Angela Bairos Pimentel

Licenciatura em Ciências da Engenharia Biomédica

Algorithm for the Parkinson's Disease Behavioural Models Characterization using a Biosensor

Dissertação para obtenção do Grau de Mestre em
Engenharia Biomédica

Orientador: Prof. Doutor Hugo Gamboa
Co-orientadora: Prof[a]. Doutora Ana Dulce Correia

Júri:

Presidente: Prof[a]. Doutora Maria Adelaide Jesus
Arguentes: Prof[a]. Doutora Carla Quintão
Vogais: Prof. Doutor Hugo Gamboa

FACULDADE DE CIÊNCIAS E TECNOLOGIA
UNIVERSIDADE NOVA DE LISBOA

Novembro, 2012

Algorithm for the Parkinson's Disease Behavioural Models Characterization using a Biosensor

Aos meus pais

Acknowledgements

This dissertation could not have been written without the immense knowledge of Dr. Hugo Gamboa who not only served as my supervisor but also encouraged and challenged me throughout my academic program. I'm very thankful to Dra. Ana Correia, my co-supervisor from Instituto de Medicina Molecular (IMM), whose support, dedication, motivation and enthusiasm helped me with no doubt during this research. Many thanks to Dr. Sérgio Cunha for his support with the biosensor, and with the development of the algorithm.

To *PLUX-Wireless Biosignals*, S.A. workers for welcome me every day and allowing me to belong to their business daily-life. A special thanks to Joana Sousa for her supervision over my work, and Neuza Nunes for her dedication and patience while helping me during the development of my work.

I'm also very thankful to IMM investigators, in special to Dr. Rui Santos for his constant support in the Institute. With the business environment lived at PLUX and the opportunity to also meet the daily work of researchers at the IMM there's no doubt that both enriched me in a personal and professional level.

To my colleagues Ricardo Chorão, Rodolfo Abreu, Diliana Santos and Nuno Costa whose knowledge in their thesis helped me in developing and improving some parts of mine. A big thank to you all. Also to André Carreiro for his contribution in the algorithm development.

Last but not least, I'm very thankful to my family. To Marco Pimentel, my cousin, godfather and role model who guides, advices and encourages me professionally. To my brother who I know that will be dedicated and strong in the next stage of his life. To the one and only, Carlos Sousa, who has been by my side everyday, and helped me during my research with patience and dedication and to my parents: *Sem vocês nunca teria chegado até aqui. Obrigada por me darem esta oportunidade, pelo vosso constante apoio e por aceitarem o caminho que escolhi. São o modelo que espero um dia conseguir vir a ser. Tenho muito orgulho em vocês. A vós dedico esta tese.*

Abstract

The neurodegenerative disease, Parkinson's Disease (PD) constitutes a major health problem in the modern world, and its impact on public health and society is expected to increase with the ongoing ageing of the human population. This disease is characterized by motor and non-motor manifestations that are progressive and ultimately refractory to therapeutic interventions. The degeneration of dopaminergic neurons emanating from the substantia nigra is largely responsible for the motor manifestations. Thus, understanding the behaviour related to this disease is an added value for the diagnosis and treatment of PD. Also, *in vivo* models are essential tools for deciphering the molecular mechanisms underpinning the neurodegenerative process. Zebrafish has several features that make this species a good candidate to study PD. In particular, the occurrence of behavioural phenotypes of treated animals with neurotoxin drugs that mimic the disease has been investigated. And, an electric biosensor, Marine On-line Biomonitor System (MOBS) is being used for the real-time quantification of such behaviour. This equipment allows quantifying the fish movements through signal processing algorithms. Specifically, the algorithm is used for the evaluation of fish locomotion detected by a series of bursts in the domain of MOBS that correspond to the zebrafish tail-flip activity. In this thesis we proceeded to the development of an algorithm affording a electrical signal discrimination between "healthy" and "ill" zebrafish and consequently improving the detection of parkinsonism-like phenotypes in zebrafish. The first approach was the improvement of the existent algorithm. However, the first analysis failed to distinguish between different behavioural phenotypes when fish were treated with the neurotoxin *6-hydroxydopamine* (6-OHDA). Consequently, we generated a new algorithm based on Machine Learning techniques. As a result, the novel algorithm provided a classification over the health condition of the fish, if the same is "healthy" or "ill" with its respective probability and the level of activity of the fish in number of tail-flips per minute. The method Support Vector Machine (SVM) was useful for the classification of the fish events.

The zero crossing rate parameter was used for the characterization of the swimming activities. The algorithm was also integrated in the platform Open Signals, and for a faster evaluation of the signals, the algorithm implementation included parallel programming methods. This algorithm is a useful tool to study behaviour in zebrafish. Not only it will allow a more realistic study over the PD research area but also test and assess new drugs that use zebrafish as animal model.

Keywords: PD, Zebrafish, MOBS, Behaviour, Machine Learning, Zero Crossing Rate, SVM.

Resumo

A doença neurodegenerativa, doença de Parkinson (PD) constitui um grave problema de saúde no mundo, e o seu impacto sobre a saúde pública e sociedade irá aumentar com o envelhecimento contínuo da população humana. Esta doença é caracterizada por manifestações motoras e não motoras, que são progressivas e em última análise refractárias às intervenções terapêuticas. A degeneração de neurónios dopaminérgicos que emanam da substância negra é em grande parte responsável pelas manifestações motoras. Assim, o estudo do comportamento relacionado com esta doença é uma mais valia para diagnóstico e tratamento da PD. Além disso, modelos *in vivo* são ferramentas essenciais para decifrar os mecanismos moleculares subjacentes ao processo neurodegenerativo. O peixe zebra tem várias características que tornam esta espécie um bom candidato para o estudo da PD. Em particular, tem-se investigado a ocorrência de fenótipos comportamentais dos animais tratados com neurotoxinas que simulam a doença. E, um biossensor eléctrico MOBS está sendo utilizado para a quantificação em tempo real de tais comportamentos. Este equipamento permite quantificar os movimentos dos peixes através de algoritmos de processamento de sinal. Especificamente, o algoritmo é usado para a avaliação da locomoção do peixe, detectado com base em variações no domínio de MOBS, que correspondem ao número de barbatanadas por minuto do peixe zebra. Nesta tese, procedeu-se ao desenvolvimento de um algoritmo que ofereça uma discriminação dos sinais eléctricos entre peixes zebra "saudáveis"ou "doentes", e consequentemente, permitir melhorar a detecção de fenótipos parkinsonianos do peixe zebra. A primeira abordagem consistiu em melhorar o actual algoritmo. No entanto, a primeira análise falhou numa distinção entre fenótipos comportamentais quando os peixes foram tratados com a neurotoxina 6-OHDA. Consequentemente, geramos um novo algoritmo baseado em técnicas de Machine Learning. Como resultado, o novo algoritmo proporcionou uma classificação sobre o estado de saúde do peixe, se o mesmo está "saudável"ou "doente", com a sua respectiva probabilidade e o nível de actividade do peixe em número de barbatanas por minuto. O método SVM mostrou-se útil para a classificação dos peixes. O parâmetro

zero crossing rate, foi útil para caracterizar o nível de actividade dos peixes. O algoritmo também foi integrado na plataforma Open Signals, e para permitir uma avaliação rápida dos sinais, a implementação do algoritmo incluiu métodos de programação em paralelo. Este algoritmo é uma ferramenta útil para estudar comportamentos no peixe zebra. Não só irá permitir um estudo mais realístico na área de investigação da PD mas também testar e avaliar novas drogas que usem o peixe zebra como modelo animal.

