\mathcal{A}_D$. b: three quarters atlas $\mathcal{A}_{\propto}$ obtained for $\propto = 0.78$rad (45°). c: lateral atlas $\mathcal{A}_L$. The red lines show the isocontours that delimit the areas where the pixels have a probability equal to 1, to 0.5 and to 0.05 to belong to the swim bladder.

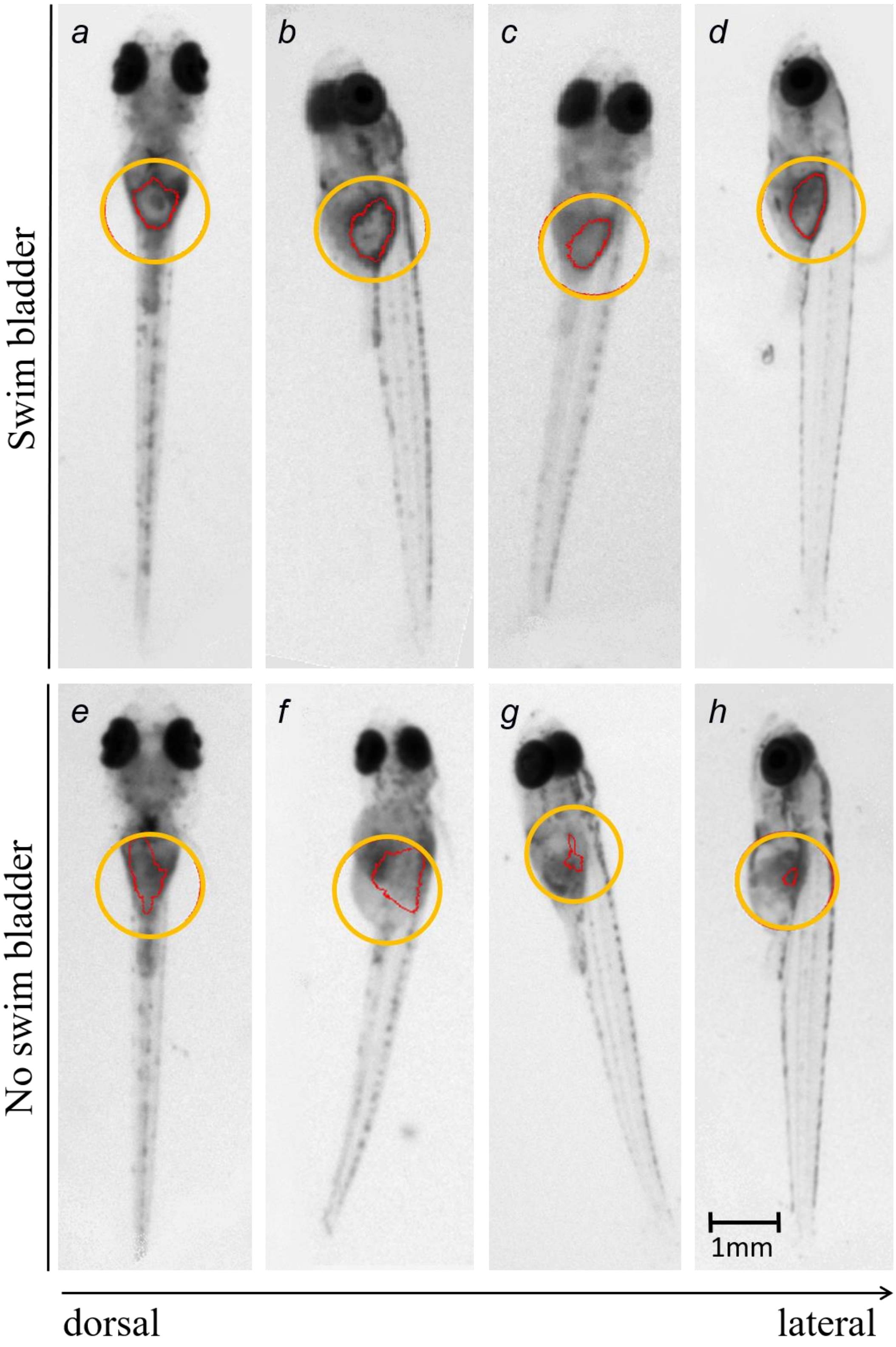


Figure 52. Swim bladder segmentation results on medaka alevins with or without a swim bladder and seen in different orientations from the dorsal view (left) to the lateral view (right). The yellow circle indicates the location of the ROI $\mathcal{C}$ and the red inner line shows the segmented most probable contour $\mathcal{SB}$ for alevins with a swim bladder from a to d, and for alevins without a swim bladder from e to h.

Extraction of the swim bladder most probable contour

Finally, the step that consists of extracting the most probable contour of the swim bladder is adapted by modifying the parameters definition. Indeed, as some healthy embryos seen in dorsal view can present a dark area in the center of their swim bladder that we don't want to detect, we selected as starting point of the shortest path a_1 the most peripheral local minimum of the first radius section. To avoid the shortest path to be on this dark area contour, we experimentally determined the value of $\mathrm{rmin} = 10$ pixels. For other embryos, this dark area is not visible in our dataset so we take the global minimum of the first radius section for a_1, and $rmin$ is set to 0. All the parameters and their associated values depending on the embryo orientation are summarized in Table 8.

As shown in Figure 52, each orientation-related method allows to optimally segment the most probable contour of the swim bladder $\mathcal{SB}$. Swim bladder characterization will now allow to distinguish cases where a swim bladder is present from those without a swim bladder.

	Parameter name	Dorsal method	Three quarter method	Lateral method
Atlas generation	Number of images used for atlas generation n	20	20	6
	Used atlas	Dorsal atlas $\mathcal{A}_D$	Atlas $\mathcal{A}_\propto$ with $\propto = c_{orient} \times \frac{\pi}{2}$	Lateral atlas $\mathcal{A}_L$
Swim bladder localization	ROI center C	R barycentre	R barycentre	R lightest point
	Starting point of the shortest path a_1	Most peripheral local minimum of the ROI first radius section	Global minimum of the ROI first radius section	Global minimum of the ROI first radius section
	Minimal radius of the shortest path research area $rmin$	10	0	0

Table 8. Parameters definition for the extraction of the swim bladder most probable contour depending on the embryo orientation.

4.3.3. Swim bladder characterization

We now need to identify if the segmented shape $\mathcal{SB}$ corresponds to a swim bladder or not, basing on descriptors of the swim bladder.

4.3.3.1. Intensity descriptors

A swim bladder is characterized by a high contrast between a dark contour and a light inner part, contrary to an embryo without any swim bladder that presents a more homogeneous intensity in the delimited shape. On the histograms of both the inner part $\mathcal{SB}$i and the contour $\mathcal{SB}$c of the segmented shape $\mathcal{SB}$, this means that a shift is visible between both distributions of pixel intensities in case of an embryo with a swim bladder (Figure 53). The following intensity-related parameters are extracted from both histograms: the minimal, maximal, average and median intensities, the intensity mode, and their piecewise differences are calculated in order to characterize the contrast between $\mathcal{SB}$ inner part and contour. We also extract the two ranges, *i.e.*, the difference between the maximal and the minimal intensities, and the ratio between them. We finally extract the pixel intensity variance of $\mathcal{SB}$, and the covering, defined as follow. We calculate the difference between the maximal value of $\mathcal{SB}$c and the minimal value of $\mathcal{SB}$i in one

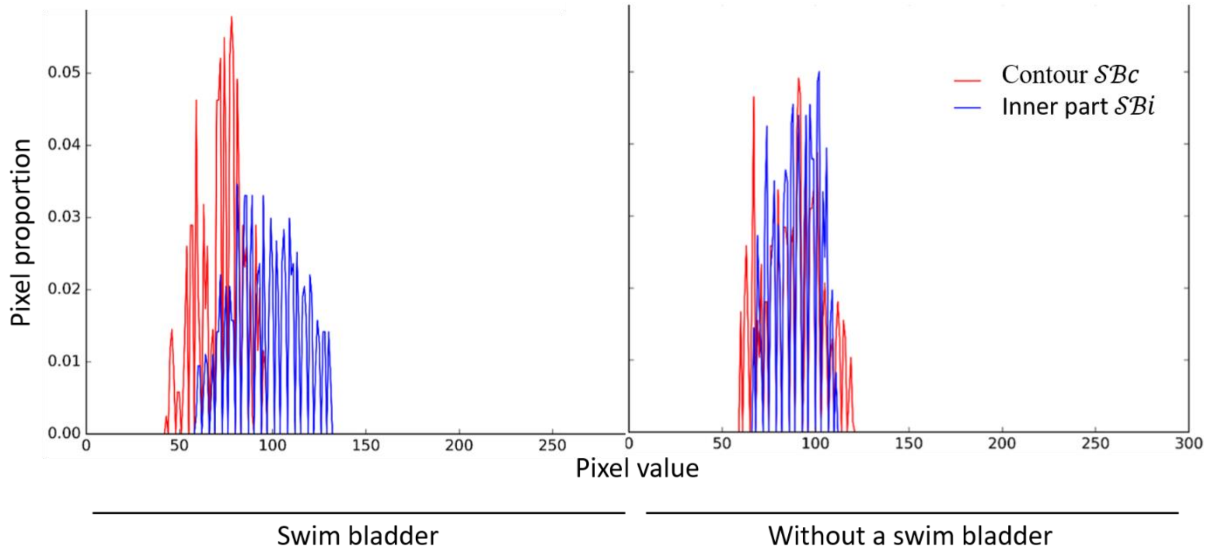


Figure 53. Histograms of the inner part $\mathcal{SB}_i$ and the contour $\mathcal{SB}_c$ of the segmented shape $\mathcal{SB}$. If a swim bladder is present, the contour is darker than the inner part, and a shift is visible between their corresponding histograms. If no swim bladder is present, this shift is not visible.

hand, and the difference between the maximal value of $\mathcal{SB}$i and the minimal value of $\mathcal{SB}$c on the other hand. The covering is defined as the ratio between both differences. These intensity-related descriptors will then be combined with other descriptors characteristic of the swim bladder morphology which are described in the following section.

4.3.3.2. Morphological descriptors

In order to characterize the swim bladder shape, the following descriptors are extracted from $\mathcal{SB}$. We refer to the *convexity* of a swim bladder, as the set difference between the convex hull of $\mathcal{SB}$ and $\mathcal{SB}$ itself [100]. We refer to the *concavity* of a swim bladder as the area of the deleted component after a morphological opening $\gamma_{\Gamma_{r_5}}$ of $\mathcal{SB}$ by a disk-shape structuring element Γ_{r_5} of size $r_5 = 5$. We furthermore consider the elongation of $\mathcal{S}$ defined by $\frac{4 \times area(\mathcal{S})}{perimeter(\mathcal{S})^2}$. All descriptors related to intensity or morphological characterization of $\mathcal{SB}$ are summarized in Table 9.

Name	Value	Name	Value
Intensity descriptors:			
$variance_i$	Variance of pixel intensity	min_diff	$min_{\mathcal{SB}_i} - min_{\mathcal{SB}_c}$
min_i	Minimal intensity	max_diff	$max_{\mathcal{S}i} - max_{\mathcal{S}c}$
max_i	Maximal intensity	av_diff	$average_{\mathcal{S}i} - average_{\mathcal{S}c}$
$average_i$	Average intensity	$mode\_diff$	$mode_{\mathcal{SB}_i} - mode_{\mathcal{SB}_c}$
$mode_i$	Histogram mode	$median\_diff$	$median_{\mathcal{SB}_i} - median_{\mathcal{SB}_c}$
$median_i$	Median intensity	$rrange$	$\frac{range_{\mathcal{S}i}}{range_{\mathcal{S}o}}$
$range_i$	$max_i - min_i$	$covering$	$\frac{h_{\mathcal{SB}_i} \cap h_{\mathcal{SB}_c}}{h_{\mathcal{SB}_i} \cup h_{\mathcal{SB}_c}}$
For i is either $\mathcal{SB}_i$ or $\mathcal{SB}_c$			
Morphological descriptors:			
$convexity$	$convex\_hull(\mathcal{SB}) - \mathcal{SB}$	$elongation$	$\frac{4 \times area}{perimeter^2}$
$concavity$	$area(\gamma_{\Gamma_{r_5}}(\mathcal{SB}) - \mathcal{SB})$		

Table 9. List of descriptors extracted for swim bladder characterization.

4.4. Assessment of the classification of alevins with and without a swim bladder

The method of swim baldder detection is assessed in this section. The experimental set-up is exposed, before presenting the classification results, and finally discussing the results.

4.4.1. Experimental setup

4.4.1.1. Experimental protocol and ground truth

The used experimental protocol is the same as the one described in Section 2.4.1.1 and Section 6.1 of the Appendix. Without any manual determination of the orientation, we record one image of each embryo at size 1500×1500.

4.4.1.2. Software and libraries

We use the same Python 2.7 environment as described in Section 2.4.1.2. We used Numpy, Scipy, Elastix [128], NetworkX and Pink libraries [106] for swim bladder segmentation, and Scikit-learn [113] for machine learning-based classification.

4.4.1.3. Dataset description and ground truth

To establish the ground truth, each fish embryo is analyzed after image acquisition on day 9. An expert observes embryos under a microscope, allowing to manipulate them and thus to see them in all possible orientations. This expert annotates each fish embryo as having a swim bladder or not.

Our experimental protocol allows to set up a total database of 406 images of embryos, among which 258 are seen in dorsal view according to the manually determined orientation (*i.e.* 63.5% belong to O_D), 119 are seen in three quarter view (29.3% belong to O_{AD}or to O_{AL}) and 29 in lateral view (7.1%). A subset of this database is used in order to generate the atlas described in Section 4.3.1 For reasons linked to the unbalanced proportions of both dorsal and lateral orientations in our total available database, we select $n = 20$ images of healthy embryos seen in dorsal view for dorsal and three quarter atlases generation, and $n = 6$ healthy embryos seen in lateral position for lateral atlas generation. The remaining 380 images constitute the validation dataset. Among those, 282 present a swim bladder according to the ground truth, and 98 do not present a swim bladder. In particular, the subset of 282 embryos with a swim bladder is composed of 195 images of fish embryos seen in dorsal view, 80 in three quarter view, and 7 in lateral view. The subset of 98 embryos without a swim bladder is composed of 43 seen in dorsal view, 39 seen in three quarter, and 16 in lateral view.

4.4.1.4. Tested classification method

A random forest classifier is defined with the following parameters, determined with the GridSearch algorithm from the Scikit learn library [113]. The number of estimators is set to 20, the maximal depth is set to 6, the minimal number of samples required to split an internal node is set to 2 and the minimal number of samples required to be at a leaf node is set to 1. The entropy criterion is chosen. In order to partially balance the subsets of embryos with and without a swim bladder, a higher weight w_- is attributed to the dataset of embryos without a swim bladder than the weight of those with a swim bladder w_+. We set $w_+ = 1$ and $w_- = 3$.

The same performance measures are used than those described in Section 3.3.1.6.

4.4.2. Classification results

An example of classification results obtained after a 5-fold cross validation process is presented in Table 10 in the form a confusion matrix that shows the distribution between embryos with and without a swim bladder according to the ground truth and to the prediction results. We reach an accuracy score of 95% in the total dataset. Moreover, the sensitivity is 90.8% and the specificity is 96.4%. For more precision, and in order to assess the results variation depending on the data partition into the training and the testing sets, a series of 500 successive 5-folds cross validations is performed. The results are presented in Figure 54 in the form of histograms of the calculated accuracy, sensitivity and specificity (as defined in Equations 3.12 to 3.15). For each of these performance criteria, a Gaussian distribution is obtained. The mean accuracy is

Prediction:	Swim bladder	No swim bladder
Ground truth:		
Swim bladder	272	10
No swim bladder	9	89

Table 10. Example of classification results after 5-fold cross validation performed on 380 tested images.

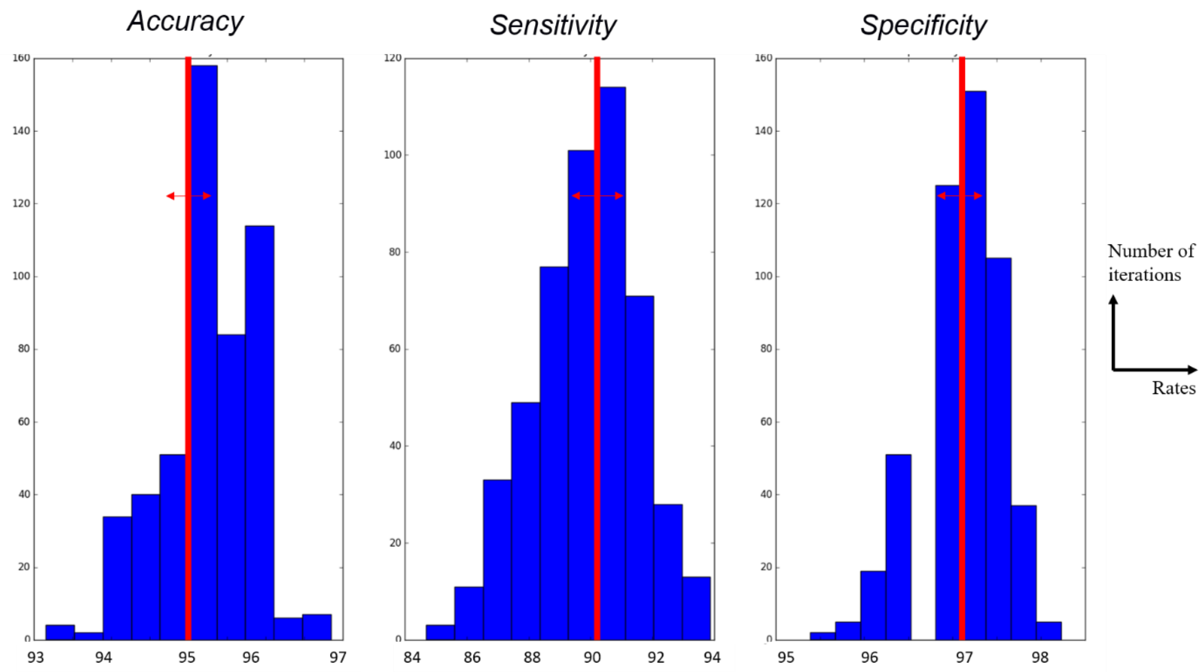


Figure 54. Histograms of the results of the swim bladder classification method after 500 successive 5-folds cross validations in terms of accuracy, sensitivity and specificity. The average value is shown with a vertical red line and the associated standard deviation with the horizontal red line. We obtain an average accuracy of 95% with a standard deviation of 0.6; an average sensitivity 90% with a standard deviation of 1.8; and an average specificity of 97% with a standard deviation of 0.5.