Palavras-chave: Doença de Parkinson, Peixe Zebra, MOBS, Comportamentos, Machine Learning, Zero Crossing Rate, SVM.

Contents

List of Figures

List of Tables

Acronyms

CNS Central Nervous System

DA Dopaminergic

DFT Discrete Fourier Transform

FFT Fast Fourier Transform

FPR False Positive Rate

hpf hours post-fertilization

IMM Instituto de Medicina Molecular

MFB Multispecies Freshwater Biomonitor

MOBS Marine On-line Biomonitor System

6-OHDA *6-hydroxydopamine*

PD Parkinson's Disease (*Doença de Parkinson*)

PSD Power Spectral Density

ROC Receiver Operating Characteristic

SVM Support Vector Machine

TPR True Positive Rate

1

Introduction

1.1 Motivation

People are living longer. Since Parkinson's Disease (PD) most commonly affects the elderly, the number of sufferers will rise substantially in the years to come. The prevalence of PD is 1% to 2% of persons older than 60 years [5]. In turn, the need for clinical and social services to care for and support patients with PD will increase at a rapid rate, with major implications for the resources that are allocated to healthcare [6].

There is currently no form of pharmacotherapy available that has shown to delay the progression of PD. However, there are a range of drugs that can treat the symptoms of the condition and consequently improve the patient's life quality. Also, the correct diagnose of PD especially in the early stages of the disease, represent quite a challenge. PD can cause a broad spectrum of symptoms and there are significant variations between patients in the way the disease manifests itself and the speed with which symptoms develop. However three symptoms are clearly fundamental: hypokinesia (reduction in movement), rigidity and tremor [6].

Despite all the recent progress in the understanding of PD, the molecular mechanisms underlying this disease are still obscure. The available *in vivo* models have failed to fully recapitulate all features of PD. However, the teleost, *Danio rerio*, has emerged as a valuable model to study different aspects associated with neurodegeneration. In particular, zebrafish display specialized neurons with direct relevance to human neuronal disorders. It has been proved that the loss of dopaminergic neurons induces changes in the behaviour of the fish, specifically decreases its level of swimming activity. Therefore the assessment of behavioural phenotypes in zebrafish can be an important contribution for studying the molecular basis of PD as well as in the drug screening analysis.

1.2 Objectives

The major aim of this work was the development of an algorithm that, once combined with the MOBS biosensor, allows to differentiate electric signals between "healthy" and "ill" zebrafish and also provide its swimming activity in number of tail-flips per minute. Hence it will improve the detection of parkinsonism-like phenotypes in zebrafish.

1.3 Thesis Overview

The structure of this thesis is schematically represented in Figure 1.1.

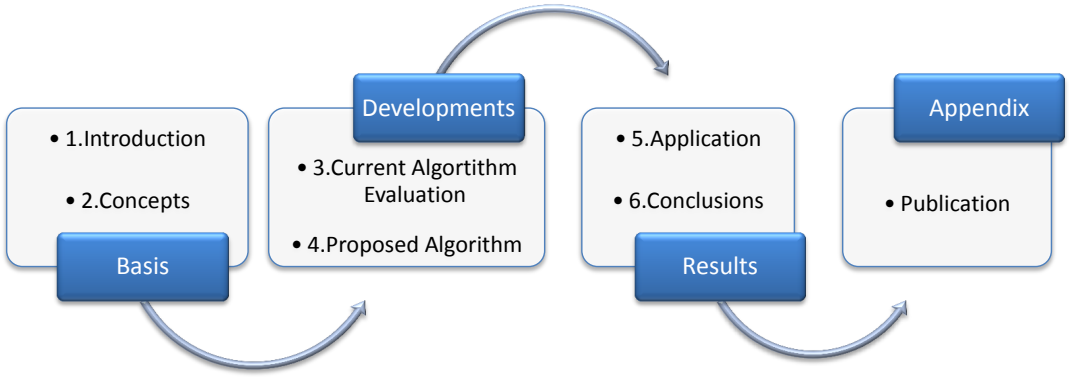


Figure 1.1: Thesis overview.

In the first two chapters the basis that support this research is reported. The motivation and objectives are presented in Chapter 1. There was an initial effort to characterize the behaviour of zebrafish using an algorithm that provided the number of tail-flips per minute. Thus, the association between the zebrafish and PD, the current algorithm used to characterize the behaviour of zebrafish, as well as the description of the biosensor MOBS are described in Chapter 2. In this chapter it is also reported machine learning techniques that were used in the implementation of the new algorithm.

Chapter 3 examines with more detail the current algorithm output using video, which required the development of a functionality in the platform Open Signals that allowed synchronism between video and signal. This detailed analysis demonstrated the need for creating a new algorithm that could simulate zebrafish behaviour as real as possible. Chapter 4 presents the development of the new algorithm using machine learning techniques as well as its validation.

The following chapters addresses the results. Chapter 5 demonstrates the application of the new developed algorithm using a new case study related with PD. Chapter 6 presents the conclusions of this research work as well as its future work. The Appendix

contains the paper published in the context of this research work.

This thesis was written using the LaTeX environment [7]. The signal acquisition uses the software *MATLAB* and the signal processing algorithms were developed in *Python*. The *Orange* software was used to build the classifier [8]. The final algorithm was also integrated in the platform Open Signals that required some knowledge in *Javascript* and *HTML*.

This dissertation was developed at *PLUX - Wireless Biosignals*, S. A. and at IMM from Lisbon University.

2

Concepts

2.1 Zebrafish and Parkinson's Disease

2.1.1 Zebrafish

Zebrafish (scientific name - *Danio rerio*) are tropical fresh water fish from Ganges region of India. They can be found in Nepal, Bangladesh, Pakistan and Myanmar [9].

Figure 2.1: Zebrafish [1].

The fish seen in Figure 2.1 is named for the five horizontal blue stripes on the side of the body. Males are torpedo shaped and have gold stripes between the blue stripes; females have a larger, whitish belly and have silver stripes instead of gold. Fully grown adults are around 3-5 cm long and 1 cm wide.

Zebrafish are omnivorous, meaning they will eat plants and animals, like zoo-plankton, insects and phytoplankton. In captivity they eat conventional flaked fish food [9].

2.1.2 Zebrafish as a model organism

Most insights into human disease are a result of experiments that would be unethical or unfeasible to perform on humans. Instead, biomedical research uses models to look at the functions of the genes involved in maintaining healthy organisms in order to obtain vital clues about the causes and progression of human diseases.

People are familiar with the use of mice and rats as model organisms (lab rats). As mammals they are very similar to humans, therefore they can be used to study complex processes underlying normal human development and diseases.

If we want to know something simple that is likely to occur in all living organisms than we can use bacteria or yeast as they are easy and cheap to look after and they're very well understood. However, sometimes they can be too simple in terms of biological organization.

Zebrafish are the ideal model organism to bridge the gap between "too simple" and "too complex". They are aquatic vertebrates and have similar body plans (and similar tissues and organs) to humans, and they are much easier and with reduced cost to breed than mice and rats. Zebrafish has a short generation time (3 months) and breed prodigiously (hundreds of offspring per female per week). They develop from a single cell in fertilized egg in about 24 hours (for a mouse it takes about 21 days). Also, the embryos are large, robust, transparent, easy to manipulate genetically and are developed outside the mother. Some drugs can even be administered by adding directly to the tank. Zebrafish mutations phenocopy many human disorders and the genome sequence of zebrafish is near completion [9].

However, besides all the advantages, zebrafish also have disadvantages when compared to other models. They are not mammals, so they are not as closely related to humans as mice. Therefore, all the new discoveries must later be verified in a mammal model [10]. It is the similarity between the genes, which scientists call conservation, or genetic homology, the reason why fish can be used to study human diseases. Hence, zebrafish can be used as a model organism.

The Central Nervous System (CNS) coordinates the activity of the body. It includes the brain and the spinal cord. Disorders in the CNS can affect control of physical movement, alteration of mood, change in sociability and absence of, or decline in communication [9].

More and more groups are becoming interested in the fact that adult zebrafish possess a high capacity for regeneration. Amazingly, spinal cord tissue can regenerate after a complete transection. In a process that takes about 6 weeks, approximately 80% of animals given a posterior injury achieve functional recovery [11]. This phenomenon is based on the striking ability of the CNS neurons to recover, traverse the lesion, and re-establish functional connections [12].

Some of the neurological disorders that can be studied with zebrafish are Hereditary

Spastic Paraplegia, Parkinson's Disease, Huntington's Disease, Motor Neuron Disease and Multiple Sclerosis. These diseases cause loss of voluntary movement control in patients. Given that their health is aggravated over time, they are called neurodegenerative disorders. At this moment there is no cure, and any treatment only slows the progression of symptoms [9].

2.1.3 Parkinson's Disease

PD was first described in 1817 by James Parkinson and is the second most common neurodegenerative disorder, after Alzheimer's disease [13]. The PD is characterized by tremor, muscle rigidity, a slowing of physical movement, and can also cause cognitive and mood disturbances. It results of the loss of nerve cells in part of the brain known as the substantia nigra. These cells are called Dopaminergic (DA) neurons as they produce the neurotransmitter - dopamine, which is used to send messages to the parts of the brain that co-ordinates movements. When around 80% of the DA neurons are lost, the symptoms of PD start to show. The cause of PD is not absolutely clear; there are some mutations associated with the loss of DA neurons and it is known that some toxins or chemicals may also cause the disease [9].

2.1.4 Parkinson's Disease in Zebrafish

The DA nervous system in zebrafish is well characterized in both embryos and adult zebrafish. DA neurons are first detected between 18 and 19 hours post-fertilization (hpf). Some toxins known to induce DA cell loss in other animal models have now also been tested in adult zebrafish, as for example, the *6-hydroxydopamine* (6-OHDA) which is a neurotoxin that induces death of the DA cells [14, 15, 16]. The swimming velocity and total distance moved decreased after exposure to this neurotoxin [17, 18]. Thus the evaluation of swimming behaviour can be related with the loss of DA cells, and consequently with PD.

2.2 Marine On-line Biomonitor System – MOBS

A biosensor is defined as a self-contained integrated device that is capable of providing specific quantitative analytical information using a biological recognition element. The main advantages are the possibility of a continuous monitoring, the high specificity and sensitivity [19].

Biosensors are an essential control and safety tool for our environmental and health quality and are commonly used in medicine. Many of today's biosensor applications are similar, in that they use living organisms which respond to toxic substances or other stressors at a much lower level than us to warn us of their presence. Under this scope, the MOBS was developed, an automated system for recording behavioural responses of

marine and fresh water species. This device was firstly applied successfully in the environmental field, and nowadays is used in the biomedical field, in particular, by sensing behavioural changes in organisms as an indication of stress or disease. Zebrafish has proved to be a suitable model candidate for this research since it has been used in medical research during the past years, e.g in development studies [20], drug toxicity assessments [21] and neurodegenerative diseases [22]. Previous studies using this electronic device were used to asses water quality [2] and testing analgesics [23].

2.2.1 The main device

MOBS is an automatic system for recording behavioural responses of marine and fresh water species. Low power electrical signals are modulated by the behavioural activities of the organisms and then monitored, processed and analysed in real time.

The device monitors changes in electric fields caused by organism movements by means of non-invasive electrodes. It is an external automated transducer designed and manufactured at Faculty of Engineering of the University of Porto (Portugal). The MOBS device can record continuously specific behavioural activities of marine fish species, such as ventilation frequency and swimming activities and can quantify electrical signatures patterns from individual organisms as well as groups of animals.

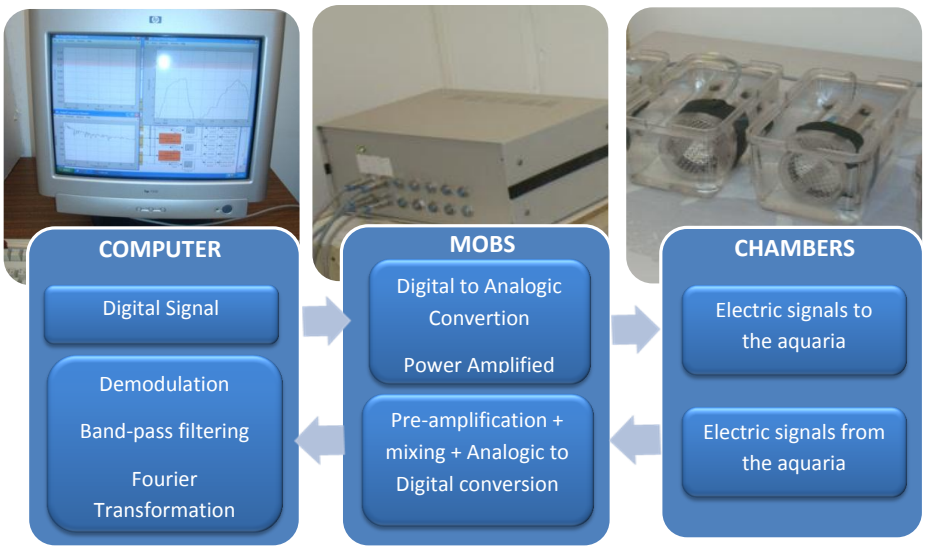


Figure 2.2: The operation diagram of the MOBS system adapted from [2].

The MOBS can manage up to 14 containers in parallel which consists of cylindrical chambers with 6 cm in diameter and 10 cm long [2]. The device injects weak analogue

electrical signals into the water of the test chambers through a pair of non-invasive stainless steel electrodes. The response is measured as a change in impedance of the water column received by another pair of non-invasive stainless steel electrodes associated with movements of the fish [23]. The electrodes are attached vertically at the aquaria walls such that they provide a homogeneous distributed electric field across the entire aquarium.