95% with a standard deviation of 0.6%, the mean sensitivity is 90% with a standard deviation of 1.8%, and the mean specificity is 97%, with a standard deviation of 0.5.

4.4.3. Discussion

To validate the automated method for swim bladder detection in medaka embryos, the program results have been compared to the gold standard of microscope-based observations. The overall accuracy of 95%, the overall sensitivity of 90.8% and the overall specificity of 96.6% reveal the feasibility of the automatic detection of the swim bladder from 2D images of medaka embryos.

The experimental protocol of our method presents the advantage to avoid the manual positioning of each embryo in the well whereas in many studies related to phenotypes classification, the protocol implies to manually place the fish embryo in either dorsal or lateral position [87, 41, 88]. After anesthesia, embryos remained in the incubation medium in their well and can have any possible orientation before images are acquired and treated. However, this protocol does not permit to control the orientation neither. In the studied database, a disproportion between natural positioning of embryos is revealed. Without any control on embryo positioning, we obtain far fewer alevins seen in lateral view compared to those seen in three quarter view and to dorsal view even more. In our total validation dataset of 380 images, we only have 23 images of embryos seen in lateral view (6%) against 119 embryos seen in three quarter view (31%) and 238 embryos in dorsal view (63%). Regarding the significant proportions of alevins seen in dorsal and three quarters orientation in the validation dataset and the overall accuracy, it can be can concluded that the classification correctly works on embryos seen in dorsal and three quarters position. Nevertheless, even if no systematic error is observed on laterally seen alevins through the cross validations, a most important subset of embryos seen in lateral view would improve the reliability of the validation process on lateral orientation. At final, the results of this study reveal that the presence or the absence of swim bladder can be detected with a satisfying accuracy in images of embryos, regardless of their orientation. For improving the precision and adapt the atlas-based method to the analysis of other relevant morphological parameters, a future work will consist in performing a three-dimensional atlas reconstruction of a healthy embryo, by making the reconstruction from dorsal and lateral atlases interpolation.

5. Discussion

The aim of this work has been to develop image processing-based methodologies to automate the teratogenicity assessment assay on fish embryo. This assay has been developed to detect strong developmental abnormalities following exposure of medaka embryos at various concentrations of a teratogenic chemical. This test relies on the analysis of embryos morphology after a 9-day exposure. A teratogenic index TI is calculated that depends on both the estimation of the LC50 (the concentration that causes the death of 50% of the exposed embryos) and of the EC50 (the concentration inducing malformations or death of 50% of the embryos). These estimations depend on the proper detection of dead embryos. We thus have focused first on the development of an automated assessment of medaka embryos mortality. To achieve this, an assay was developed to detect cardiac arrests in medaka embryos based on the analysis of pixel intensity variation from video sequences. EC50 calculation also relies on detection of embryo malformations. We focused on two of the most observed malformations: axial malformations and the absence of a swim bladder. Robust detection assays were developed to assess these phenotypes with the objective of robust experimental conditions and minimizing the manual intervention from the operator.

Section 5.1 is concerned with the first technical challenge, namely the ground truth subjectivity. We explain how subjectivity was quantified during our studies and its consequence on the reliability of the results. We discuss the perspective of using Deep Learning to reduce the subjectivity impact on embryos classification. All the developed procedures rely on the analysis of images and videos. Plates with one embryo per well are placed on the acquisition platform and a camera moves above each well and takes an image and a video of the whole well. In such context, the second technical challenge is the information loss incurred going from interactive 3D microscope-based to 2D image-based observations. This point is discussed in Section 5.2. In a third section 5.3, we present some perspectives for improving the test performance. In particular, functional assessment in medaka embryos could be developed by quantitative measurement of the cardiovascular function, and the embryos behavior should be analyzed. Finally, we discuss the way of assessing the whole teratogenicity method in Secti5.3.3on 5.4.

5.1. Ground truth subjectivity

The first problematic raised during this dissertation is the subjectivity linked to manual annotations, and thus to the ground truth. This subjectivity raises the question of the annotations reliability, which has an influence on the program development. Indeed, annotations have an influence on features selection. Typically, when seeking features that are representative of an anomaly we need to detect, images are manually screened and classified according to the annotations (for example, images with and without an edema according to the microscope-based ground truth). Image-based features, coefficients and thresholds applied on these features are then optimized to distinguish between both cases (for example, the distance between the swim bladder border and the beginning of the tail is a feature, and a minimal value can be associated to this feature to detect an edema). If the ground truth changes, the way of considering each feature is impacted, and thus, coefficients and thresholds also change. For this reason, subjectivity is a problem we need to carefully handle.

5.1.1. Subjectivity quantification

In this work, two different annotations were used: the image-based (or video-based) observations, used to assess the error of the detection programs, and the microscope-based observations, used as ground truth for assessing the efficacy of the whole automated method. Thus, two subjectivity assessments were conducted.

A first assessment of the inter-operator subjectivity based on video observation was performed during the validation of the mortality test and is presented in Section 2.4.2. A dataset of 200 video sequences of medaka embryos was observed by three operators. Each embryo was annotated by the three observers according to the presence or the absence of a beating heart, based solely on video. At the end of the experiment, differences were noted on 6 videos, for a rate of 3%. This subjectivity rate can be considered acceptable for the assessment of the program efficacy.

In a similar way, the inter-operator subjectivity assessment was performed on microscope-based annotations. A set of 143 embryos were successively analyzed under a microscope by three trained operators who observed the embryos under a microscope. The following endpoints were analyzed: mortality, presence of edemas, eyes, axial malformations, swim bladder, and any significant development delay. Size and pigmentation are also qualitatively analyzed. At the end of mortality assessment, 2 embryos were differently classified by the different observers (1.3%), while at the end of the morphological assessment, 30 embryos were differently classified (20%). The percentages of malformed embryos measured by the three observers on the entire dataset varied from 40% to 60% depending on the operator. Inter-operator subjectivity seems to have a low impact on the mortality assessment, as the search for a beating heart is an easy task. By contrast, this study reveals the significant impact of subjectivity on the morphological assessment, which is a reason we decide to automate the process.

During automation, the first impact of subjectivity, *i.e.* of annotations change, is on the choice of the features (which parameter to consider to detect an edema? The distance between the swim bladder bottom and the tail for example). The second impact is on the way to consider each feature to discriminate the data (which value must this distance have to reveal an edema?). In other worlds, annotations change impacts the coefficients and thresholds applied to features to correctly classify data. It is not conceivable to manually adapt coefficients and thresholds each time a change appears on the ground truth, as it is long and tedious. For this reason, we decide to use a supervised machine learning-based classification method for classifying healthy and malformed alevins with a method that can be retrained easily with new annotations. Under this strategy, features are measured (distances, area, ratios, angles, etc.), then a random forest classifier is trained to automatically find the most relevant features and their associated thresholds (test functions optimization as described in Section 1.2.2), by learning from annotations, used as training labeled data. The training step is very fast and can be repeated as often as necessary in response to changes in the annotations.

Random forest is an efficient way to quickly adapt the classification model to the annotations, by training the model on labeled data coming from these annotations. Nevertheless, if annotations are biased by subjectivity, the classification model may inherit this bias. To lessen the impact of inter-operator subjectivity on the morphological assessment, a morphological assessment could be performed on a set of embryos according to the following procedure: three observers would annotate a set of embryos under a microscope according to the presence or absence of a malformation. A consensus would be made between the three observers conclusions, and would constitute the final ground truth. As such a process is time-consuming and require the availability of several trained operators at the same time, it should be staggered through time. Resulting consensus data would progressively replace the labeled data coming from the analysis of a single operator and be used for the classification model training. The idea would be to progressively optimize the reliability of the classification model that would be trained on these new consensus data. Alternatively, when multiple observers are not available at the same time, it is still useful to augment the training dataset based on new data annotated by different, trained observers, even if each annotation is only performed once. Assuming observers are reasonably consistent, this would progressively build a statistically significant ground truth. This would still progressively be improving the reliability of the classification model. This is the method we plan to use.

While these strategies can help limit the influence of subjectivity on the classification results, the chosen features themselves are not adaptive. To improve the sensitivity of the test, new features will have to be developed. These features can be manually developed as we have done, or they could be automatically identified by using deep-learning-based methods.

5.1.2. Deep learning for embryos classification

In this manuscript, some alevins phenotypes classification problems were tackled by developing empirical (for mortality assessment) and machine learning-based approaches (for malformations detection). These studies illustrate the consensus in image-related research to the effect that different classification, recognition and learning tasks require different image representations in order to extract the desired information from data and interpret them. The central challenge when learning from images is thus to find relevant data representative features, which are specific to the purpose of the study [129]. In this work, we chose to manually design (to "engineer") the features, which up to 2015 (when this work was started) was the most common practice to characterize data, but also requires time and significant domain knowledge in the concerned field. A more recent Machine Learning practice that does not require hand-tuned features is Deep Learning [37, 130, 131].

Deep Learning refers to a subset of machine learning and artificial intelligence based on the building and the training of large artificial neural networks to represent data. An Artificial Neural Network (ANN) is defined as a computational model that approximates the structure and functions of biological neural networks. Each individual neuron of the ANN corresponds to a nonlinear processing unit. Neurons are arranged in layers within the ANN. Each layer takes

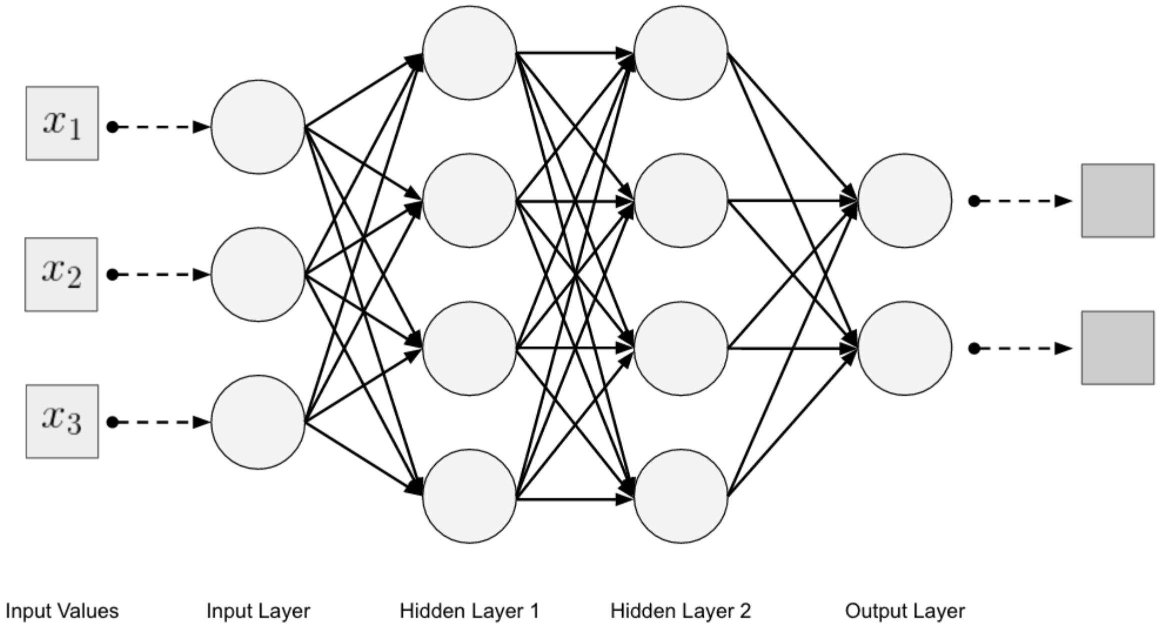


Figure 55. Principle of an Artificial Neural Network (ANN). The data is given as input to neuron of the input layer (left). Each neuron corresponds to a non-linear treatment unit that extracts features from the data. Outputs of all neurons of the input layer are given to neurons of the following layer, and so on, until reaching the final output layer. As the data is passed through deeper layer of the neural network, higher level parameters and thus features of the data are analyzed.

as input the output of the previous layer and is responsible for extracting or grouping features. If most modern machine learning-based algorithms present such a structure in layers of processing units, learning algorithms, such as ANN architectures, are considered to be *deep* if they include more than 3 hidden layers, if they use modern activation layers, improved optimization algorithms and other techniques that make these architecture efficient and effective (Figure 55). Deep ANN have become better approximations of actual known neural architectures, particularly of the visual cortex, through the use of convolutional layers. As the data is passed through successive layers of the network, higher levels parameters of the data are grouped and higher levels of abstraction can be represented. For this reason, deep networks can model complex relationships between input and output and to extract useful patterns from data, without requiring feature engineering.

In image recognition and classification, the principle of ANN could be vulgarized as follow. The input layer of the algorithm takes the initial image as input. The image is considered in its entirety, meaning as a spatial distribution of pixel intensities. Each pixel is treated by a neuron of the first hidden layer that passes the corresponding intensity to all the neurons of the second hidden layer. Neurons of the second hidden layer analyze the intensity variations between neighbor pixels. Neurons of the third hidden layer begin to analyze intensity variations in a group of neighbor pixels, allowing to extract first basic shapes (as lines). Neurons of the fourth hidden layer analyze the links between them, allowing to extract more complex shapes by combination of several lines. The process continues until reaching the final output layer and obtaining a complex characterization of the image.

Whereas ANNs have been proposed many decades ago (the prototype of a single artificial neuron, called the Perceptron, was proposed in 1958 [132]) and multi-layer architectures exists since the late 1980s [133], they have become popular in recent years in many research fields including speech recognition [134], natural language processing [135, 136], pattern recognition [137, 138], image and video recognitions [139, 140, 141], and life sciences [142, 143, 144, 145]. Deep learning has become a very successful branch of machine learning, that excels when the working data are unstructured, sparse, and large [146]. Among scientific fields that investigate the usefulness of deep learning, we find medical and pharmaceutical studies [147]. In particular, deep learning approaches can help to establish links between the modelling of a molecular structure of a chemical and a particular effect of this chemical. Thus, it is especially useful for drug design and toxicity prediction [148, 149, 150, 151, 152, 153]. Considering the applications of deep learning, in particular for image classification [154, 155], it appears promising for improving the test performance. Firstly, it would facilitate and accelerate the process of data characterization, in particular for the recognition of complex phenotypes such as edemas. Secondly, the idea behind deep learning is to discover multiple levels of representation in data, leading to more abstracted concepts. More abstract concepts are generally invariant to most local changes of the input data. For categorical concepts as our binary classification into healthy and malformed embryos, more abstract representations could detect categories that cover more varied phenomena and thus they could have greater predictive power [156]. Nevertheless, Deep-Learning approaches require vastly more annotated data than classical learning methods (around several hundreds of thousand data used for training). This implies a high computing power. Annotations are also typically required to be more precise, e.g. including segmentation masks to highlight the regions of interest. Such conditions are not easy to obtain and is likely to explain why deep learning techniques have not been commonly used in the scope of fish embryo studies for now. For our purpose, a higher number of data would be necessary to investigate the use of deep learning approaches: a dataset 100 times larger than those used in this work should be reasonable.

To conclude, if literature highlights the performance of deep learning approaches in many scientific domains including image analysis and pattern recognition, only a few applications are related to toxicity screening for now, and none on fish embryo phenotypes classification. Nevertheless, the development of high throughput screening assays based on the classification of a large number of fish embryos images definitely appears to be adapted to the development of deep learning techniques. After gathering a wider set of images, a research axis could consist in manually classify cropped images of several complex embryos phenotypes and train a neural network on this database.

5.2. The technical challenge of information loss from 3D interactive observations under a microscope and 2D image-based observations

The main technical issue raised by our experimental setup is the information loss between what can be observed under a microscope by manipulating embryos (termed here 3D interactive, since they can be observed from multiple points of view), and what remains visible on the 2D acquired images and videos processed by our proposed assays. Indeed, when observing embryos under a microscope, embryos are seen in color. The operator can zoom in on details and manually make the focus. Finally, he has the possibility to change the position of the anesthetized embryos in order to detect all anomalies. Thus, resulting annotations are more reliable. In comparison, when images and videos are acquired, the resolution is fixed, and each embryo is seen from a single orientation in its well. Depending on the orientation of the embryo in the image, some information can be occulted (for example an axial malformation, of a beating heart occulted by the dark eyes in eggs). For this reason, quantifying the information loss between 3D interactive observations under a microscope and 2D video- or image-based observations appears necessary.

5.2.1. Assessing the amount of information loss between 3D and 2D observations

To assess the efficacy of an automated assessment method, it is important to quantify the error due to image processing (*software classification error*); the error due to the acquisition process (*information loss*), and the error of the whole anomaly detection method (*overall error*).

To do so, two different annotations were considered during our studies: the microscope-based annotations, which we consider as the ground truth since the embryo can be observed from all possible orientations under a microscope, and image- or video-based annotations, where the embryo is observed solely on the acquired image or video. The following assessments were made.

- On the one hand, the software classification error was quantified by comparing the programs results to the image- or video-based annotations. In the case of the detection of cardiac arrests, the obtained error rate is less than 2% (Section 2). For the axial malformations detection assay, the error rate is close to 3%. (Section 3).
- On the other hand, the information loss was quantified by comparing the two annotations. For both the detection of cardiac arrests and the detection of axial malformations, they differ by 10-12%. Most of time, these differences correspond to a heartbeat that is not visible on video (especially for eggs where embryos are folded into the chorion), or an axial malformation that is not visible on images of alevins because of the orientation. For the swim bladder detection, no difference was noticed between the two annotations, meaning that the swim bladder is visible regardless to the alevin orientation.
- Finally, the overall error was measured by comparing the program results to the microscope-based ground truth. For the mortality assessment, the overall error rate as measured on 566 newly generated videos was 18% (Section 2.4.4). For the axial malformations detection assay, we achieve an overall error rate of 15%. For the swim bladder detection, the overall error rate is equivalent to the software classification error rate, which is 5%.

With software classification error rates no higher than 5%, we note that our assays always achieve results that are comparable to those obtained by human when the analysis is performed on the same data. This validates the relevance of the developed features for each anomaly.

When no significant information loss exists between the microscope-based and image-based observations, as it is the case for swim bladder detection, we obtain a satisfactory overall error rate of 5%. This highlights the fact that, when looking at an anomaly that is visible regardless to the alevin orientation on images, an acceptable overall error rate can be achieved. In contrast, for the detection of cardiac arrests and axial malformations, the overall error is in majority due to the information loss between microscope-based and video- or image-based observations: not all heartbeats or all axial malformations that are detectable under a microscope are visible in the corresponding videos or images. This results in a decrease in the overall specificity of the mortality assay (a few alive alevins are correctly identified) and in the overall sensitivity of the axial malformations detection assay, (a few axial malformations are detected). To compensate this information loss, malformations must be analyzed in their entirety, so that an embryo that presents an anomaly which is visible on image could be detected even if another anomaly is not visible on the image.

5.2.2. Global morphological assessment instead of individual malformation assessment to limit the information loss from 3D to 2D observations

Since each malformation is properly detected by the program when visible on images, it is pointless to try to improve this specific malformation detection. An option to improve the overall sensitivity of the test is to develop new features. Nonetheless, any new feature will be sensitive to the information loss. To limit this information loss, malformations must be analyzed in their entirety.

Until now, studied malformations have been individually assessed, leading to an overall accuracy percentage for each detection test that is more or less satisfying. However, a screening test does not necessarily require the detection of each possible abnormality. Screening aims at alerting if at least one abnormality is detected, regardless of its type. If an exposed embryo shows several abnormalities, the absence of detection of a particular anomaly is not a problem since another one is detected. For example, an alevin may have an edema and an axial malformation which are difficult to detect on the image, but it may also not have a swim bladder, which could be more easily detected by the program. Thus, by combining all developed features and by gathering all malformations appearance into a binary pair of classes, *i.e.*, “malformed” vs. “healthy” may increase the sensitivity of the whole assessment method. An alevin with several anomalies will have a higher probability to be detected by the program.

The study described in [41] tends to confirm this hypothesis. In this project, a so-called “Two-third classification” is performed that consists of (i) classifying embryos images between “dead”, “chorion” and “other”, then (ii) classifying images of the “other” category according to the presence of defects. For the second classification, each phenotype was considered individually, including the “Normal” (healthy) phenotype and each analyzed defect (axial malformations, yolk, etc.). Each phenotype leads to a binary classifier, which classifies an image as positive if the concerned phenotype is detected and as negative otherwise. Two strategies were compared to evaluate the presence of a “Normal” phenotype. The first consists of considering that each image that has never been classified in the positive class for any defect phenotype belongs to the “Normal” class. If this procedure allows to distinguish each defect of alevins, it also tends to accumulate errors made by the other classifiers and thus results in extensive over- or under-estimation of the proportion of “Normal” phenotypes. The second strategy is to use the binary classifier built for the direct recognition of the “Normal” phenotype. This second strategy appeared more robust for this application.

To test this hypothesis in our project, we should gather all features to build a random forest classifier based on a new training set of images of alive alevins where each image will be labeled as “malformed” or “healthy” according to the microscope-based ground truth. This means all possible malformations will be considered, even those that were not specifically studied in this work. Features combination is especially useful as a same feature can be relevant for the detection of different malformations. For instance, features related to the size of the alevin (Section 3.2.2.1) could be useful to detect large edemas. Some features related to the eyes (gap between eyes or eyes circularity in Section 4.2.2) were initially developed for orientation identification but can also be used for the detection of eyes malformations. In the first instance, we plan to train the classifier on around 500 available labeled images. Then, the classifier should be tested on new unlabeled data to estimate the accuracy of the automated malformation detection method. Second, as a classification model is more reliable when trained on more data, the classifier should be trained on a larger dataset. A training dataset of 3000 data is deemed reasonable, but it could be augmented as long as it improves the classification accuracy. If no evolution is noticed on the accuracy after three trainings, the classification model should be considered as stable, and training should be stopped.

This new model is expected to increase the sensitivity of our automated malformation detection method. Nonetheless, some alevins with a single malformation will probably remain undetected. To estimate the impact on the efficacy of the overall teratogenicity test, we need to assess the impact of this information loss on the calculation of the teratogenicity index.

5.2.3. Consequences of the 3D-2D information loss on TI calculation

The efficacy of the global teratogenicity test depends on its ability to correctly classify teratogenic and non-teratogenic chemicals, according to their teratogenicity index TI. Thus, to maximize the test performance, it is important to be as precise as possible in the teratogenicity index calculation. However, information loss between 3D interactive microscope- and 2D image-based observations has an impact on the automated detection of anomalies. Thus, it could have an impact on the precision of TI calculation. Detecting a higher number of dead embryos than reality for all concentrations of the tested chemical could result in a shift of the dose-response curve, leading to a decrease of the measured LC50 (Figure 56a). In a similar way, automated malformations assessment is expected to be less sensitive than the one obtained with a visual assessment made under a microscope, as all malformations are not visible in images. This may result in an increase in the measured EC50 (Figure 56b). Combining both LC50 decrease and EC50 increase could result in a decrease of the calculated TI (according to Equation 1.1). Nonetheless, a change in TI does not necessarily result in a decrease of the global test sensitivity. In a further validation step, the efficacy of the manual and of the automated methods should be compared, and the impact of automation on the precision of the teratogenicity test should be assessed. This is discussed in Section 5.4.

5.2.4. A perspective to overcome information loss for morphological assessment: tomography reconstruction for 3D atlas building

To answer the problem related to the alevin orientation and information loss, a process was presented in Section 4 of this manuscript consisting of generating atlases of a healthy alevin seen in several orientations. Such atlases represent alevin anatomy. Automated registration to a reference alevin together with these atlases has allowed us to easily identify organs and body parts. For better accuracy, it would be useful to generate a 3D atlas of a healthy alevin.

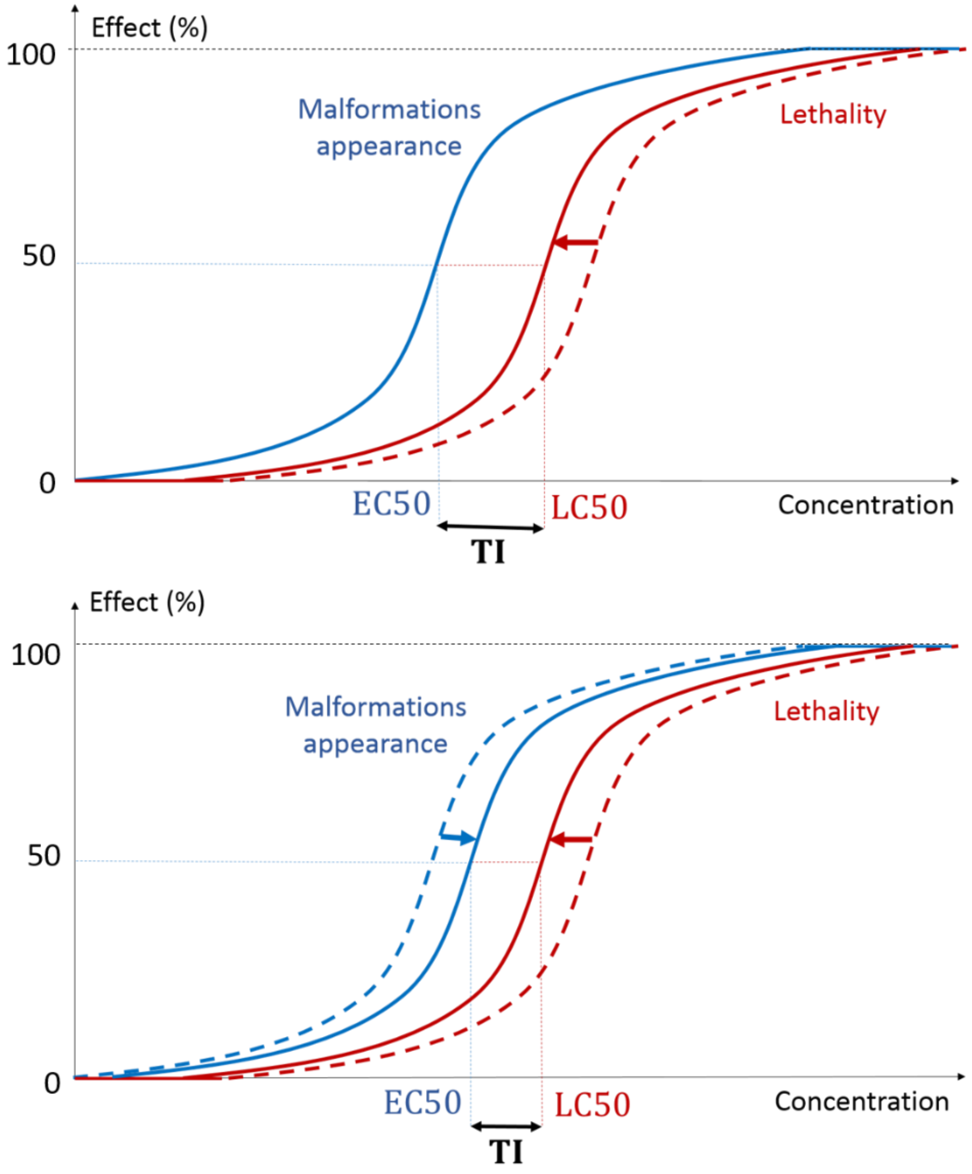


Figure 56. Consequence of the embryos assessment results on the TI calculation. a: shift of the mortality curve when detecting too many dead embryos. b: shift of both the mortality and the malformations when detecting too many dead embryos and too many malformed embryos. We notice that the shifts result in a reduction of the teratogenicity index TI wich is the ratio between LC50 and EC50.

In the context of fish-embryos based studies, robust 3D atlas generation has been performed on zebrafish using confocal imaging. This optical imaging technique uses a spatial pinhole to block out-of-focus light in image formation, allowing to capture multiple two-dimensional images at different depths inside the alevin. These multiple images can then be used to perform three-dimensional reconstruction of alevins [90]. In lieu of confocal imaging, a solution would consist of using a sparse tomographic reconstruction from the two dorsal and lateral atlases used as 2D projections of the alevin. Tomographic reconstruction is defined as a type of multidimensional inverse problem where the challenge is to yield an estimate of a specific 3D system from a finite number of 2D projections. In other worlds, the tomography process maps an internal parameter of an "object" using cross sections or slices, based on external non-invasive measurements and on computer-assisted calculations. The mathematical basis for tomographic imaging was introduced by Johann Radon [157, 158, 159] and is widely used in medical imaging [160, 161]. A notable example of applications is the reconstruction of computed tomography (CT) images where projection images of patients are obtained by propagating X-rays through many orientations of the patient [160, 162, 163]. More precisely, the system can be described as follow. The patient is placed into a rotating X-ray tube, composed of a X-ray source and a detector. The path of X-rays through the patient, from the source to the detector (at a distance d) constitutes the considered line section (Figure 57). When passing through the patient, the X-rays are attenuated. The exit beam intensity depends on the crossed tissues and can be measured by integrating the signal intensity along the line section between X-ray source and the detector:

$$I_d = I_0 \exp[-\int_0^d \mu(s; \bar{E}) ds], \tag{5.1}$$

where I_0 is the initial X-ray intensity, I_d is the projected X-ray intensity after crossing the tissue, and μ is the linear attenuation coefficient which is function of the location s along the line section and of the effective energy $\bar{E}$ at location s. This process is repeated for each rotation of angle ϕ of the rotated X-ray tube.

In our study, the transparent alevin can be assimilated to the patient and the light to the X-rays beam. Consider the representation of the alevin on the three-dimensional space represented by the orthonormal frame (O, x, y, z). Two perpendicular projections of the alevin are available in this frame (the two atlases) and correspond to the dorsal projection supported by the plan (O, x, z), and the lateral projection supported by the plan (O, y, z). Each pixel $p_1(x_{p_1}, 0, z_{p_1})$ of the dorsal projection corresponds to an intensity $I_d^D(p_1)$. Similarly, each pixel

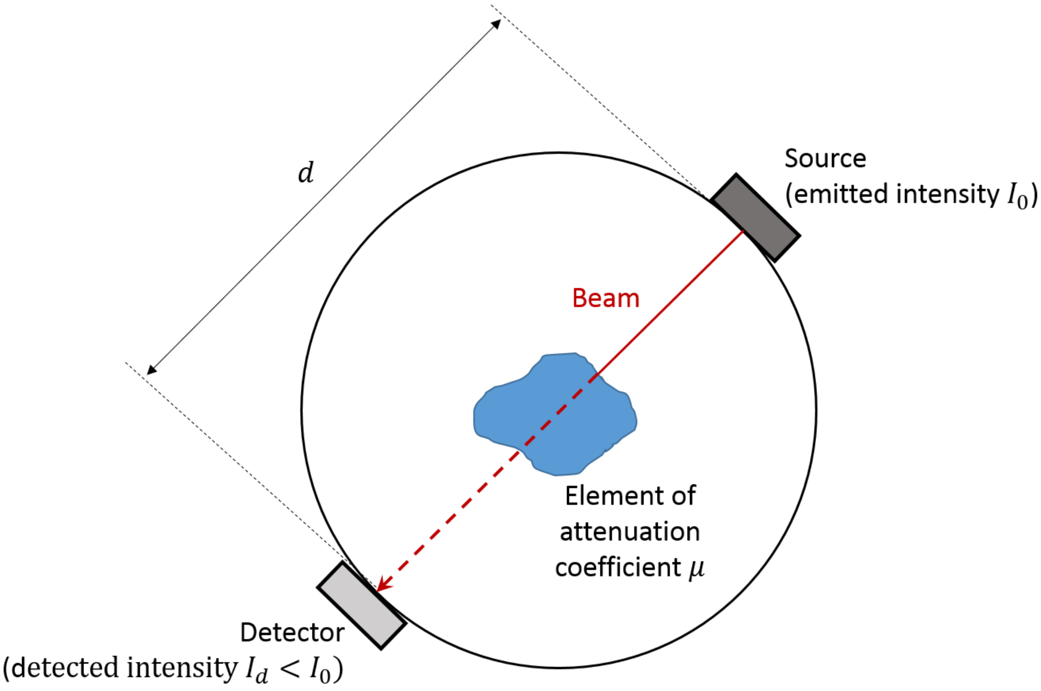


Figure 57. Tomography principle. The X-rays beam is transmitted along a distance d from the source to the detector, and is attenuated by passing through the element with an attenuation coefficient μ.

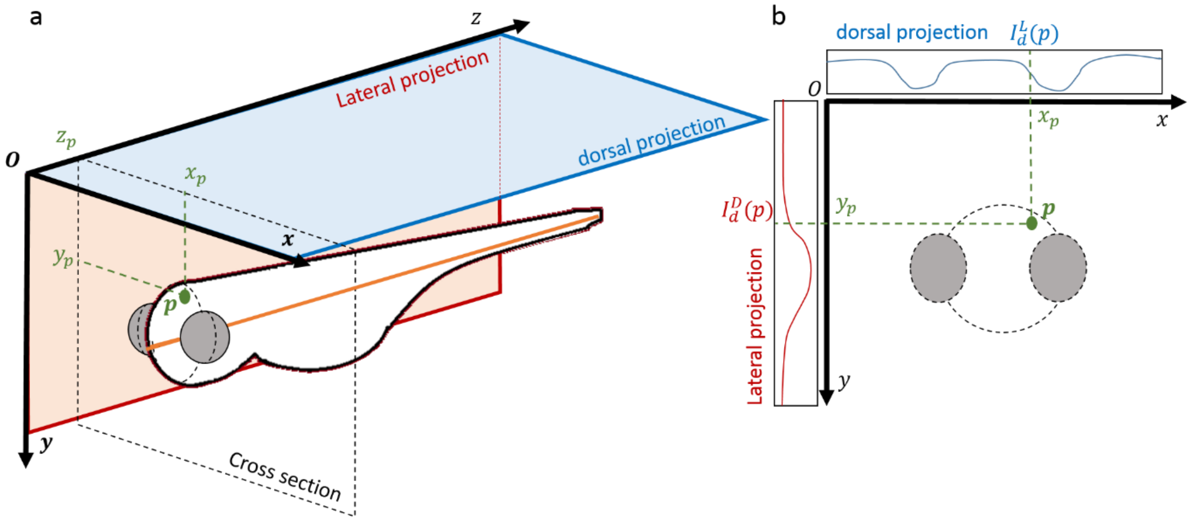


Figure 58. Principle of reconstruction tomography of a 3D alevin from the 2D dorsal and lateral projections. a: representation of an alevin in the three-dimensional orthonormal frame (O, x, y, z). b: representation of the 2D cross section at position z_p in the orthonormal frame (O, x, y). The relation between the point $p(x_p, y_p, z_p)$ and its corresponding intensities in the dorsal projection $I_d^D(p)$ and in the lateral projection $I_d^L(p)$ is shown.