The main device is controlled via an USB port by external processing software which produces signals in the digital domain (at 48000 samples/s or 48 kHz). These are converted by the main device into analogical electrical signals, power amplified and transmitted to the independent testing units at which they are conducted into the water by a pair of non-invasive stainless steel electrodes – Figure 2.2. In response to the behavioural signatures of the organisms, the amplitudes of the electrical signals are modulated and then received by a second pair of electrodes. In the main device they are amplified and converted back to the digital domain at 48000 samples/s, before filtered, demodulated and down-sampled at 100 Hz by the external computer software. Then, they are analysed in the frequency domain (Fourier transform with proper windowing) in chunks of about 10 s.

- **Discrete Fourier Transform (DFT)**: The frequency domain allows a different vision over the signal, and simplifies some operations like convolution and correlation. It is defined as:

$$A_r = \sum_{k=0}^{N-1} X_k exp(-2j\pi rk/N) \quad \text{with} \quad r = 0, 1, \ldots, N-1 \tag{2.1}$$

  where A_r is the r^{th} coefficient of the DFT and X_k denotes the k^{th} sample of the time series which consists of N samples and $j = \sqrt{-1}$. Also worth mentioning the Fast Fourier Transform (FFT) which is a method for efficiently computing the DFT of time series (discrete data samples)[24].

Upon processing, the system provides a signal in the frequency band of 0.2 Hz to 40 Hz that is correlated with the fish activity. As the harmonics are relevant to obtain signal shapes, they defined the cut-off frequency of the filters at around 45 Hz. This allows to obtain a clear representation of the direct time domain signal and its frequency spectrum, which is suitable to broaden the range of pattern recognition algorithms that can be used afterwards [2].

2.2.2 Other biosensor

Another biosensor similar to this one is the Multispecies Freshwater Biomonitor (MFB), which is based on the detection of impedance changes in the water across a test chamber due to movements of an organism in an alternating electrical field. The MFB is the first multi-species aquatic biomonitor available in the European market. It has been applied to

several kinds of freshwater organisms, mainly to test behavioural effects to the exposure of pharmaceutical effluents and to pollution detection on aquatic invertebrates and fish. These studies were analysed using the FFT [25, 26].

Yet one of the advantages from this biosensor related to MOBS is the fact that in order to prevent the organisms from touching the electrodes, the chambers walls are covered with nylon netting (50μm) [27].

2.3 Behaviour in Zebrafish

Behaviour is the final outcome of a sequence of neurophysiological events including stimulation of sensory and motor neurons, muscular contractions, and release of chemical messages [27]. On-line biomonitors frequently use behaviour as an end point, which provides a visual and, thus, measurable response at the whole-organism level. This method generates fast and sensitive results that can be integrated in many biological functions [28].

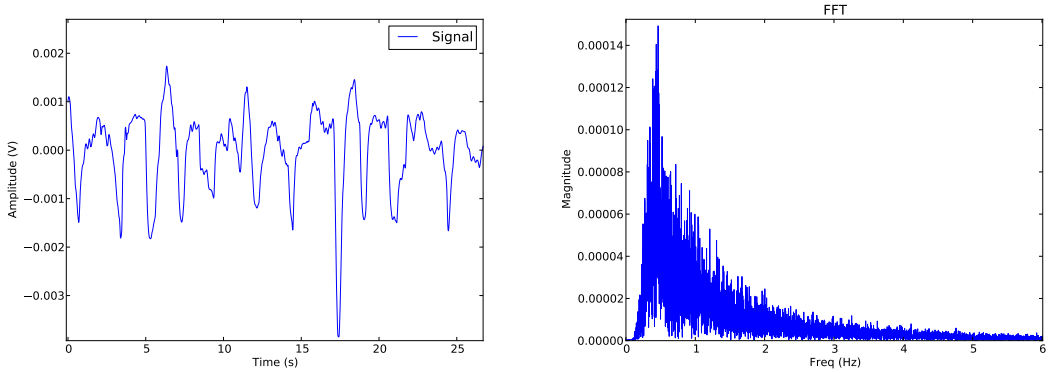
There is a lack of studies on complex behaviour in zebrafish; although it is recognised as having great potential as a model for understanding the genetic basis of human behavioural disorders. One area of interest has been the effect of drugs on behaviour and also the studying of social behaviour, learning and memory.

The number of behavioural studies of zebrafish looks set to increase, and many researchers whose primary expertise is in genetics or development biology are using behavioural protocols as a paradigm for testing the reinforcing properties of drugs of abuse. One of the problems with designing and conducting behavioural experiments is demonstrating that the results are a valid measure of the behaviour under consideration. Thus there is a need for adequate controls, in order to ensure that the results are not due to unrelated artefacts, for example, outside disturbance, either visual or auditory and acclimatisation. The behaviour may also vary according to the time of the day at which observations are recorded, especially in relation to matting behaviour and feeding regime [15]. The next subsections describe the behaviour studied with MOBS.

2.3.1 Locomotion

A typical activity using zebrafish in the time domain of MOBS is shown in Figure 2.3(a). The amplitude of the fish activity in the time domain is in the order of the mV.

Locomotion can be presented as a series of bursts in the time domain, and can cover a broad frequency spectrum, at which ventilation is occasionally present. Although the strong bursts can cover a broad frequency spectrum, still most of the energy is located in the range between 0 Hz and 1 Hz as seen in Figure 2.3(b). In contrast, the spectrum for locomotion looks often like a random and unstructured signal for an inexperienced user. A clear separation between the signals for ventilation and locomotion in this fish cannot be ensured [2].



(a) Time domain (b) Frequency domain

Figure 2.3: Locomotion of a "healthy" fish represented in time and frequency domain (as "healthy" is meant that is neither "ill" nor transgenic).

2.3.2 Ventilation

Ventilation consists in opening and closing of mouth/operculum and causes only very local disturbances in the water. The smaller the distance between electrodes and organism, the better the corresponding electric field can be identified and quantified. Typically ventilation generates waves of triangular shape with a higher frequency and smaller amplitude than most of the energy located for locomotion. Ventilation can be detected and quantified by frequencies and thus requires a clear peak in the frequency spectrum [2]. However, ventilation will not be studied with zebrafish given its high level of activity.

2.4 Current Algorithm

An algorithm is a sequence of instructions designed to solve a problem [29]. The current algorithm used to characterize the behaviour of zebrafish consists in the evaluation of a specific locomotion behaviour of zebrafish, with a series of bursts in the domain of MOBS corresponding to the zebrafish tail-flip activity. Thus the outcome reflects the number of tail-flips per minute per individual fish [23].

The algorithm process uses the derivative of the signal in the time domain. This will allow the detection of the behaviour tail-flip, with representative peaks of the derivative that characterize the strong bursts. These peaks are detected using the standard deviation of the signal multiplied by a factor, to allow the comparison between the two parameters standard deviation and derivative, given that, the behaviour tail-flip can be detected. However, this algorithm detection compared with the actual fish behaviour requires confirmation, and this can be accomplished by using video synchronized with the signal in the time domain.

Besides the multiplicative factor, other thresholds are used to limit the maximum and minimum amplitude of the fish activity.

Figure 2.4 presents an example of a "healthy" fish behaviour associated with its derivative. The fish strong bursts result in signal (blue) variations and consequently provide defined peaks of the difference (green). Thus the algorithm output (red) will detect these peaks using a threshold that is defined by the standard deviation multiplied by a factor (black). To refer that the difference, the standard deviation and the algorithm output were amplified in this case to simplify visualization.

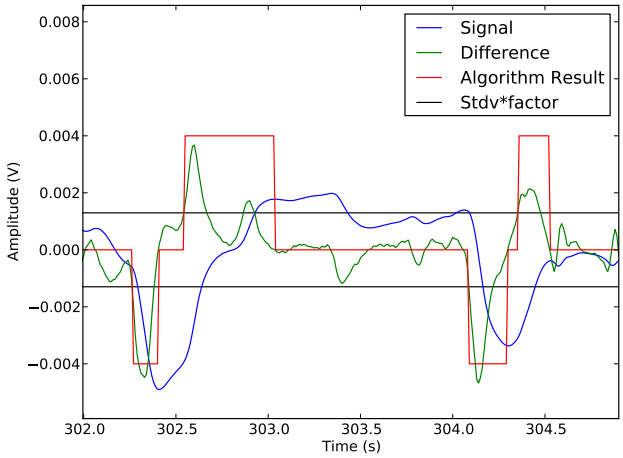


Figure 2.4: Algorithm process. The signal is represented in blue, the difference in green, the algorithm output in red and the standard deviation multiplied by a factor in black.

For an easy behaviour analysis, the algorithm is created with -1, 0 and 1 values as seen in Figure 2.4 (red). The values -1 and 1 are attributed if the difference exceeds the standard deviation, and passes to 0 when the difference is null. The 0 value is maintained until the difference exceeds again the standard deviation. Finally the algorithm will count the number of resulting transitions 0/1, 0/-1 and divide it by the total time of the signal in minutes providing the number of tail-flips per minute, of an individual fish.

2.4.1 Need for improvement

The pre-defined thresholds (multiplicative factor, maximum and minimum amplitude for the fish activity) are one of the reasons for confirmation and improvement. The algorithm only provides one type of behaviour, the tail-flips, which is a measurement of the fish activity (the higher the number of tail-flips, the more active the fish is). Nevertheless the possibility to study other behaviour (e.g. swimming and ventilation) may turn this algorithm more advantageous and complete for future works.

A more detailed analysis in the signal compared to the actual fish behaviour is necessary, which requires synchronism between signal and video. Possible errors from the main device that are visible in the signal need to be detected and filtered.

In the work performed by Correia et. al (2012) [18], a new transgenic line of zebrafish

was developed to study the DA neurons. This transgenic line was treated with the neurotoxin 6-OHDA and behavioural effects investigated with the MOBS biosensor. It was demonstrated that the drug induces behavioural changes that were related to the death of DA neurons. The use of an improved algorithm could contribute as a more sensitive tool in the detection of behavioural phenotypes associated with the loss of the DA neurons. Thus it is essential to confirm if the actual algorithm is in fact detecting the right behaviour - the tail-flips. To develop a new algorithm, Machine Learning techniques are suggested.

2.5 Machine Learning

Machine Learning enables the extraction of implicit, previous unknown, and potentially useful information from data [30].

By Arthur Samuel (1959), machine learning is the field of study that gives computers the ability to learn without being explicitly programmed. A more recent definition by Tom Mitchell (1998) says: "A computer program is said to learn from experience E with respect to some task T and some performance measure P, if its performance on T, as measured by P, improves with experience E" [3].

Machine learning is used do extract information from the raw data in databases - information that is expressed in a comprehensible form and can be used for a variety of purposes. The process is one of abstraction: taking the data, warts and all, and inferring whatever structure underlies it. With machine learning we can use tools and techniques that are used for finding, and describing, structural patterns in data [30].

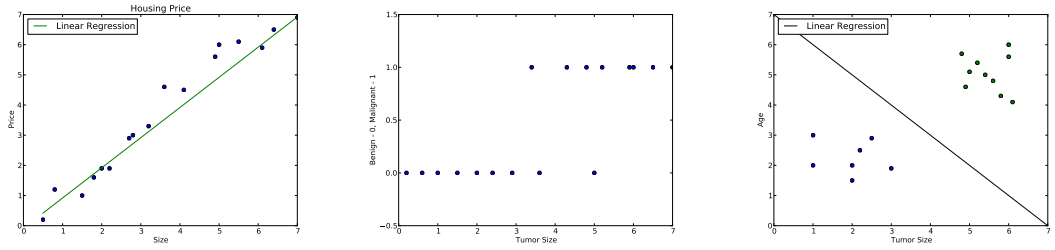
There are different types of machine learning algorithms, the main two types are: unsupervised and supervised learning.

2.5.1 Unsupervised Learning

With unsupervised learning it is intended to let the computer learn by it self. The right answers are not labelled in the data, there is no such supervisor and there is only input data. Finding some structure is possible using clustering algorithms which allows groups separations [3, 31].

2.5.2 Supervised Learning

The idea is to teach the computer how to do something. The right answers are provided in the data set to the algorithm. In a sense, the scheme operates under supervision by being provided with the actual outcome for each of the training examples. In this type of machine learning the regression and classification problems are included.



(a) Linear Regression

(b) Classification

(c) Classification using two input variables. Blue represents benign tumor and green malignant tumor.

Figure 2.5: Supervised learning examples. Adapted from [3].

2.5.2.1 Regression Problems

Predict continuous valued output, for example predict the price of a house according to its size using linear regression - Figure 2.5(a). In cases where the linear model is too restrictive, one can use for example a quadratic or a higher-order polynomial, or any other non-linear function of the input, this time optimizing its parameters for best fit.

Given a training set with m training examples we can represent x as the input variable/feature, y as the output variable or target variable and $h_\theta(x)$ our hypothesis which estimates the output y. It is used to make predictions. Related to Figure 2.5(a) 17 training examples are used, with the size of the house as the input variable and the price as output.

Linear Regression

When the output and all input variables are numeric, linear regression is a natural technique to consider. Also when using more than one variable it is important to consider that there might be a single variable that does all the work and the others are irrelevant or redundant.

The hypothesis using one input variable as seen in Figure 2.5(a) can be expressed as:

$$h_\theta(x) = \theta_0 + \theta_1 x \tag{2.2}$$

Where θ_0 and θ_1 are the parameters used so that $h_\theta(x)$ is close to the output y when using our training examples. Here, the machine learning program optimizes the parameters, θ, such that the approximation error is minimized, that is, our estimates are as close as possible to the correct values given in the training set. In many cases, there is no analytical solution and we need to resort to iterative optimization methods. The most commonly used are gradient descent and normal equation [31].

The success of supervised learning can be judged by trying out the concept description that is learned on an independent set of test data for which the true classifications are known but not made available to the machine [30].