$p_2(0, y_{p_2}, z_{p_2})$ of the lateral projection corresponds to an intensity $I_d^L(p_2)$. To each point $p(x_p, y_p, z_p)$ of the three-dimensional space, a doublet of projected intensities $(I_d^D(p), I_d^L(p))$ is associated. The predicted intensity of the point p can be calculated by interpolation of these two projected intensities (Figure 58). Thus, reconstruction tomography could deliver the three-dimensional internal structure of the entire medaka embryo organism (3D atlas). Note that classical reconstruction methods require a large number of projections. However, recent iterative reconstruction methods can cope with very few projections. These techniques are called tomosynthesis or limited-angle tomography rather than CT reconstruction. They are particularly used in medical imaging for breast imaging [164, 165].

Once the 3D atlas is built, it should be used for analyzing 2D images of alevins. The orientation coefficient c_{orient} previously described could be used to find the plan of the 3D atlas that corresponds to the analyzed 2D image. This plan would be considered as the 2D atlas for the considered orientation. Then, comparison would be made between the 2D image and this atlas to reveal abnormalities. Such process would allow to extend the detection programs already developed to the analysis of other organs, by optimizing the precision, and circumventing the difficulty linked to alevin orientation [166].

5.3. Improvement of the teratogenicity test performance

In this work, we especially have focused on the development of the morphological assessment of medaka embryos. To improve its sensitivity, more features must be developed in order to analyze other malformations such as edemas. However, the analysis of the embryo morphology does not allow to detect every abnormality that could occur during the embryonic development of medaka. To detect subtle anomalies and to improve the performance of the teratogenicity test, a study could be conducted not only on the morphology, but on the function of organs, through functional assessment of embryos.

5.3.1. Case of alevins edemas

Among malformations that are considered during the visual assessment made under a microscope, the detection of edemas remains problematic. An edema is a swelling of the body

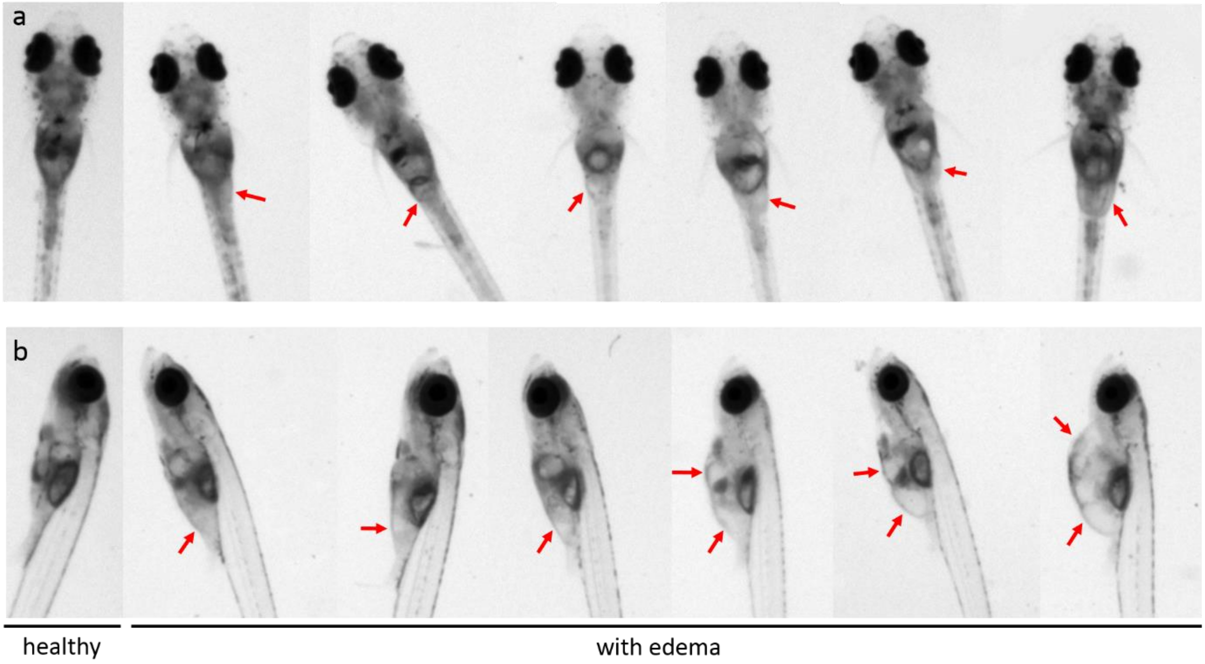


Figure 59. Variability of edemas appearance for alevins seen in dorsal view in a, and for alevins seen in lateral view in b, compared to healthy alevins. Red arrows indicate the location of edema which are thinner on the left side, and larger on the right side.

due to an excess of fluid. This malformation represents about 40% of the malformed alevins, which is significant. However, even when observing medaka embryos under a microscope, edemas are not clearly visible. When visible, the high variability of their appearance complicates their detection. An edema can appear at different locations, around the heart (pericardiac edema) or within the trunk region. The size of the edemas varies from the most prominent swelling to the thinnest bubble (Figure 59). In this paragraphe, we refer to three different types of edemas: the large, the intermediary, and the thinnest edemas. If large edemas are easy to characterize by measuring the yolk size, other edemas remain difficult to detect, which is a problem since they represent around 75% of all edemas. Intermediary edemas refer to edemas which are visually detectable for a trained operator on images, but which cause minimal swelling of the yolk. They appear in the trunk region, and can easily be mistaken for a swim bladder. When seen in dorsal view, they have the form of a slight protrusion at the frontier between the trunk and the tail, whereas healthy alevins present a spindlely shape (Figure 59a). These edemas cannot be reliably detected with size related features and are difficult to characterize. Concerning thinnest edemas, they are especially difficult to detect even under a microscope. The information loss due to the 2D acquisition causes these edemas to become invisible on the acquired image. In addition, all edemas are generally more visible on alevins seen in lateral view, which represent less than 10% of all embryos (Figure 59b).

Thus, edema is a frequent malformation that should be considered by the automated method. Currently, some developed features related to alevin size, such as maximal width, may be useful for the detection of large edemas. Intermediary case of edema could also be detected by analyzing the protrusion of the yolk, especially when seen in dorsal view. Features such as length and width from the lower part of the trunk, should be extracted. However, most edemas remain very difficult to detect with the information loss due to the current acquisition settings. Other indirect features should be found to detect such anomaly.

5.3.2. Orientation as a descriptor of alevin health?

Alevin orientation is a recurrent concern for the correct morphological assessment of medaka embryos. Fir this reason, we proposed a method to automatically identify the orientation of alevins (described in Section 4.2.2). The program validation involved to build a validation dataset representative of the alevins orientations. Images were manually classified into four categories: the dorsal orientation class O_D, the lateral orientation class O_L and the two intermediary orientation classes O_{AD} and O_{AL} (for “almost dorsal” and “almost lateral” respectively). This step highlighted the disproportion between the different alevins orientations on images. In particular, two observations were made:

- when looking at healthy alevins, we noticed that more than 80% are seen in dorsal view while less than 4% is seen in lateral view;
- when looking at all alevins seen in lateral view, less than 15% are healthy. Most of the alevins seen in lateral view show important malformations, as axial malformations or edemas. Large edemas seem to make the alevin tip over in the lateral side. Similarly, axial malformations influence the alevin balance.

While these observations have not been yet precisely quantified, they raise the hypothesis that the mere alevin orientation could be used as an indicator of the presence of a developmental anomaly in medaka embryo, even if no obvious malformation is present. Healthy alevins seem to naturally align themselves to the dorsoventral position. Alevins seen in lateral view and that do not present a visible malformation are sufficiently unusual that they raise the possibility that they are victim of some anomaly that could explain their uncommon position. We could make the hypothesis that the disequilibrium is caused by a small asymmetrical edema. Following this assumption, alevins detected as seen in lateral view should be considered as malformed.

5.3.3. Eggs particular case

Assessment of eggs morphology is challenging for both the operator and the automated analysis. At 9dpf, if unhatched, the embryo is tightly folded inside the chorion, making the mortality and malformations assessment difficult to perform. As described in Section 2.2.4, the program is able to automatically distinguish eggs from alevins. However, in the acquired videos and images, heartbeat and malformations are generally occulted by the folded embryo’s body, making the automated analysis of eggs less reliable. A solution would be to manually remove the chorion and thus force the hatching before observation and image acquisition. Unfortunately, this process is tedious and complex, especially because medaka chorion is hard and there is a risk to wound the embryo during this process. Such lesions could be wrongly attributed to an effect of the tested chemical during the analysis. Another option is to use a protease treatment (pronase) to digest the chorion. However, this implies to expose embryos to the enzymatic solution, which could damage them.

If no information can be obtained on eggs by assessing the mortality or the presence of malformations, it is necessary to obtain information in another way. Medaka eggs normally hatch between 8 and 9 days. Thus, a delay in the hatching process at day 9 might be considered as a developmental delay. Nevertheless, is such developmental delay leading to a developmental anomaly? Under normal husbandry conditions, a medaka embryo with a slight hatching delay generally develops without morphological abnormality afterwards. In this case, hatching delay is not a developmental abnormality. Nevertheless, when exposed to some compounds such as thyroid disruptors, embryos never hatch. In this case, the delayed hatching is relevant for teratogenic assessment. How to distinguish both cases? A way to identify a developmental delay that is caused by a tested chemical is to compare the number of eggs in the exposed population and in the control population. Indeed, a high number of eggs in the control group can reveal an artefact problem due to exposure conditions such as a shock during transportation or a temperature change. On the contrary, if eggs hatch normally in the control group whereas many eggs are present in the exposed population, this reveals an effect of the chemical on the development of embryos. In such cases, eggs are generally still not hatched a few days later. This observation tends to confirm a significant effect of the chemical. To overcome this difficulty, the teratogenicity assessment must include a measurement of the hatching rate for each exposure condition. The hatching rate of exposed embryos would then be normalized by the hatching rate of controls. After normalization, if a low hatching rate is measured for a concentration of a studied chemical, the chemical could be considered as having an effect on medaka embryos at this concentration. As our software is able to distinguish eggs from alevins, it is possible to fully automate this process. As images are named depending on the exposure condition, hatching rates could be automatically calculated for each condition, and normalized by the hatching rate measured in controls.

5.3.4. Quantitative assessment of cardiovascular function

In our work, we limited the cardiac function assessment to the detection of a heartbeat. Nevertheless, other cardiovascular parameters may be indicators of an impaired cardiovascular system development such as arrhythmia. A quantitative study of cardiovascular parameters (measure of the heart frequency, of the blood flow throughput) would allow to complete the qualitative assessment currently performed on the medaka embryos.

In our method, a signal of the pixel intensity related to time is recorded in the heart region of the alevin during 1 second. By recording the same videos during a longer duration, this signal would reveal the periodic heartbeat and may allow to measure a heart frequency. Medaka hearts normally beat at a frequency of around 130 beats per minute (bpm) [96]. We estimate that at least 10 beats are required to properly measure the frequency and detect arrhythmia, meaning 5 seconds would be sufficient for a control. To ensure the correct analysis of embryos that show bradycardia, a 10-second-long video seems reasonable. This is the usual duration for such study [84, 89, 96]. The frequency can be measured by Fourier analysis of this signal. This method allows to decompose a potentially noisy and complex periodic signal in its frequency domain using the well-known Fourier Transform, and thus to extract the frequency which is the most represented in the signal (Figure 60). This method is fast to implement and so highly used in recent studies for heart frequency extraction [84, 89, 92, 96]. Nevertheless, it limits the analysis to the heart frequency. Another study analyzes the intensity signal in the time domain to reveal periodic intensity changes: the periodic heartbeat. The frequency is measured by calcu-

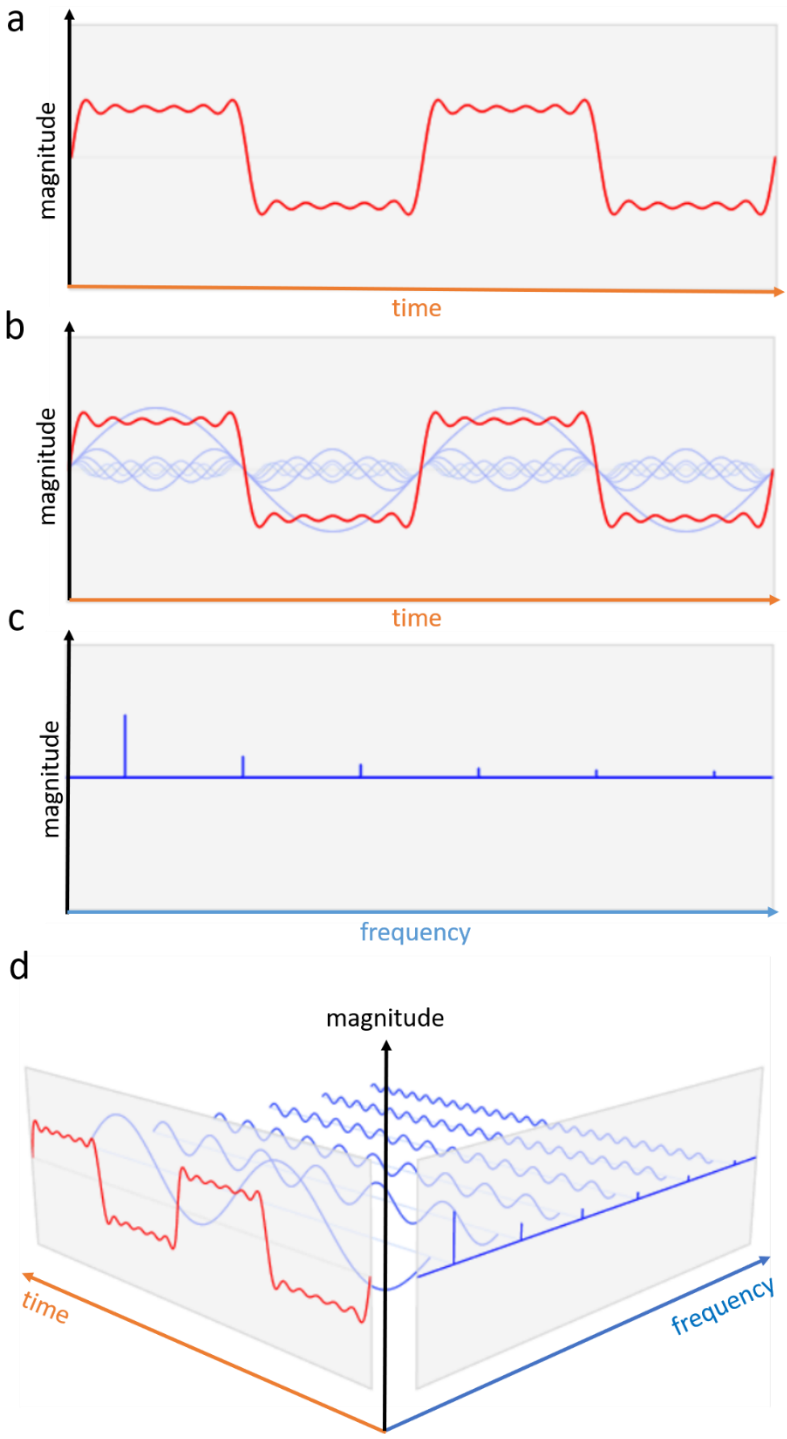


Figure 60. Representation of a complex continuous one-dimensional signal in the time domain and on the frequency domain by application of the Fourier Transform. a: the initial complex signal represented in the time domain. b: decomposition of the complex signal into a series of sinusoidal signals with increasing frequencies. c: resulting representation of the complex signal in the frequency domain. Each peak represents a sinusoidal signal, from the signal with the highest magnitude but the lowest frequency (the fundamental signal we want to extract), to the signal with the lowest magnitude but the highest frequency (noise). d: representation of the whole process of signal decomposition for representation in the frequency domain.

lating the average beat-to-beat interval (obtained by measuring the time interval between successive local maxima in the signal), and by calculating its reciprocal value. By comparing the frequencies measured in exposed embryos and in controls, we could reveal tachycardia and bradycardia. The regularity in heart contractions is also analyzed by the authors by measuring the Root Mean Square of Successive Differences (RMSSD) which is a time-domain method that can be applied for the short-term assessment of heart rate variability, and thus the detection of arrhythmia [80]. If this method requires more development, it bring information not only on the heart frequency but on the regularity of the heartbeat.

In [96], the authors also highlight the feasibility of measuring the heart frequency by motion analysis performed on videos of the arteries in the alevin's tail. The motion areas are segmented and the optical flow is analyzed using the Färneback's algorithm [167]. Then, the analysis of speed variation in the blood flow allows to detect the heart contractions and to measure a heart frequency. This observation is relevant as we plan to investigate blood flow analysis from video sequences of the blood vessels of the embryo. Indeed, as illustrated in [81], blood vessels can be observed with high-resolution video to estimate the erythrocytes (red blood cells) velocity. In this study, a black and white CCD camera is used to record high-resolution videos of under-pigmented zebrafish mutants observed with an inverted microscope using infrared illumination for optimizing embryo settling. Videos frames are interlaced, meaning two fields of lines are generated for a same frame: a field displaying the odd lines acquired during 20ms, and a second field displaying the even lines, acquired during the following 20ms. As erythrocytes move between the two fields acquisition, the subtraction of both fields of a same frame leads to the generation of a shift vector. The vector length gives the distance travelled by the cells during the 20ms. Extracting the motion from blood vessels also shows the vessel contour. Thus, the vessel diameter can be measured by analyzing the intensity profile on the vessel cross-section. Combining velocity estimation and vessel diameter could bring a quantitative information on the blood flow.

Nevertheless, for blood vessels analysis, the spatial and temporal resolution must be sufficiently high to ensure motion smoothness, which requires the use of a high-resolution high-speed camera (In [96], the resolution is 2µm per pixel and the recorded frame rate is 100fps). The current experimental settings are not compatible with these requirements (about 12µm of resolution and a frame rate of 30fps) so additional steps must be added for such study. As many endpoints are already analyzed on day 9, involving a high experimentation time, additional steps cannot be performed this day. For this reason, we plan to record 10-second long videos of

6dpf embryos while still in egg form, observed under a motorized epifluorescence upright microscope (NIKON Eclipse Ni), and with a high-resolution camera recording 4 megapixels' images at 100 frame per second (Hamamatsu ORCA-Flash4.0 V3 Digital CMOS C13440-20CU). At this stage of development, the embryo is still in egg form, so easiest to manipulate as it does not move. We avoid the use of anesthesia that might disrupt the heart frequency. The embryo's yolk is still prominent and blood vessels are easily visible. Information about the heart frequency and its regularity might be extracted from the analysis of periodic changes in the intensity signal in the vessels [96]. Moreover, at 100fps, a frame is recorded every 10ms, which, according to [81], is sufficient to extract shift vectors and cells velocity estimation. Thus, an estimation of the blood flow might be obtained by combining the cells velocity and the diameter of blood vessels. However, we know that studies aiming to analyze the blood flow and to extract a heart frequency from blood vessels are usually performed on arteries, where especially veins are visible on the yolk at this stage of development. If expected parameters cannot be measured by this process, the analysis of the periodic intensity variation should be performed on the heart region to extract the heart frequency and detect arrhythmia as done in [80].

5.3.5. Behavioral assessment

Subtle abnormalities are not visible under a microscope. We may expect to detect some of them by another indirect way. We make the assumption that a developmental anomaly of the nervous system leads to an alteration of the embryo's behavior. Based on this postulate, several laboratories have addressed the issue of behavioral analysis [168, 169, 170]. Behavior analysis is based on the observation of the effect observed on embryo's behavior when subject to a light stimulus. It is mostly conducted on zebrafish. Two types are distinguished: Photo Motor Response (PMR) is measured before hatching, and Late Motor Response (LMR) is measured after hatching. On controls, PMR allows to detect the reflex movement provoked by the light change, while LMR allows to detect the decrease in the alevin activity when passing from a dark to a bright period. As most of embryos hatched the day of the assessment, the LMR measurement is the most adapted to our experimental setup. In our laboratory, an assessment test was developed that may reveal differences in 9dpf medaka alevins behavior that were exposed or not to teratogenic chemicals. Embryos are subject to a succession of bright and dark periods and parameters related to their motion, such as the duration spent in slow or fast motions, are automatically measured with a video acquisition system (Section 6.1.3 in the Appendix).

Comparing control alevins to alevins that were exposed to neuroactive chemicals revealed variations of some motion-related parameters that could be used as features for the behavioral assessment of medaka embryos. Nevertheless, a lot of data are recorded during this process and most of them are not interpreted yet. A supervised machine learning method could be tested to discriminate embryos that were exposed to neuroactive drugs or not, according to the presence of an effect on alevin locomotion. Behavior analysis should be performed on a set of embryos exposed to neuroactive drugs, and on controls, to measure parameters related to locomotion (the features). Data should be labeled according to the exposure condition. A model could be trained on these labeled data and tested on another set of unlabeled data.

5.4. Assessment of the teratogenicity test

An automated method has been developed for the assessment of chemical teratogenicity on medaka fish embryo. This method allows to assess the embryo mortality and to detect some of their malformations. To increase its sensitivity, the test may be completed by adding the analysis of new features representative of other malformations, of functional and behavioral alterations (Figure 61). The final validation of the automated teratogenicity assessment should assess the practical use of the automated method and its benefit compared to the manual method described in Section 1.1.3.2. We expect the teratogenicity assessment to detect and reject strong teratogenic chemicals while maximizing specificity. Thus, the performance of the automated and of the manual methods, meaning their ability to discriminate teratogenic from non-teratogenic chemicals, should be assessed and compared.

To do so, a list of reference chemicals, *i.e.*, chemicals whose the teratogenic effect is known on human, will be used. For each chemical, a concentration range will have to be considered, including a control plate. The efficacy of the manual method will be assessed as described in Section 1.1.3.2. For the automated method, all recorded images and videos will be considered, including the unusable images and videos, eggs and alevins, alive and dead embryos, malformed and healthy embryos. The automated mortality and malformations assessments will have to be applied on the whole set of images and videos, allowing to identify each embryo as being an egg or an alevin (or untreated if the data is unusable), and each alevin as being malformed or healthy. The classification between alive and dead embryos will be used to calculate a LC50, and the classification between malformed and healthy embryos will be used

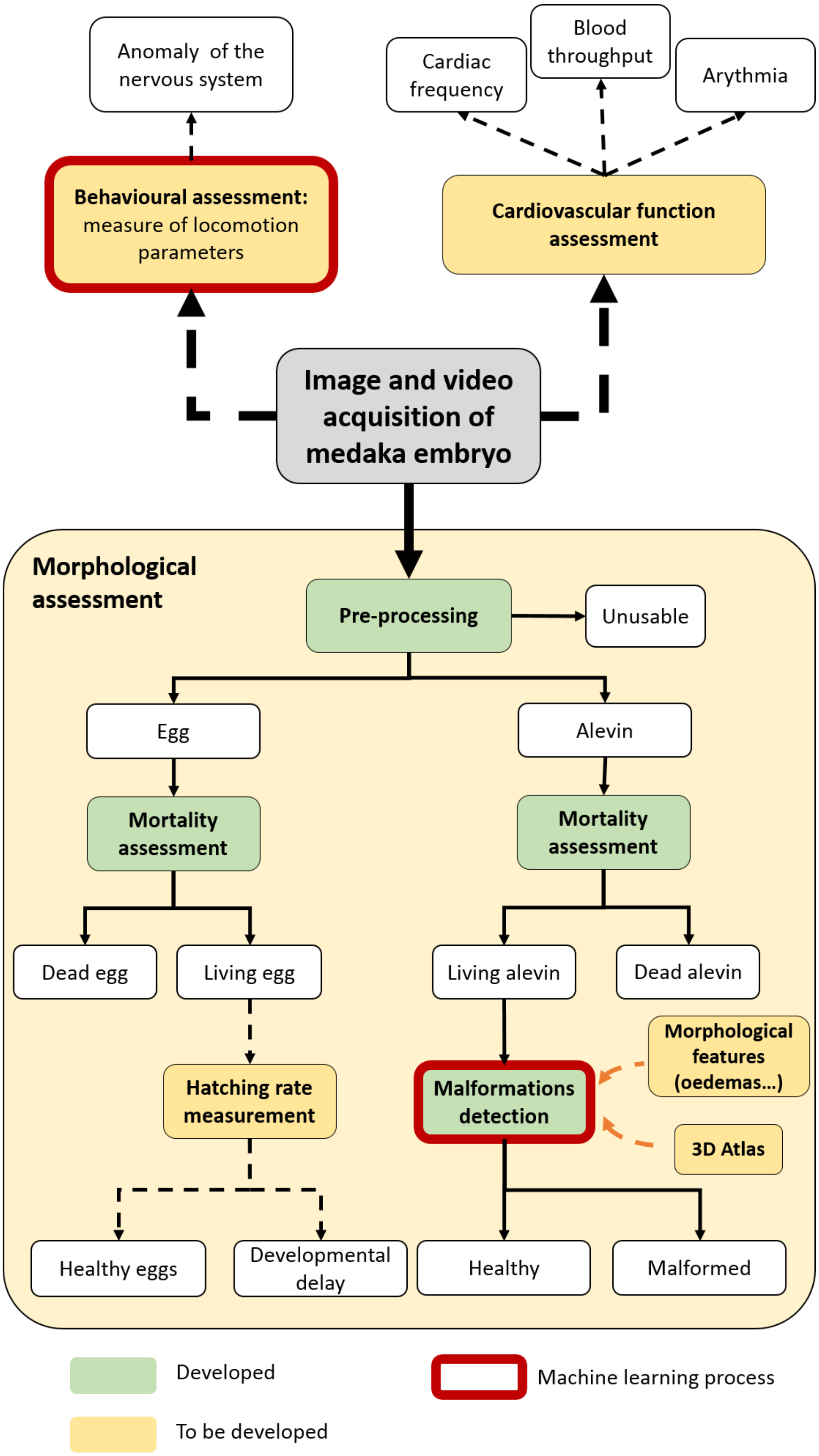


Figure 61. Flowchart of the final automated teratogenicity assessment assay. Each embryo should undergo behavioral assessment, cardiovascular assessment and morphological assessment to detect abnormalities and calculate a teratogenicity index, which will be used for the detection of teratogenic chemicals.

to calculate a EC50. The ratio between the two indices will give the teratogenicity index TI. A TI threshold should be determined that allows to discriminate teratogenic from non-teratogenic chemicals as well as possible. The performance of both manual and automated methods should be compared in terms of overall accuracy, sensitivity and specificity to assess the benefit of the automated method compared to the manual method.

6. Appendix

Here, complementary data used during this work are exposed. Section 6.1 details the material and methods used for embryo culture, chemical exposure, behavioral assessment and image acquisition for morphological assessment. Section 6.2 describes the development of a personalized sample rack for optimization of the image acquisition conditions.

6.1. Material and methods

6.1.1. Embryo culture

Medaka eggs are bought from Amagen (UMS 3504 CNRS / UMS 1364 INRA). They are incubated in culture medium composed of:

- 17.1mM NaCl,
- 4.02mM KCl,
- 3.6mM CaCl2,
- 3.3mM MgSO4,

dissolved in osmosis water. Methylene blue is added to the culture medium (about 4mg/L) for its antibacterial property. When medaka eggs are received at the laboratory, they are observed under a stereomicroscope (model Leica MZ 12 5 with objective Plan APO 0.63 ×) to eliminate dead eggs. Embryos are incubated for 9 days at a temperature of 27°C and subject to a light cycle composed of 14 hours of light and 10 hours of obscurity. The incubation medium is replaced every 2 days during these nine days using a Hamilton MICROLAB STAR automated device.

6.1.2. Chemical exposure

At day one, alive eggs at *blastula* stage (Section 1.1.2.2) are manually placed into a 24-well plate, one egg per well, and incubated in 2mL of culture medium containing increasing concentrations of the chemical of interest. Five concentrations are tested for each chemical, plus a zero concentration for control. One plate is prepared for each concentration, for a total of six plates per chemical. Chemicals that were used for this study are listed in Table 11.

Chemical name	**CAS number**
Amantadine hydrochloride	665-66-7
Amaranth	915-67-3
6-Aminonicotinamide	329-89-5
Azelaic acid	123-99-9
Caffeine	58-08-2
Captopril	62571-86-2
Colchicine	64-86-8
Cromolyn sodium	15826-37-6
Cyclopamine	4449-51-8
Cyclophosphamide	6055-19-2
Dimethyl Sulfoxide	67-68-5
Ethylene-d_4 thiourea	106-18
5-Fluorouracil	51-21-8
Hydroxyurea	127-07-1
Hydroxyzine dihydrochloride	2192-20-3
Lactitol	585-86-4
L-Ascorbic Acid	50-81-7
Lithium chloride monohydrate	231-212-3
Metoprolol	5692-17-7
Sodium cyclamate	139-05-9
Urethane	51-79-6

Table 11. List of chemicals used during this project.

6.1.3. Behavioral assessment

After a 9-day exposure, the behavior of medaka embryos is assessed using the 4th generation model of Zebrabox video tracking system (Viewpoint, Lyon, France). The 24-well plates containing embryos are individually placed into the Zebrabox. The acquisition system is illustrated in Figure 62 and is composed of:

- an infrared camera that acquires grey levels images even during dark periods, by receiving the light produced by an infrared LED;
- a filter in front of the camera allows to avoid the captor saturation during light periods;
- two mirrors that allow to lengthen the optical path and to record a video of the whole plate;
- a diffusing light support to homogeneously illuminate the plate.

When placing in the system, the embryos are left 10 minutes in light to limit the impact of the environment change on embryos behavior. The embryos are then subjected to a light stimulus which is a succession of 30 seconds of light periods and 30 seconds of dark periods. The embryos movement is recorded during 6 minutes. The recorded video is simultaneously treated by the image processing software Zebralab for the real time tracking of the embryos in their respective well. The tracking appears in real time in red on the video. In particular, three different movements are distinguished which are fast movement, slow movement and absence of movement (static).

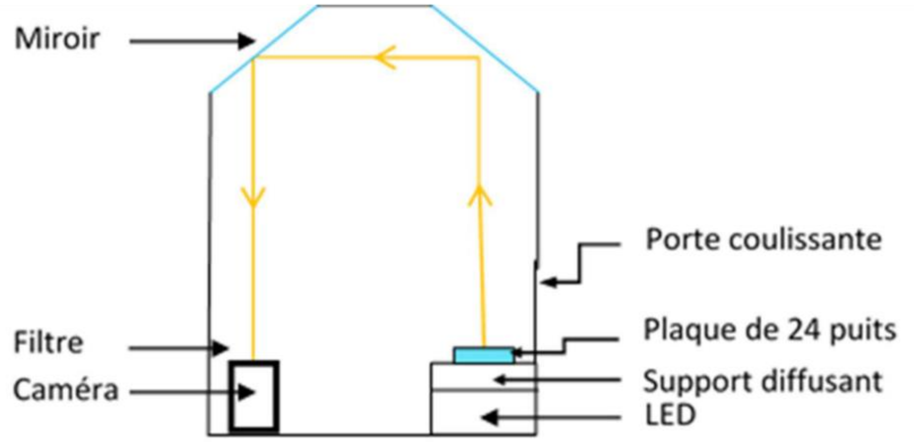


Figure 62. Zebrabox video acquisition system for alevin behavioral analysis.

Eleven parameters are analyzed during the behavioral assessment:

- *largdist ON*, *largdist OFF* and *largdist Tot* are the distances in fast movement spent by the embryo during light period, during dark period and the total respectively;

- *smldist ON*, *smldist OFF* and *smldist Tot* are the distances in slow movement spent by the embryo during light period, during dark period and the total respectively;

- The total travelled distance *totdist* (sum of largdist Tot and smldist Tot);

- *largdur*, *smldur* and *inadur* are the duration spent by the embryo in fast movement, the duration in slow movement and the duration being static respectively;

- *inact* is the number of times the embryo is static.

6.1.4. Image acquisition for morphological assessment

The ninth day of incubation, 1.5mL of incubation medium is removed from each well before anesthetizing fish embryos with 70µL of tricaine (final concentration of 0,18 g/L in a total volume of 0.57mL). The plate is then placed under an acquisition platform composed of a light platform, for underneath illumination of the plate, and a moving monochrome camera (objective Nikkon AF MICRO NIKKOR 60mm 1:2:8D Kipon NIX-C). Embryos are manually centered on the well. Then, image and video recording is automatically performed under the control of a Visilog Visual Basic script. For each well, we record a 8-bit image of the whole well of size 1500×1500 pixels, and a 8-bit video sequence at 30frames per second with the same dimensions over a duration of 1 second. The image resolution is about 12µm. Platform control and image data acquisition were performed using FEI Visilog 7.

6.2. Optimization of image acquisition: development of a personalized sample rack

6.2.1. Context

The teratogenicity assessment assay described in this thesis analyzes medaka embryos placed in flat bottom 24-well plates. Before image acquisition, embryos are anesthetized and finally fall at the bottom of the well. Most of time, the embryo touches the wall of the well. As explained in Section 2.2, such embryo appears partially occulted or deformed because of some border effects. Indeed, a healthy embryo may take the curved shape of the wall and be wrongly detected as malformed. Such artefacts can cause important bias for image analysis. The program developed allows to recognize and to reject such image. To avoid the rejection of too many images, each embryo is manually centered in its well before acquisition. This process is tedious and time-consuming. To optimize this step, we developed a rounded bottom 24-well plate. In such well, the anesthetized alevin falls at the center of the well without manual intervention.

6.2.2. Objective and constraints

Because of the different automates used during the test, the development of such device asks to meet shape and weight requirements (Figure 63a).

To be compatible with the image and video acquisition platform, the device must be transparent so that light can cross the wall and spread inside the wells. The rounded bottom of the wells must be perfectly smooth as any defect or shadow on the curved wall would be visible on the acquired image. In particular, conical bottom are forbidden as the angle at the center would be visible and superimposed to the embryo on the acquired image.

For incubation process, a lid is needed to limit medium evaporation. This lid must allow gas exchanges. Adsorption of the chemical’s molecule on the plate walls also has to be as limited as possible. Two solutions are the use of polypropylene single use plates (low adsorption), or the use of glass plates reusable after a cleaning process.

A first prospection revealed that there is no plate that corresponds to all these criteria in the market.