2.5.2.2 Classification problems

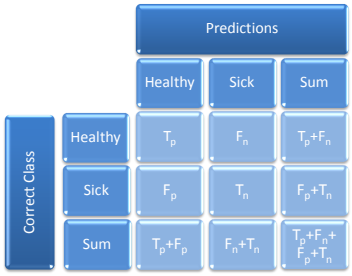
This technique intends to predict discrete valued outputs, for example predict if a tumour is benign or malign according to the tumour size – Figure 2.5(b). It is also possible to use more than one input variable to predict the output as seen in Figure 2.5(c), which uses two input variables, the tumor size and age, to classify if the tumor is benign or malignant.

Classification problems can use two classes (e.g predict if a tumor is benign or malignant), or multi-classes. From figures 2.5(b) and 2.5(c), the aim is to infer a general rule, coding the association between the input attributes and its output. That is, the machine learning system fits a model to the past data to be able to estimate the tumor malignancy for a new situation [3, 31]. Using two classes it is important that our hypothesis is given in terms of probability, so that the class that presents higher probability will be chosen.

Classification Performance

The data produced by a classification scheme during testing are counts of the correct and incorrect classifications from each class. This information is then normally displayed in a confusion matrix - Table 2.1.

Table 2.1: Confusion Matrix. T_p and T_n are the number of true and negative examples respectively. F_p and F_n the number of false positives and negatives respectively.



A confusion matrix is a form of contingency table showing the differences between the true and predicted classes for a set of labelled examples. Considering T_p and T_n the number of true positives and true negatives respectively, F_p and F_n the number of false positives and negatives respectively, there are measures that can be extracted from the confusion matrix:

$$Accuracy = \frac{T_p + T_n}{T_p + F_p + T_n + F_n} \tag{2.3}$$

$$Sensitivity = \frac{T_p}{T_p + F_n} \tag{2.4}$$

$$Specificity = \frac{T_n}{T_n + F_p} \tag{2.5}$$

It is relevant to choose one classifier that maximizes its accuracy when the testing set is

applied. The accuracy from equation 2.3 is the proportion of correctly classified examples among all data classified. The sensitivity - equation 2.4, also called True Positive Rate (TPR) is the number of detected positive examples among all positive examples, e.g. the proportion of healthy people correctly diagnosed as healthy. The specificity - equation 2.5, is the proportion of detected negative examples among all negative examples, e.g. the proportion of sick correctly recognized as sick [8]. A good way of visualising a classifier's performance is with the Receiver Operating Characteristic (ROC) curve – Figure 2.6.

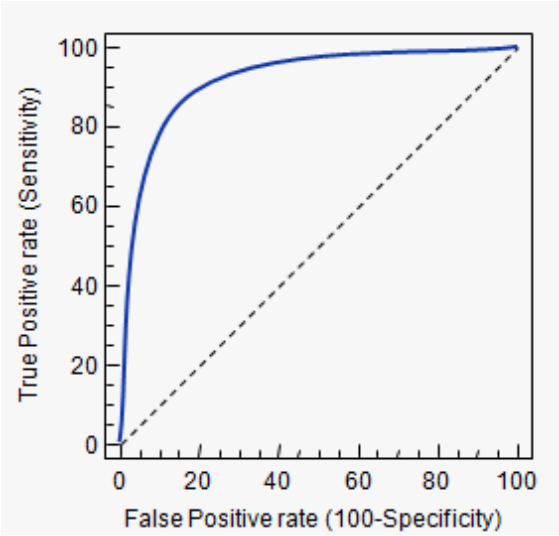


Figure 2.6: ROC curve example, from [4].

It consists in plotting the sensitivity according to the False Positive Rate (FPR) (1-specificity) for different cut-off points of a parameter [32]. Each point on the ROC curve represents a sensitivity/specificity pair corresponding to a particular decision threshold. The ROC curve shows how the number of correctly classified positive examples varies with the number of incorrectly classified negative examples [33]. A test with perfect discrimination (no overlap in the two distributions) has a ROC curve that passes through the upper left corner (100% sensitivity, 100% specificity). Therefore the closer the ROC curve is to the upper left corner, the higher the overall accuracy of the test [34].

A possible classifier is the Support Vector Machine (SVM), a powerful technique for general (non-linear) classification, regression and outlier detection with an intuitive model representation. SVM was developed by Cortes and Vapnik (1995) for binary classification. Their approach may be roughly sketched as follows:

- **Class separation**: basically, we are looking for the optimal separating hyper-plane between the two classes by maximizing the *margin* between the classes closest points (Figure 2.7)- the points lying on the boundaries are called *support vectors*, and the middle of the margin is our optimal separating hyperplane;
- **Overlapping classes**: data points on the "wrong" side of the discriminant margin

are weighted down to reduce their influence;

- **Non-linearity**: when we cannot find a linear separator, data points are projected into an (usually) higher-dimensional space where the data points effectively become linearly separable (this projection is accomplished via kernel techniques);

- **Problem solution**: the whole task can be formulated as a quadratic optimization problem which can be solved by known techniques.

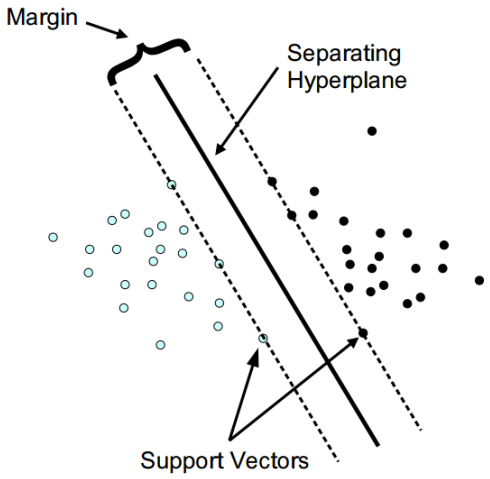


Figure 2.7: Classification for SVM(linear separable case).

An algorithm able to perform all these tasks is called a *Support vector machine* [35].

There are at least three reasons for the success of the SVM: its ability to learn well with only a very small number of free parameters, its robustness against several types of model violations and outliers, and last but not least its computational efficiency compared with several other methods (e.g. Logistic regression) [36]. As for disadvantages, if the number of features is much greater than the number of samples, the method is likely to give poor performance. Also SVM do not directly provide probability estimates, these are calculated using five-fold cross-validation, and thus performance may suffer [37].

Besides SVM another method that is very used in classification is the Naïve Bayes classifier. Naïve Bayes classifier is a supervised learning algorithm based on applying Bayes theorem with the "naïve" assumption of independence between every pair of features. Bayes' rule says that if you have a hypothesis H and evidence E that bears on that hypothesis, then:

$$P[H|E] = \frac{P[E|H]P[H]}{P[E]} \tag{2.6}$$

where $P[A]$ denotes the probability of an event A and $P[A|B]$ denotes the probability of A conditional on another event B. The evidence E is the particular combination of attribute values. Let's call n pieces of evidence $E_1, E_2, ..., E_n$ respectively. Assuming that

these pieces of evidence are independent (given the class), their combined probability is obtained by multiplying the probabilities:

$$P[H|E] = \frac{P[E_1|H] \times P[E_2|H] \times ... \times P[E_n|H] \times P[H]}{P[E]} \tag{2.7}$$

This method goes by the name of Naïve Bayes because it is based on Bayes' rule and "naïvely" assumes independence [30]. These classifiers have worked quite well in many real-world situations, such as document classification and spam filtering. They require a small amount of training data to estimate the necessary parameters.