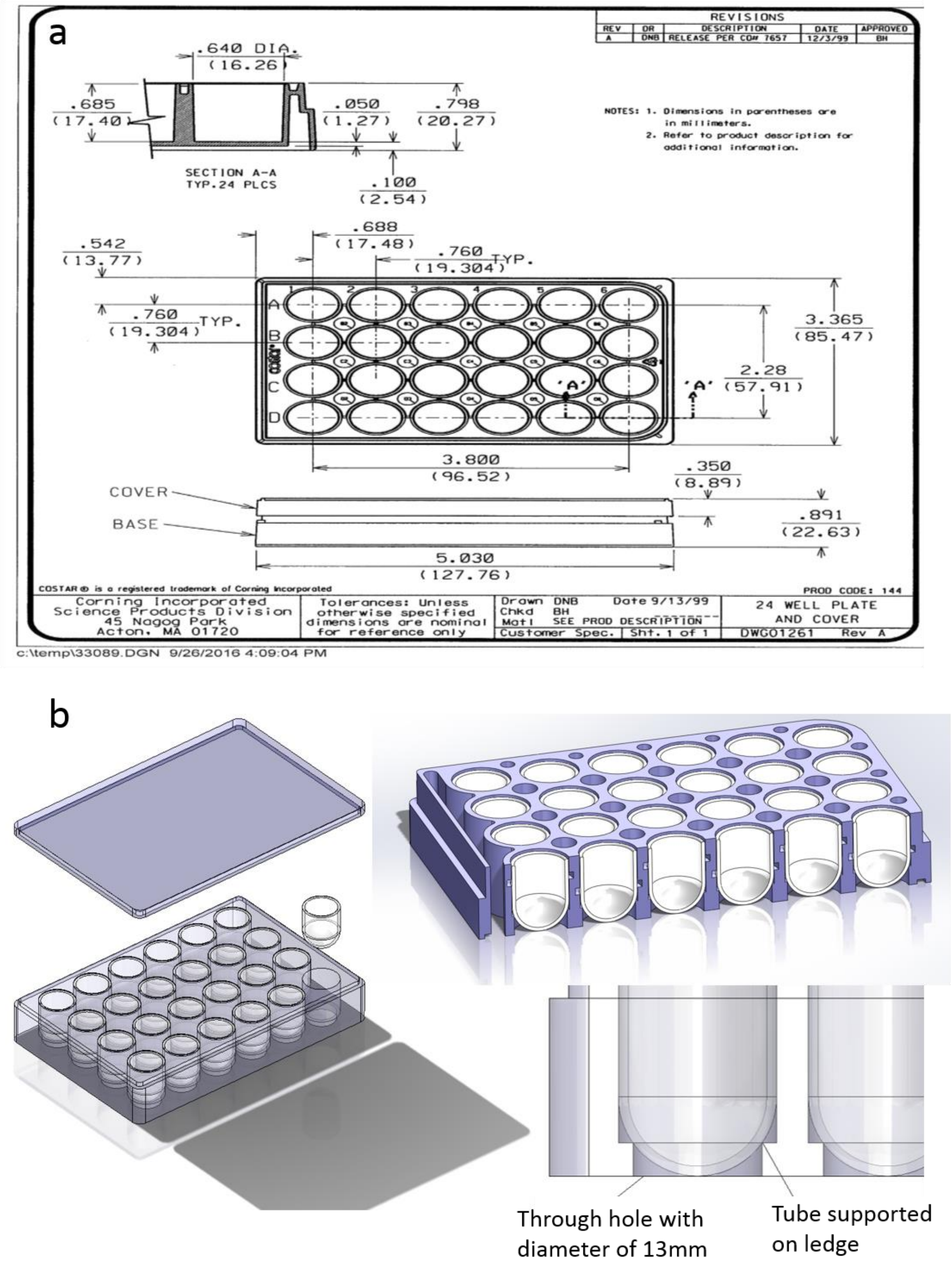

a
REVISIONS
REV DR DESCRIPTION DATE APPROVED
A DNB RELEASE PER CO# 7657 12/3/99 BH
NOTES: 1. Dimensions in parentheses are in millimeters.
2. Refer to product description for additional information.
.640 DIA. (16.26)
.685 (17.40)
.050 (1.27)
.798 (20.27)
SECTION A-A TYP.24 PLCS
.100 (2.54)
.542 (13.77)
.688 (17.48)
.760 (19.304) TYP.
.760 (19.304) TYP.
1 2 3 4 5 6
A B C D
'A' 'A'
3.365 (85.47)
2.28 (57.91)
3.800 (96.52)
.350 (8.89)
COVER
BASE
.891 (22.63)
5.030 (127.76)
COSTAR® is a registered trademark of Corning Incorporated
PROD CODE: 144
Corning Incorporated Science Products Division 45 Nagog Park Acton, MA 01720
Tolerances: Unless otherwise specified dimensions are nominal for reference only
Drawn DNB Date 9/13/99
Chkd BH
Mat'l SEE PROD DESCRIPTION
Customer Spec. Sht. 1 of 1
24 WELL PLATE AND COVER
DWG01261 Rev A
c:\temp\33089.DGN 9/26/2016 4:09:04 PM
b
Through hole with diameter of 13mm
Tube supported on ledge

Figure 63. Conception of the rounded bottom 24-well plates. a: standard 24-well plate COSTAR® used as model for shape requirements. b: design of the device by a computer-aided design software.

6.2.3. Description of the device

To answer to all technical constraints, we decided to design and develop a sample rack specifically adapted to our needs.

Several prototypes of the sample rack have been conceived (Figure 63b).

- The first version of the prototype V1 is made in polymethyl methacrylate (PMMA), has 24 holes with ledges, adapted to the inclusion of test tubes. Horizontal grooves were carved on both sides to ensure the grip during the rack manipulation by the Hamilton MICROLAB Star device (Figure 64a and Figure 65a).
- A second version of the prototype V2 was designed with holes in between the wells for weight optimization and proper detection of alevins during the behavioral analysis as explained in Section 6.2.4.1 (Figure 65b).
- A third version V3 limits the well movement with silicone ring-shaped seals. It was designed for optimization of image acquisition described in Section 6.2.4.2 (Figure 65c).

In the market, it is feasible to find test tubes with the expected width and thickness, but not with the expected length of 2cm. Two solutions are possible: either truncate pre-existing test tubes to the expected length or to make them up from scratch by the glass maker (involves a mold creation). For the first test, sodocalcic glass test tubes were provided by Dutscher society and truncated by V.E.R.A.L. glass-maker society (Figure 64b). After the prototype validation, the creation of a tube mold could be planned depending on the prices.

For rack covering, standard covers COSTAR® Universal square lid are supplied by Corning society (Figure 64a).

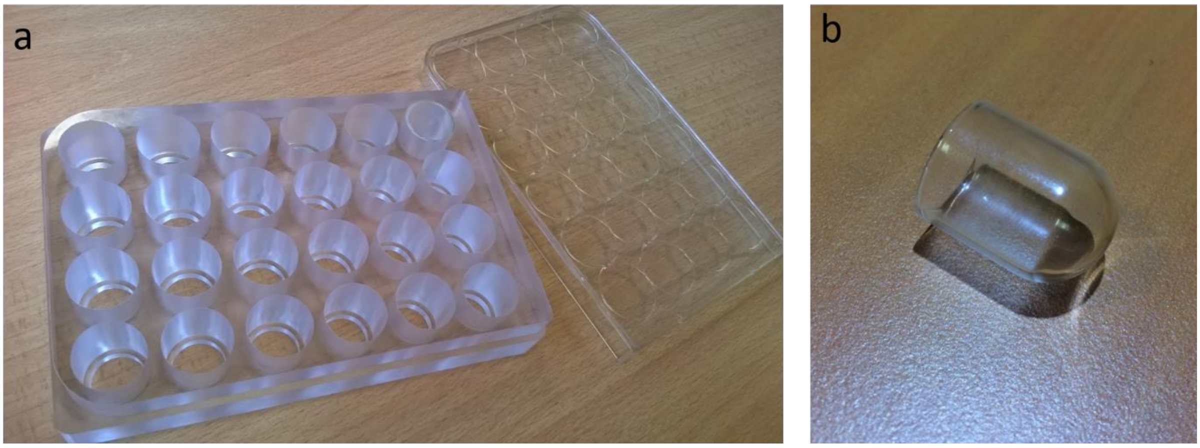


Figure 64. First version of the sample rack developed. a: sample rack with the standard COSTAR® cover. b: example of a truncated test tube in sodocalcic glass.

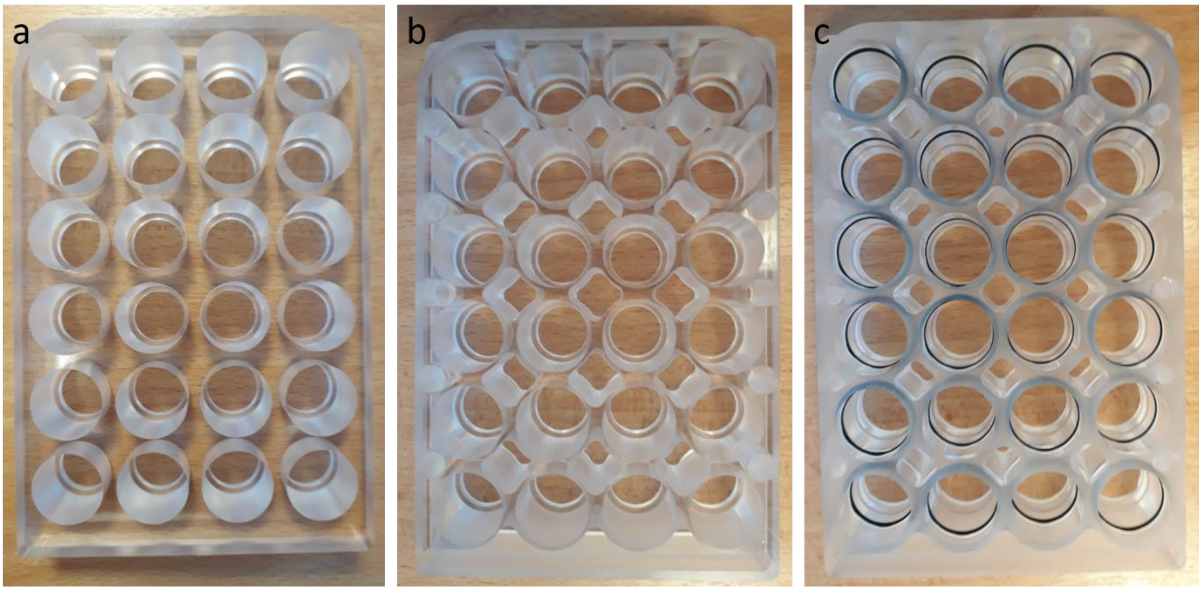


Figure 65. Three versions of the developed sample rack. a: the sample rack V1. b: the drilled version of the sample rack V2. c: the sample rack V3 with silicone ring-shaped seals.

6.2.4. Prototypes validation in experimental conditions

6.2.4.1. *Alevins behavioral analysis*

When assessing the behavior using the prototype V1, the curved wall of the tubes causes the deviation of the light waves between light table, the air and the glass of the tubes (refraction phenomenon illustrated in Figure 67). The incident ray arrives on the medium boundary with an angle of incidence a_i, and is deflected into a refracted ray with an angle of refraction a_r. This phenomenon leads to a thick dark circle all around the well borders, as fewer light rays are received by the camera in this region (Figure 66b), compared to the image obtained by using a standard 24-well plate (Figure 66a). The pixel intensity in this circle is similar to the alevin intensity. As the alevin tracking is based on the detection of the variation of pixel intensity in the well, the software is unable to detect the alevin in the dark circle. Moreover, the presence of the dark circle causes a strong contrast in the circle border, which is very sensitive to vibrations due to external environment. These vibrations cause a high pixel intensity variation in the circle frontier, that is interpreted by the program as a movement of the alevin.

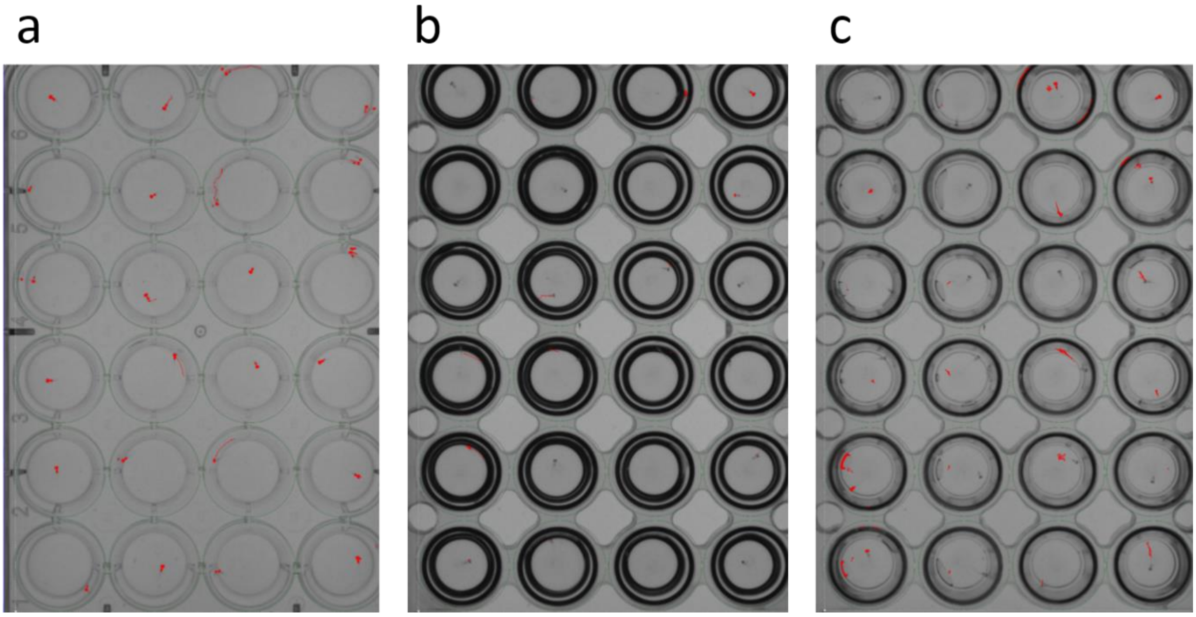


Figure 66. Results of the acquisition during behavioral assessment. a: classical 24-well plate with flat bottom. b: sample rack V2. c: sample rack V2 immerged in water. We notice that the dark circles due to light refraction on the curved wall of the wells are reduced when the sample rack is immerged.

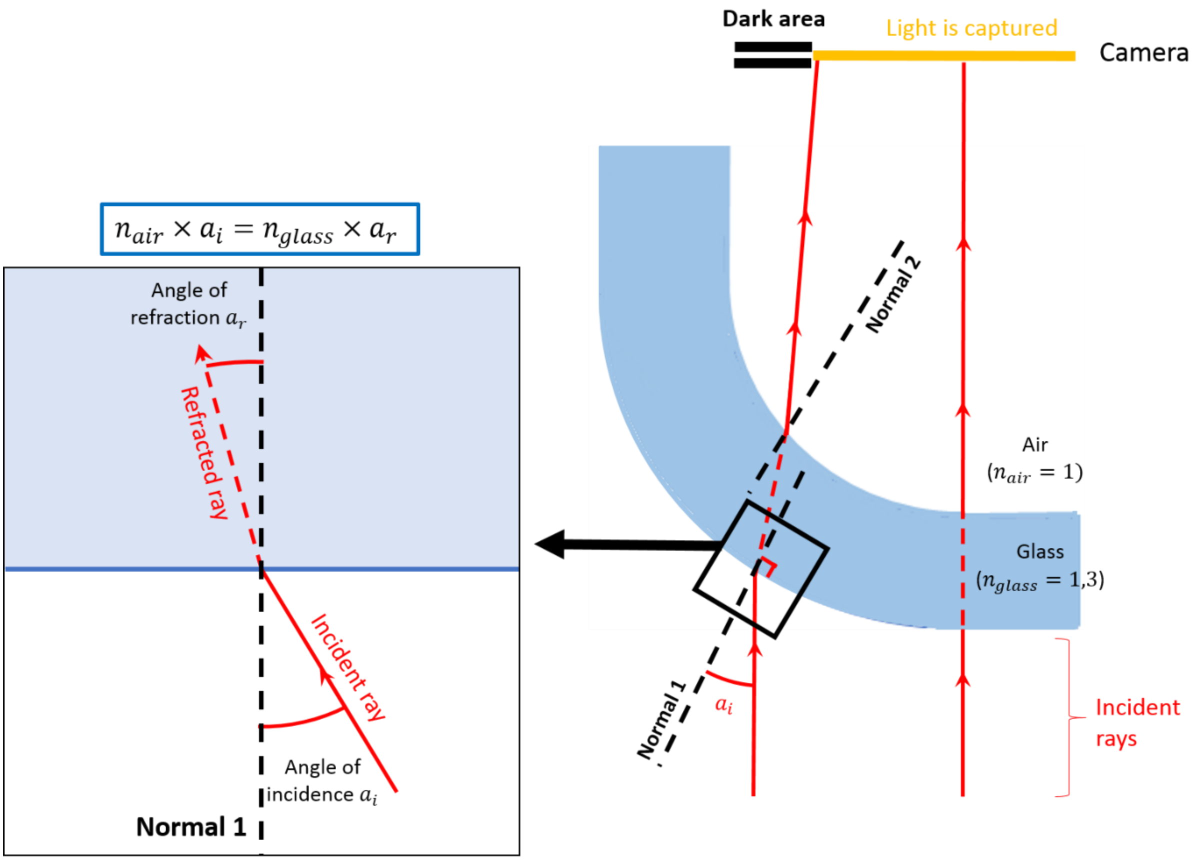


Figure 67. Refraction of light on the glass tubes. Upright incident rays come from the light platform located below the plate. At the center of the tube, rays are perpendicular to the glass. Thus, they are not deviated. On the curve, rays are not perpendicular. They are deviated twice: at the first interface between air and glass and at the second interface between glass and air. This results in an area where no ray is captured by the camera.

To prevent the appearance of this circle, we decided to immerge the device in water to increase the value of the refracted angle a_r, and thus to partially realign the incident and the refracted rays. To do so, the second version of the rack was designed. This rack V2 was drilled to allow the immersion of wells into the water, avoiding air bubbles to be stuck between the glass wall of the wells and the rack itself (Figure 66c).

A behavioral analysis was performed with this device on alevins exposed to increasing concentrations of an anesthetic (tricaine). The effect was revealed y the resulting dose-response curve, which allows to validate the use of the device in experimental conditions.

6.2.4.2. *Alevins morphological analysis*

When images are recorded with the acquisition platform described in Section 6.1.4, and with the sample rack V2, all alevins are perfectly centered in the sample rack. Nevertheless, all truncated test tubes, used as wells, present an imprint in their center. This imprint is visible on the acquired image and is systematically superimposed to the centered alevins (Figure 68). As the image properties of the imprint and the alevin are similar, the image processing program is not able to distinguish them.

To remove the imprint from the well bottom, we decided to acquire two images of each well: one with the alevin inside and one without. The aim is to subtract the image of the empty well from the image with the alevin.

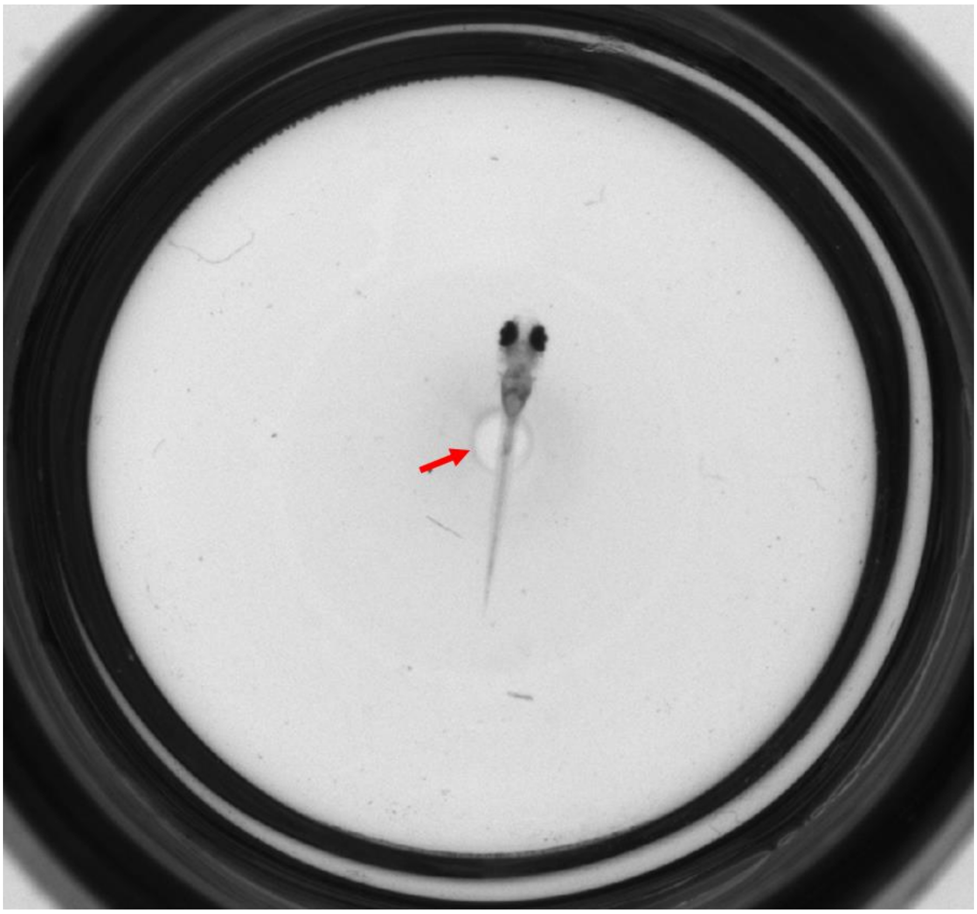


Figure 68. Example of image acquired for the morphological assessment with sample rack V2. The alevin appears centered in the well without any manual intervention. The red arrow shows the imprint present at the bottom of the well and that appears superimposed with the alevin.

To ensure the two images are perfectly aligned before the subtraction, two different methods were tested:

- the physical method which consists in adding silicone ring-shaped seals inside the rack holes so that the truncated test tubes are maintained unmoving inside their rack holes (sample rack V3 shown in Figure 65c). Tests revealed that the seals are not sufficient to ensure the wells alignment in the two successively acquired images, so this method was not retained;

- the computerized method which consists of image registration.

To perform image registration, points of reference are required to make the correspondence between the two images we want to register. Thus, glass wells were marked by pen. Then, two images of the sample rack V2 were successively acquired with and without alevins. The image with the alevin is denoted I^{alevin} and the image of the empty well is denoted I^{empty}. Marks are segmented in the two images, leading to the two binary masks M^{alevin} and M^{empty}. The transformation $T_{M^{alevin} \rightarrow M^{empty}}$ required to pass from the mask M^{alevin} to the mask M^{empty}. This transformation is then applied to I^{alevin}, leading to the registered image I_{reg}^{alevin} that can be subtracted to I^{empty}. The result is denoted I^f. This process is illustrated in Figure 69.

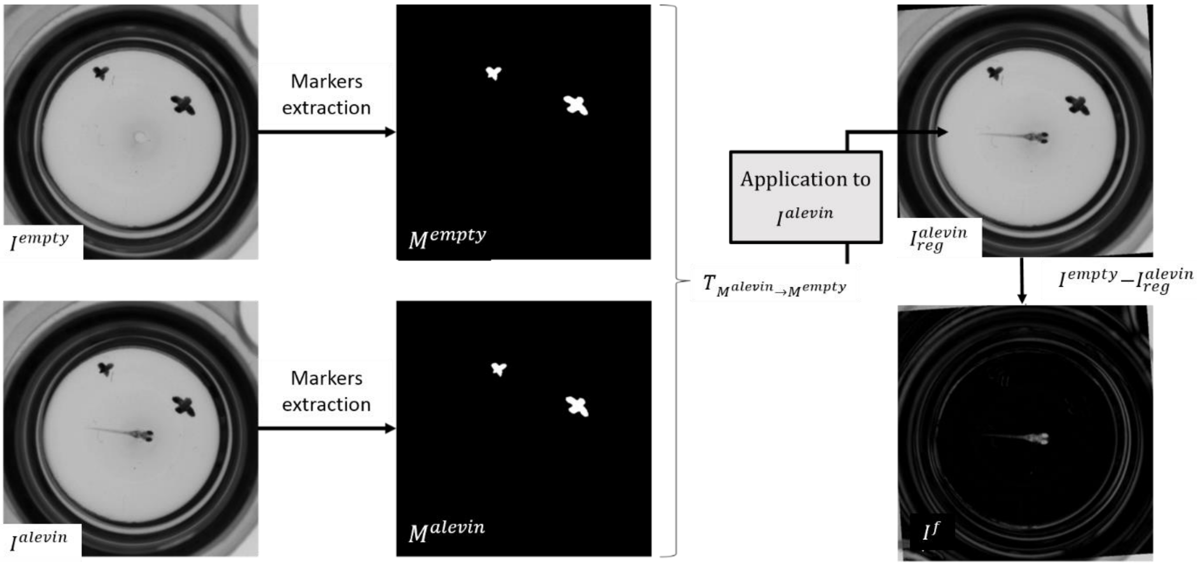


Figure 69. Removing process of the well imprint by image registration and image subtraction.

Several steps still need to be performed. The process must be completed to obtain a proper image of the alevin without the imprint. Then, the analysis program must be tested on the resulting image to assess the impact of the imprint removing on the alevin segmentation and classification efficacy. Finally, the process will have to be performed on a more important number of images (about 30 images at least) for to validate the method.

6.2.5. Conclusion and perspectives

Several versions of a sample rack have been conceived and tested in different experimental conditions corresponding to the steps of the teratogenicity test. Currently, the use of the sample rack V2 was validated for the behavioral assessment and still need to be validated for the morphological assessment. The perspective at medium term is to test if image subtraction allows the correct segmentation and analysis of the alevin. If so, this step could be added as a pre-processing step to a future version of the software. At long term, the behavior and morphological analysis will have to be performed in a significant number of data by testing reference chemicals (positive and negative).

Bibliography

[1] E. Barbeau, “Méthode alternative à l’expérimentation animale pour l’identification de substances chimiques altérant le développement embryonnaire,” 2015.

[2] N. Vargesson, “Thalidomide-Induced Teratogenesis: History and Mechanisms,” *Birth Defects Research (Part C)*, pp. 140–156, 2015.

[3] W.M.S. Russell and R.L. Burch, *The Principles of Humane Experimental Technique*. 1959.

[4] M. Balls, “The Three Rs and the Humanity Criterion. An Abridged Version of The Principles of Humane Experimental Technique by WMS Russell and RL Burch.” Fund for the Replacement of Animals in Medical Experiment (FRAME)., 2009.

[5] “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.

[6] “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.

[7] 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.

[8] 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.

[9] E.K. Balon, “Alternative ways to become a juvenile or a definitive phenotype (and on some persisting linguistic offenses),” *Environmental Biology of Fishes*, pp. 17–38, 1999.

[10] 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.

[11] M. Halder *et al.*, “Regulatory Aspects on the Use of Fish Embryos in Environmental Toxicology,” *Integr. Environ. Assess. Manag.*, vol. 6, no. 3, 2010.

[12] 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.

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

[14] A.R. Cossins and D.L. Crawford, “Fish as models for environmental genomics,” *Nat. Publising Group*, vol. 6, pp. 324–332, 2005.

[15] K. Dooley and L.I. Zon, “Zebrafish: a model system for the study of human disease,” *Current Opinion in Genetics & Development*, pp. 252–256, 2000.

[16] L. Gunnarsson, A. Jauhiainen, E. Kristiansson, O. Nerman, and D.G. Larsson, “Evolutionary conservation of human drug targets in organisms used for environmental risk assessments,” *Environ. Sci. Technol.*, no. 42, pp. 5807–5813, 2008.

[17] K.C. Brannen, J.M. Panzica-Kelly, T.L. Danberry, and K.A. Augustine-Rauch, “Development of a Zebrafish Embryo Teratogenicity Assay and Quantitative Prediction Model,” *Birth Defects Res. Part B*, vol. 89, pp. 66–77, 2010.

[18] J.S. Ball *et al.*, “Fishing for teratogens: a consortium effort for a harmonized zebrafish developmental toxicology assay,” *Toxicological Sciences*, pp. 210–219, 2014.

[19] E. Krupp, “Screening of developmental toxicity – Validation and predictivity of the zebrafish embryotoxicity assay (ZETA) and strategies to optimize de-risking developmental toxicity of drug candidates,” *Toxicology Letters*, p. 39, 2016.

[20] S.A.B. Hermsen, E-J. van den Brandhof, L.T.M. van der Ven, and A.H. Piersma, “Relative embryotoxicity of two classes of chemicals in a modified zebrafish embryotoxicity test and comparison with their in vivo potencies,” *Toxicology in vitro*, pp. 745–753, 2011.

[21] 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.

[22] 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.

[23] T. Iwamatsu, "Stages of normal development in the medaka Oryzias latipes," *Mechanisms of Development*, pp. 605–618, 2004.

[24] J.M. Spitsbergen and M.L. Kent, "The State of the Art of the Zebrafish Model for Toxicology and Toxicologic Pathology Research—Advantages and Current Limitations," *Toxicol. Pathol.*, vol. 31, pp. 62–87, 2003.

[25] P. Kamavisdar, S. Saluja, and S. Agrawal, "A Survey on Image Classification Approaches and Techniques," *Int. J. Adv. Res. Comput. Commun. Eng.*, vol. 2, pp. 1005–1009, 2013.

[26] R. Maree, P. Geurts, J. Piater, and L. Wehenkel, "Biomedical Image Classification with Random Subwindows and Decision Trees," presented at the Computer Vision for Biomedical Image Applications CVBIA, 2005, pp. 220–229.

[27] T.S. Korting, L.M. Garcia Fonsec, E.F. Castejon, and L.M. Namikawa, "Improvements in Sample Selection Methods for Image Classification," *Remote Sens.*, vol. 6, pp. 7580–7591, 2014.

[28] M.R. Anderberg, *Cluster Analysis for Applications*, Academic Press. 1973.

[29] E.W. Forgy, "Cluster analysis of multivariate data: efficiency versus interpretability of classifications," *Biometrics*, vol. 21, pp. 768–769, 1965.

[30] J. Cousty, L. Najman, Y. Kenmochi, and S. Guimaraes, "Hierarchical segmentations with graphs, quasi-flat zones, minimum spanning tree and saliency maps," *Journal of Mathematical Imaging and Vision*, pp. 479–502, 2018.

[31] X. Zhu and A.B. Goldberg, "Introduction to Semi-Supervised Learning," *Synthesis Lectures on Artificial Intelligence and Machine Learning*, pp. 1–130, 2009.

[32] A. L. Tarca, V. J. Carey, X-w. Chen, R. Romero, and S. Drăghici, "Machine learning and its applications to biology," *PLoS Comput. Biol.*, 2007.

[33] C. Sommer and D.W. Gerlich, "Machine learning in cell biology – teaching computers to recognize phenotypes," *Journal of Cell Science*, pp. 1–11, 2013.

[34] N.B. Gunter *et al.*, "Automated detection of imaging features of disproportionately enlarged subarachnoid space hydrocephalus using machine learning methods," *NeuroImage: Clinical*, 2018.

[35] T.M. Deist *et al.*, "Machine learning algorithms for outcome prediction in (chemo)radiotherapy: An empirical comparison of classifiers," *Medical Physics*, pp. 3449–3459, 2018.

[36] K. Kourou, T.P. Exarchos, and K.P. Exarchos, "Machine learning applications in cancer prognosis and prediction," *Computational and Structural Biotechnology Jounal*, pp. 8–17, 2015.

[37] I. Goodfellow, Y. Bengio, and A. Courville, *Deep Learning*, MIT Press. 2016.

[38] T. Hastie, R. Tibshirani, and J.H. Friedman, *The elements of statistical learning: data mining, inference, and prediction*, Springer. 2009.

[39] R. Liu *et al.*, "Automated Phenotype Recognition for Zebrafish Embryo Based In Vivo High Throughput Toxicity Screening of Engineered Nano-Materials," *Plos One*, vol. 7, no. 4, pp. 1–10, 2012.

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

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

[42] B.E. Boser, I.M. Guyon, and V.N. Vapnik, "A training algorithm for optimal margin classifiers," in *Proceedings of the fifth annual workshop on Computational learning theory*, 1992, pp. 144–152.

[43] C. Cortes and V. Vapnik, "Support-vector networks," *Mach. Learn.*, vol. 20, pp. 273– 297, 1995.

[44] V. N. Vapnik and V. Vapnik, *Statistical learning theory*, Wiley., vol. 2. 1998.

[45] O. Chapelle, P. Haffner, and V. N. Vapnik, "Support Vector Machines for Histogram-Based Image Classification," *IEEE Transactions ON Neural Networks*, pp. 1055–1064, 1999.

[46] O. Stern *et al.*, "Automatic localization of interest points in zebrafish images with tree-based methods," in *6th IAPR International Conference (PRIB), Proceedings*, Delft, Netherlands, 2011.

[47] L. Breiman, "Bagging Predictors," *Mach. Learn.*, vol. 24, pp. 123–140, 1996.

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

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

[50] N. Otsu, “A threshold selection method from gray-level histograms,” *Automatica*, vol. 11, no. 285–296, pp. 23–27, 1975.

[51] C. Tomasi and R. Manduchi, “Bilateral filtering for gray and color images,” in *IEEE International Conference on Computer Vision*, Bombay, India, 1998, pp. 839–846.

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

[53] J. Serra, *Image Analysis and Mathematical Morphology - Volume II : Theoretical Advances*. Academic Press, London, 1988.

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

[55] P. Soille, *Morphological Image Analysis, principles and applications*. Springer, 1999.

[56] L. Vincent, “Morphological area openings and closings for grey-scale images,” *Shape in Picture*, pp. 197–208, 1994.

[57] P. Soille and H. Talbot, “Directional morphological filtering,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, pp. 1313–1329, 2001.

[58] S. Beucher and C. Lantuéjoul, “Use of watersheds in Contour Detection,” in *Int. Workshop on Image Processing*, Rennes, France, 1979.

[59] J. Cousty, G. Bertrand, L. Najman, and M. Couprie, “Watershed cuts: minimum spanning forests and the drop of water principle,” *IEEE transactions on Pattern Analysis and Machine Intelligence*, pp. 1362–1374, 2009.

[60] J. Cousty and L. Najman, “Incremental algorithm for hierarchical minimum spanning forests and saliency of watershed cuts,” in *Proceedings of the 10th International Symposium on Mathematical Morphology (ISMM)*, Verbania-Intra, Italy, 2011, pp. 272–283.

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

[62] S. Chen, M. Zhao, G Wu, C. Yao, and J. Zhang, “Recent Advances in Morphological Cell Image Analysis,” *Computational and Mathematical Methods in Medicine*, pp. 1–10, 2012.

[63] Y. Al-Kofahi, W. Lassoued, W. Lee, and B. Roysam, “Improved Automatic Detection and Segmentation of Cell Nuclei in Histopathology Images,” *IEEE Transactions on Biomedical Engineering*, pp. 841–852, 2010.

[64] J. Rittscher, “Characterization of Biological Processes through Automated Image Analysis,” *Annual Review of Biomedical Engineering*, pp. 315–344, 2010.

[65] F. Jean, A.-C. Roudot, and D. Parent-Massin, “An automatic method for the evaluation of xenobiotic toxicity on haematopoietic progenitors,” *Computer Methods and Programs in Biomedicine*, pp. 1–8, 2000.

[66] A.Benzinou, Y. Hojeij, and A.C. Roudot, “Digital image analysis of haematopoietic clusters,” *Computer Methods and Programs in Biomedicine*, pp. 121–127, 2005.

[67] A. Benzinou, Y. Hojeij, and A.C. Roudot, “Automatic Cellular Aggregates Quantification for Toxicology Using Statistical Learning,” in *2nd International Conference on Information & Communication Technologies*, 2006, pp. 1557–1561.

[68] M.R. Lamprecht, D.M. Sabatini, and A.E. Carpenter, “CellProfiler^{TM}: free, versatile software for automated biological image analysis,” *BioTechniques*, pp. 71–75, 2018.

[69] I.W.T. Selderslaghs, A.R. Van Rompay, W. De Coenb, and H.E. Witters, “Development of a screening assay to identify teratogenic and embryotoxic chemicals using the zebrafish embryo,” *Reproductive Toxicology*, pp. 308–320, 2009.

[70] 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.

[71] 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.

[72] 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.