Naïve Bayes classifiers can be extremely fast compared to more sophisticated methods. The decoupling of the class conditional feature distributions means that each distribution can be independently estimated as a one dimensional distribution. In turn this helps to alleviate problems stemming from the curse of dimensionality [38]. However, there are many datasets for which Naïve Bayes does not do well. Because attributes are treated as though they were independent given the class, the addition of redundant ones skews the learning process [30].

2.5.3 Feature Extraction

There are many features/parameters that can be used as input variables in our problem. Besides the current algorithm output in section 2.4 the following features were also computed:

- **Zero Crossing Rate** – It is defined as the number of time-domain zero crossings within a defined region of signal, divided by the number of samples of that region [39]. The zero crossing process consists in counting the number of times that the signal changes sign, meaning, it counts when the signal passes from negative to positive and from positive to negative.

- **Standard Deviation** – The standard deviation is equal to the square root of the variance and measures how much variation exists from the signal average. A small value of standard deviation indicates that the points tend to be very close to the average, whereas a high value that the points are very spread out and more apart from the average. Considering a signal defined over a finite time window with length N, and represented as time series $[x(n)]$, the standard deviation σ can be represented using the average μ [40]:

$$\sigma = \sqrt{\frac{1}{N}\sum_{n=0}^{N-1}[x(n)-\mu]^2} \quad \text{where} \quad \mu = \frac{1}{N}\sum_{n=0}^{N-1}[x(n)] \tag{2.8}$$

- **Histogram** – Given an univariate sample $S = x_1, x_2, ...x_n$, this one can be processed to form a histogram and thereby gain insight into the distribution of the data. Let

χ be the set of possible distinct values in S. For each $x \in \chi$ the relative frequency is:

$$f(x) = \frac{\text{the number of} \quad x_i \in S \quad \text{for which} \quad x_i = x}{n} \tag{2.9}$$

A discrete-data histogram is a graphical display of the relative frequency where each distinct value in the sample appears [41]. One possible parameter to extract from the histogram is the maximum number of occurrences which represents the maximum value of the numerator in equation 2.9.

- **Periodogram** – Is based on the definition of the Power Spectral Density (PSD) as seen in equation 2.10. One of the first uses of the PSD, has been in determining possible "hidden periodicities" in time series, which may be seen as a motivation for the name of this method [42, 43]. A possible parameter to extract from the PSD is the maximum power spectral density which represents the maximum value from equation 2.10.

$$P_{xx}(f) = \frac{1}{N} |\sum_{k=0}^{N-1} X_k exp(-2j\pi rk/N)|^2 \tag{2.10}$$

where N is the number of examples, and $\sum_{k=0}^{N-1} X_k exp(-2j\pi rk/N)$ the DFT already defined in equation 2.1.

2.5.4 Performance Measures

Performance tests are used to validate machine learning models and algorithms. A possible statistical test is *leave one out*; for a given dataset of m instances, only one instance is left out as the validation set (instance) and training uses the $m - 1$ instances. We then get m separate pairs by leaving out a different instance at each iteration. The results of all m judgements, one for each member of the dataset, are averaged, and that average represents the final error estimate.

This procedure is an attractive one for two reasons. First, the greatest possible amount of data is used for training in each case, which presumably increases the chance that the classifier is an accurate one. Second, the procedure is deterministic: no random sampling is involved. There is no point in repeating it 10 times, or repeating it at all: the same result will be obtained each time. Set against this is the high computational cost, because the entire learning procedure must be executed m times and this is usually infeasible for large datasets. Nevertheless, *leave-one-out* seems to offer a chance of squeezing the maximum out of a small dataset and getting as accurate an estimate as possible [30, 31].

Another statistical measure is the correlation coefficient which is a numerical value that indicates the degree and direction of relationship between two variables; the coefficients range in value from $+1$ (perfect positive relationship) to 0 (no relationship) to -1 (perfect negative or inverse relationship) [44].

Often in the study of behavioural ecology, and more widely in science, we require to statistically test whether the central tendencies (mean or median) of 2 groups are different from each other on the basis of samples of the 2 groups [45].

A used statistical test is the Mann-Whitney U Test which is a non-parametric test that can be used in place of an unpaired t-test. It is used to test the null hypothesis that two samples come from the same population (i.e. have the same median) or, alternatively, whether observations in one sample tend to be larger than observations in the other [46].

3

Current Algorithm Evaluation

In this chapter the data improvements of MOBS before applying the current algorithm are presented. The zebrafish behaviour are analysed using the platform Open Signals that will enable synchronism between video and signal. The thresholds used in the current algorithm are also tested and new suggestions are made regarding the usefulness of the algorithm.

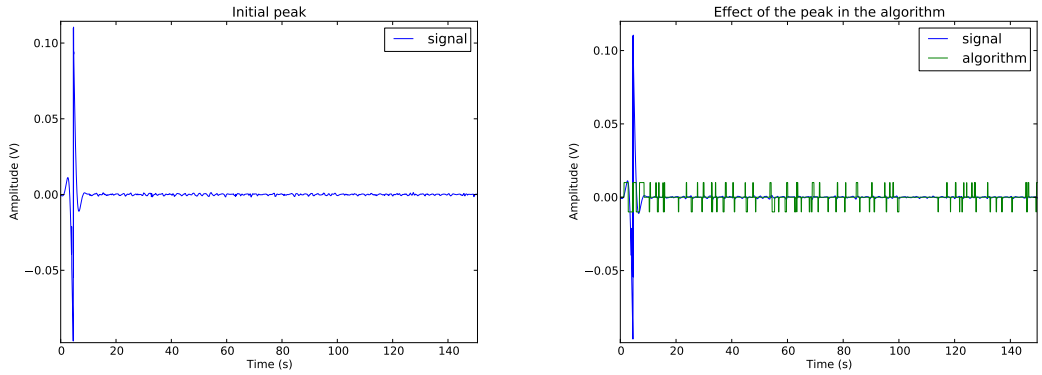
3.1 Preparing the Data

3.1.1 Start Peak

After starting the main device to visualize the fish locomotion, it is noticed in the time domain, an initial peak of higher amplitude than the fish activity. This peak is characteristic of the main device. Following this peak the fish activity is measured. The delay from the main device until the fish activity is displayed is approximately 30 seconds, and considering this, the current algorithm contained only the analysis of the signal after 30 seconds. However it was noticed that the peak was still present - Figure 3.1(a). The presence of this peak certainly changes the algorithm output as seen in Figure 3.1(b).