[73] R. Peravali *et al.*, “Automated feature detection and imaging for high-resolution screening of zebrafish embryos,” *BioTechniques*, vol. 50, no. 5, 2011.

[74] 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.

[75] S. Maanasi Kalasekar *et al.*, “Identification of environmental chemicals that induce yolk malabsorption in zebrafish using automated image segmentationSharanya,” *Reproductive Toxicology*, pp. 20–29, 2015.

[76] J.H. Westhoff *et al.*, “Development of an Automated Imaging Pipeline for the Analysis of the Zebrafish Larval Kidney,” *PLOS ONE*, pp. 1–13, 2013.

[77] C. Hans, C.W. McCollum, M.B. Bondesson, J-A. Gustafsson, S.K. Shah, and F.A. Merchant, “Automated Analysis of Zebrafish Images for Screening Toxicants,” in *35th Annual International Conference of the IEEE EMBS*, Osaka, Japan, 2013, pp. 3004–3007.

[78] K.L Yozzo, G.M. Isales, T.D. Raftery, and D.C. Volz, “High-Content Screening Assay for Identification of Chemicals Impacting Cardiovascular Function in Zebrafish Embryos,” *Environmental science and technology*, p. 11302−11310, 2013.

[79] A. Vogt *et al.*, “Automated image-based phenotypic analysis in zebrafish embryos,” *PMC*, pp. 656–663, 2010.

[80] Pylatiuk C. *et al.*, “Automatic Zebrafish Heartbeat Detection and Analysis for Zebrafish Embryos,” *Zebrafish*, pp. 379–383, 2014.

[81] T. Schwerte and B. Pelster, “Digital motion analysis as a tool for analysing the shape and performance of the circulatory system in transparent animals,” *The Journal of Experimental Biology*, pp. 1659–1669, 2000.

[82] C.G Burns, D.J. Milan, E.J. Grande, W. Rottbauer, C.A. MacRae, and M.C. Fishman, “High-throughput assay for small molecules that modulate zebrafish embryonic heart rate,” *Nat. Chem. Biol.*, vol. 1, pp. 263–2364, 2011.

[83] J.K. Leet *et al.*, “High-Content Screening in Zebrafish Embryos Identifies Butafenacil as a Potent Inducer of Anemia,” *Plos One*, pp. 1–10, 2014.

[84] A. Letamendia *et al.*, “Development and Validation of an Automated High- Throughput System for Zebrafish In Vivo Screenings,” *Plos One*, vol. 7, pp. 1–13, 2012.

[85] R. Alshut *et al.*, “Methods for Automated High-Throughput Toxicity Testing Using Zebrafish Embryos,” in *Lecture Notes in Computer Science*, Springer Berlin Heidelberg, 2010, pp. 219–226.

[86] D. Arslanova, T. Yang, X. Xu, S.T. Wong, C.E. Augelli-Szafran, and W. Xia, “Phenotypic analysis of images of zebrafish treated with Alzheimer’s g-secretase inhibitors,” *BMC Biotechnology*, pp. 10–24, 2010.

[87] 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.

[88] S. Deal *et al.*, “Development of a quantitative morphological assessment of toxicant-treated zebrafish larvae using brightfield imaging and high-content analysis,” *Journal of Applied Toxicology*, 2015.

[89] W.K. Martin *et al.*, “High-Throughput Video Processing of Heart Rate Responses in Multiple Wild-type Embryonic Zebrafish per Imaging Field,” *Scientific Reports*, pp. 1–14, 2018.

[90] C. Pardo-Martin, A. Allalou, J. Medina, P.M. Eimon, C. Wählby, and M. Fatih Yanik, “High-throughput hyperdimensional vertebrate,” *Nature Communications*, 2013.

[91] C. Pardo-Martin, T-Y. Chang, B.K. Koo, C.L. Gilleland, S.C. Wasserman, and M.F. Yanik, “High-throughput in vivo vertebrate screening,” *Nat. Methods*, pp. 634–636, 2011.

[92] E. Teixido, T.R. Kießling, E. Krupp, C. Quevedo, A. Murian, and S. Scholz, “Automated Morphological Feature Assessment for Zebrafish Embryo Developmental Toxicity Screens,” *Toxicological Sciences*, pp. 1–12, 2018.

[93] E. Jacob, “Influence of hypoxia and of hypoxemia on the development of cardiac activity in zebrafish larvae,” *Am J Physiol Regul Integr Comp Physiol*, pp. 911–917, 2002.

[94] P. Rombough, “Gills are needed for ionoregulation before they are needed for O2 uptake in developing zebrafish, Danio rerio,” *The Journal of Experimental Biology*, pp. 1787–1794, 2002.

[95] E. Puybareau, M. Léonard, and H. Talbot, “An automated assay for the evaluation of mortality in fish embryo,” in *Mathematical Morphology and Its Applications to Signal and Image Processing*, Reykjavik, Iceland, 2015.

[96] E. Puybareau, H. Talbot, and M. Léonard, “Automated heart rate estimation in fish embryo,” in *5th International Conference on Image Processing Theory, Tools and Applications IPTA*, 2015.

[97] H. Heijmans, *Morphological image operators*. Boston: Academic Press, 1994.

[98] F. Meyer and S. Beucher, “Morphological segmentation,” *J. Vis. Commun. Image Represent.*, pp. 21–46, 1990.

[99] R. Deriche, “Using canny’s criteria to derive a recursively implemented optimal edge detector,” *International Journal of Computer Vision*, pp. 167–187, 1987.

[100] R.A. Jarvis, “On the identification of the convex hull of a finite set of points in the plane,” *Information processing letters*, pp. 18–21, 1973.

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

[102] D. G. Lowe, “Object recognition from local scale-invariant features,” in *IEEE International Conference on Computer Vision*, Kerkyra, Greece, 1999, pp. 1150–1157.

[103] P. Salembier, A. Oliveras, and L. Garrido, “Antiextensive connected operators for image and sequence processing,” *IEEE Transactions on Image Processing*, pp. 555–570, 1998.

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

[105] S. Van der Walt *et al.*, “The scikit image contributors : Scikit image : Image processing in Python,” *PeerJ*, pp. 1–18, 2014.

[106] M. Couprie, L. Marak, and H. Talbot, “Pink image processing library,” presented at the 4th European meeting on Python in Science (Euroscipy), Paris, 2011, pp. 1–4.

[107] G. Bradski, “The opencv library,” *Dr. Dobb’s Journal of Software Tools*, pp. 120–126, 2000.

[108] H. Talbot, H. Phelippeau, M. Akil, and S. Bara, “Efficient poisson denoising for photography,” in *Proceedings of the 16th IEEE International Conference on Image Processing*, 2009, pp. 3881–3884.

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

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

[111] 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, pp. 331–340.

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

[113] F. Pedregosa *et al.*, “Scikit-learn: Machine Learning in Python,” *Journal of Machine Learning Research*, pp. 2825–2830, 2011.

[114] 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.

[115] “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.

[116] S. Ali, J. Aalders, and M.K. Richardson, “Teratological Effects of a Panel of Sixty Water-Soluble Toxicants on Zebrafish Development,” *Zebrafish*, pp. 129–141, 2014.

[117] M. S. Yuea, R.E. Peterson, and W. Heidemana, “Dioxin inhibition of swim bladder development in zebrafish: Is it secondary to heart failure?,” *Aquatic Toxicology*, pp. 10–17, 2015.

[118] S. Klein, U.A. van der Heide, I.M. Lips, M. van Vulpen, M. Staring, and J.P.W. Pluim, “Automatic segmentation of the prostate in 3D MR images by atlas matching using localized mutual information,” *Medical Physics*, pp. 1407–1417, 2008.

[119] T. Rohlfing, R. Brandt, R. Menzel, and C.R. Mauer Jr., “Evaluation of atlas selection strategies for atlas-based image segmentation with application to confocal microscopy images of bee brains,” *NeuroImage*, pp. 1428–1442, 2004.

[120] J.E. Iglesias and M.R. Sabuncu, “Multi-atlas segmentation of biomedical images: A survey,” *Medical Image Analysis*, pp. 205–219, 2015.

[121] E.W. Dijkstra, “A Note on Two Problems in Connexion with Graphs,” *Numerische Mathematik*, pp. 269–271, 1959.

[122] B. Appleton and H. Talbot, “Globally Optimal Geodesic Active Contours,” *Journal of Mathematical Imaging and Vision*, pp. 67–86, 2005.

[123] J.B.A. Maintz and M.A. Viergever, “A survey of medical image registration,” *Medical Image Analysis*, pp. 1–36, 1998.

[124] H. Lester and S.R. Arridge, “A survey of hierarchical non-linear medical image registration,” *Pattern Recognition*, pp. 129–149, 1999.

[125] D. Mattes, D.R. Haynor, H. Vesselle, T.K. Lewellen, and W. Eubank, “PET-CT image registration in the chest using free-form deformations,” *IEEE Transactions on Medical Imaging*, pp. 120–128, 2003.

[126] P. Thevenaz and M. Unser, “Optimization of Mutual Information for Multiresolution Image Registration,” *IEEE Trans. Image Process.*, pp. 2083–2099, 2000.

[127] F. Maes, D. Vandermeulen, and P. Suetens, “Medical Image Registration Using Mutual Information,” *Proceedings of the IEEE*, pp. 1699–1722, 2003.

[128] S. Klein, M. Staring, K. Murphy, M.A. Viergever, and J.P.W. Pluim, “elastix: A Toolbox for Intensity-Based Medical Image Registration,” *IEEE Transactions on Medical Imaging*, pp. 196–205, 2010.

[129] W. Zhang *et al.*, “Deep Model Based Transfer and Multi-Task Learning for Biological Image Analysis,” *IEEE Transactions on Big Data*, pp. 1–20, 2016.

[130] Y. LeCun, Y. Bengio, and G. Hinton, “Deep learning,” *Nature*, vol. 521, pp. 436–444, 2015.

[131] J. Schmidhuber, “Deep learning in neural networks: An overview,” 2015.

[132] R. Frank, “The perceptron: a probabilistic model for information storage and organization in the brain,” *Psychological review*, p. 386, 1958.

[133] Y. LeCun *et al.*, “Backpropagation applied to hand written zip code recognition,” *Neural Computation*, pp. 541–551, 1989.

[134] G. Hinton *et al.*, “Deep Neural Networks for Acoustic Modeling in Speech Recognition: The Shared Views of Four Research Groups,” *IEEE Signal Process*, pp. 82–97, 2012.

[135] G.E. Dahl and A. Acero, “Context-Dependent Pre-Trained Deep Neural Networks for Large-Vocabulary Speech Recognition,” *IEEE Transactions on Audio Speech Language Processing*, pp. 30–42, 2012.

[136] T. Luong, R. Socher, and C.D. Manning, “Better Word Representations with Recursive Neural Networks for Morphology,” in *Proceedings of the 17th Conference on Computational Natural Language Learning*, Sofia, Bulgaria, 2013, pp. 104–113.

[137] A. Krizhevsky, I. Sutskever, and G.E. Hinton, “ImageNet classification with deep convolutional neural networks,” in *Proceedings of the 25th International Conference on Neural Information Processing Systems*, Lake Tahoe, NV, USA, 2012, pp. 1097–1105.

[138] A. Isin, C. Direkoglu, and M. Sah, “Review of MRI-based brain tumor image segmentation using deep learning methods,” *Procedia Computer Science*, pp. 317–324, 2016.

[139] M.D. Zeiler and R. Fergus, “Visualizing and Understanding Convolutional Networks,” in *Proceedings of the European Conference on Computer Vision*, Zurich, Switzerland, 2014, pp. 818–833.

[140] J. Donahue *et al.*, “Long-term recurrent convolutional networks for visual recognition and description,” in *Proceedings of the Computer Vision and Pattern Recognition*, Boston, MA, USA, 2015, p. 677.

[141] I. Sutskever, O. Vinyals, and Q.V. Le, “Sequence to Sequence Learning with Neural Networks,” in *Proceedings of the Advances in Neural Information Processing Systems*, Montreal, QC, Canada, 2014, pp. 3104–3112.

[142] S. Webb, “Deep learning for biology,” *Nature*, pp. 555–557, 2018.

[143] C. Cao *et al.*, “Deep Learning and Its Applications in Biomedicine,” *Genomics Proteomics Bioinformatics*, pp. 17–32, 2018.

[144] C. Angermueller, T. Parnamaa, L. Parts, and O. Stegle, “Deep learning for computational biology,” *Molecular systems biology*, pp. 1–16, 2016.

[145] “Opportunities and obstacles for deep learning in biology and medicine,” *Jounal of the Royal Society Interface*, pp. 1–47, 2018.

[146] Y. Wu and G. Wang, “Machine Learning Based Toxicity Prediction: From Chemical Structural Description to Transcriptome Analysis,” *International Journal of Molecular Science*, pp. 1–20, 2018.

[147] S. Ekins, “The Next Era: Deep Learning in Pharmaceutical Research,” *Pharmaceutical Research*, pp. 2594–2603, 2017.

[148] F. Zhong *et al.*, “Artificial intelligence in drug design,” *Science China Life Sciences*, pp. 1191–1204, 2018.

[149] A. Korotcov, V. Tkachenko, D.P. Russo, and S. Ekins, “Comparison of Deep Learning With Multiple Machine Learning Methods and Metrics Using Diverse Drug Discovery Datasets,” *Molecular Pharmaceutics*, pp. 1–33, 2018.

[150] L. Zhang, J. Tan, D. Han, and H. Zhu, “From machine learning to deep learning: progress in machine intelligence for rational drug discovery,” *Drug Discovery Today*, pp. 1–6, 2017.

[151] D. Jimenez-Carretero *et al.*, “Tox_(R)CNN: Deep learning-based nuclei profiling tool for drug toxicity screening,” *PLoS Computational Biology*, pp. 1–23, 2018.

[152] Y. Xu, Z. Dai, F. Chen, S. Gao, J. Pei, and L. Lai, “Deep Learning for Drug-Induced Liver Injury,” *Journal of Chemical Information and Moeling*, pp. 2085–2093, 2019.

[153] L. Pu, M. Naderi, T. Liu, H-C. Wu, S. Mukhopadhyay, and M. Brylinski, “eToxPred: a machine learning-based approach to estimate the toxicity of drug candidates,” *BMC Pharmacology and Toxicology*, pp. 1–15, 2019.

[154] T-H. Chan, K. Jia, S. Gao, J. Lu, Z. Zeng, and Y. Ma, “PCANet: A Simple Deep Learning Baseline for Image Classification?,” *Journal of Latex Class Files*, pp. 1–15, 2014.

[155] K. He, X. Zhang, S. Ren, and J. Sun, “Deep Residual Learning for Image Recognition,” in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Las Vegas, USA, 2016, pp. 770–778.

[156] Y. Bengioy, A. Courville, and P. Vincent, “Representation Learning: A Review and New Perspectives,” *IEEE Transactions on Software Engineering*, pp. 1–30, 2014.

[157] E.T Quinto, *The Radon Transform, Inverse Problems, and Tomography*, American Mathematical Society Short Course. 2005.

[158] A.G. Lindgren and P.A. Rattey, “The inverse discrete Radon transform with applications to tomographic imaging using projection data,” *Advances in Electronics and Electron*, pp. 359–410, 1981.

[159] S. Kim and A.K. Khambampati, “Mathematical concepts for image reconstruction in tomography,” in *Industrial Tomography: Systems and Applications*, Woodhead Publishing Series in Electronic and Optical Materials., 2015, pp. 305–346.

[160] S. Laporte, W. Skalli, J.A. de Guiseb, F. Lavaste, and D. Mitton, “A Biplanar Reconstruction Method Based on 2D and 3D Contours: Application to the Distal Femur,” *Computer Methods in Biomechanics and Biomedical Engineering*, pp. 1–6, 2003.

[161] W. Yu and G. Zheng, “Atlas-Based 3D Intensity Volume Reconstruction from 2D Long Leg Standing X-Rays: Application to Hard and Soft Tissues in Lower Extremity,” in *Intelligent Orthopaedics*, Springer., 2018, pp. 105–112.

[162] D. Fleischmann and F.E. Boas, “Computed tomography—old ideas and new technology,” *European Society of Radiology*, pp. 510–517, 2011.

[163] H. Lamecker, T.H. Wenckebach, and H-C. Hege, “Atlas-based 3D-Shape Reconstruction from X-Ray Images,” in *Proceedings of the 18th International Conference on Pattern Recognition (ICPR)*, 2006.

[164] J.T.Dobbins and H.P. McAdams, “Chest tomosynthesis: Technical principles and clinical update,” *European Journal of Radiology*, pp. 244–251, 2009.

[165] Y. Zhang *et al.*, “A comparative study of limited-angle cone-beam reconstruction methods for breast tomosynthesis,” *Medical Physics*, pp. 3781–3795, 2006.

[166] R. Maksimovic, S. Stankovic, and D. Milovanovic, “Computed tomography image analyzer: 3D reconstruction and segmentation applying active contour models — ‘snakes,’” *International Journal of Medical Informatics*, pp. 58–59, 2000.

[167] G. Farneback, “Two-Frame Motion Estimation Based on Polynomial Expansion,” in *Proceedings of the 13th Scandinavian Conference on Image Analysis*, Sweden, 2003.

[168] D. Marcato *et al.*, “An Automated and High-throughput Photomotor Response Platform For Chemical Screens,” *IEEE*, pp. 7728–7731, 2015.

[169] T-H. Chen, Y-H. Wang, and Y-H. Wu, “Developmental exposures to ethanol or dimethylsulfoxide at low concentrations alter locomotor activity in larval zebrafish: Implications for behavioral toxicity bioassays,” *Aquatic Toxicology*, pp. 162–166, 2011.

[170] P.T. Gauthier and M.M. Vijayan, “Nonlinear mixed-modelling discriminates the effect of chemicals and their mixtures on zebrafish behavior,” *Scientific Reports*, pp. 1–11, 2018.