This situation was solved by using the algorithm furthermore in the signal. Given that, instead of considering 30 seconds before the analysis, the algorithm only acts in the signal after 40 seconds. This guarantees that the initial peak is not presented, and that the evaluation of the algorithm is not corrupted by this peak. The result is shown in Figure 3.2.

These changes will contribute with two possible variations in the current algorithm output:



(a) Initial peak from the main device

(b) Algorithm effect

Figure 3.1: Initial peak from the main device and its effect in the algorithm output.

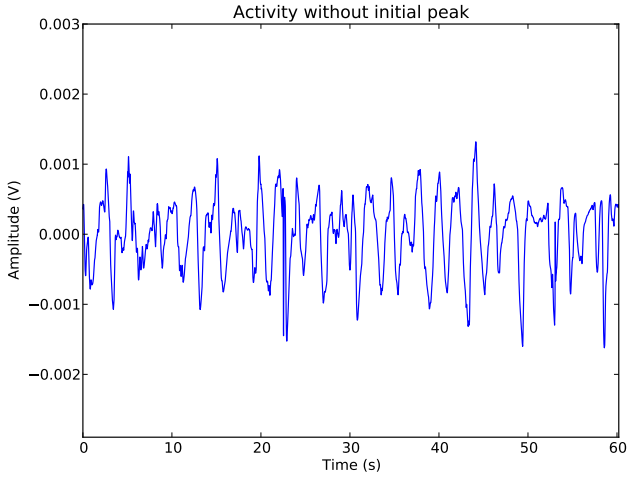


Figure 3.2: Signal without the initial peak from the main device.

- Increase of the algorithm output - tail-flips per minute. This situation happens due to the standard deviation that decreases because of the absence of the initial peak. Given that, more peaks from the derivative will be detected as tail-flips. It is important then to ascertain that the threshold used to allow the behaviour detection in the algorithm, the multiplicative factor (see section 2.4), is in fact the correct one to detect the tail-flips.

- Decrease of the algorithm output - tail-flips per minute. This happens because the transitions detected by the algorithm from this initial peak are no longer counted - Figure 3.1(b). Consequently the number of tail-flips decreases.

It is noticed most often an increase in the algorithm output, meaning that there is a higher number of transitions due to the standard deviation decrease, than the number of transitions removed from the initial peak. One of the disadvantages of taking more time to remove this peak is the time precision that the user wants to maintain; however is of

greater importance the absence of this peak in the algorithm evaluation.

3.1.2 Error Peaks Detection

Another difficulty related to the main device occurs during the recording of the fish activity. It was noticed in the time domain the presence of peaks with much higher amplitude than the fish activity - Figure 3.3.

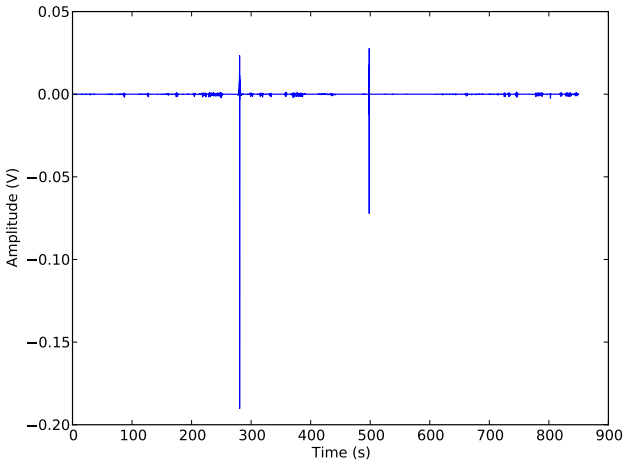
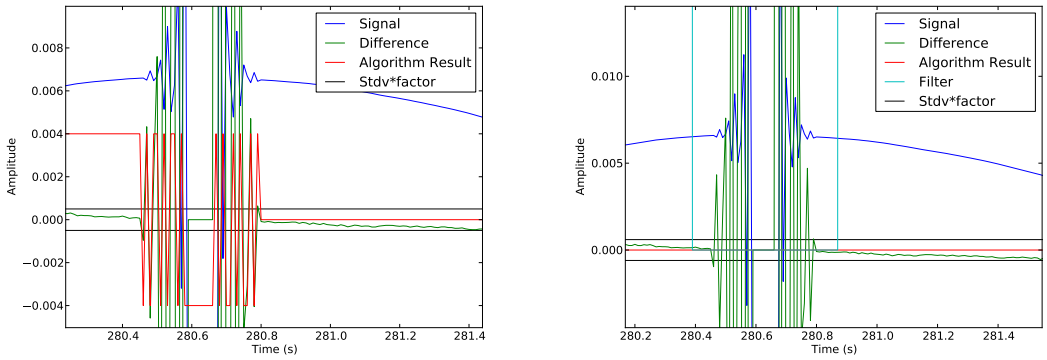


Figure 3.3: Artefacts of the main device or software with higher amplitude than the amplitude of the fish activity.

Since we can record more than one chamber at the same time, it was possible to visually identify these peaks in each chamber at the same time. Given that, we can say that the problem was not from one chamber in particular, but from the main device itself or from the computer software. The impact of these peaks on the results is well noticed in Figure 3.4(a).

The idea to solve this problem was by the application of a filter. The fact that this peak is of higher amplitude than the fish activity, turns it easy to identify. Then for the filter process, it is used 0 values when those peaks are detected and 1 values otherwise. In the end the filter is multiplied with the signal to exclude these peaks for further analysis. The filter result is shown in Figure 3.4(b).

Again, because these peaks are not included in the algorithms behaviour detection, the standard deviation will decrease and the multiplicative factor needs verification. However is not noticeable an increase in the algorithm output as in the previous section but a decrease. This happens because there was a higher number of transitions removed by these error peaks, than the number of transitions added from the decrease of the standard deviation.



(a) Algorithm effect

(b) Algorithm with filter

Figure 3.4: Artefacts of the main device, its effect in the algorithm result with and without the filter. Signal enhanced from 3.3.

3.2 Synchronism

The signal in the time domain is delayed in relation to the instant of acquisition start. This delay is caused by the main device. Given that, it is difficult to compare a video where the fish movements are present, with its respective signal from MOBS.

3.2.1 Open Signals

The Open Signals is a platform designed and programmed by *PLUX - Wireless Biosignals*, S. A. It is a useful tool for this research, because it will allow synchronism between signal and video.

Using Open Signals, synchronism is possible with a visible stimulus in the signal and video. This stimulus must be sufficient to not be confused with the fish activity as shown in Figure 3.5. A touch in the chamber is a possible stimulus and to not corrupt the signal from the fish activity for further analysis, this stimulus should be produced at the end of the recording.

With this platform, the user can navigate freely through the signal and video independently (without both being synchronized yet). The synchronism is accepted after the user locks both signal directly in the window and video using the lock button (Figure 3.5). After the right time is selected in accordance to the stimulus made, it will be possible to analyse the signal variations in comparison to the fish movements in the video. Navigating in one datum will automatically progress the other in the same way allowing the study of their behaviour more precisely.

3.2.2 Time Precision

A test was made to access the main device time precision. Behavioural tests lasts 15 minutes. It was then decided, using Open Signals, to perform a precision test for 30

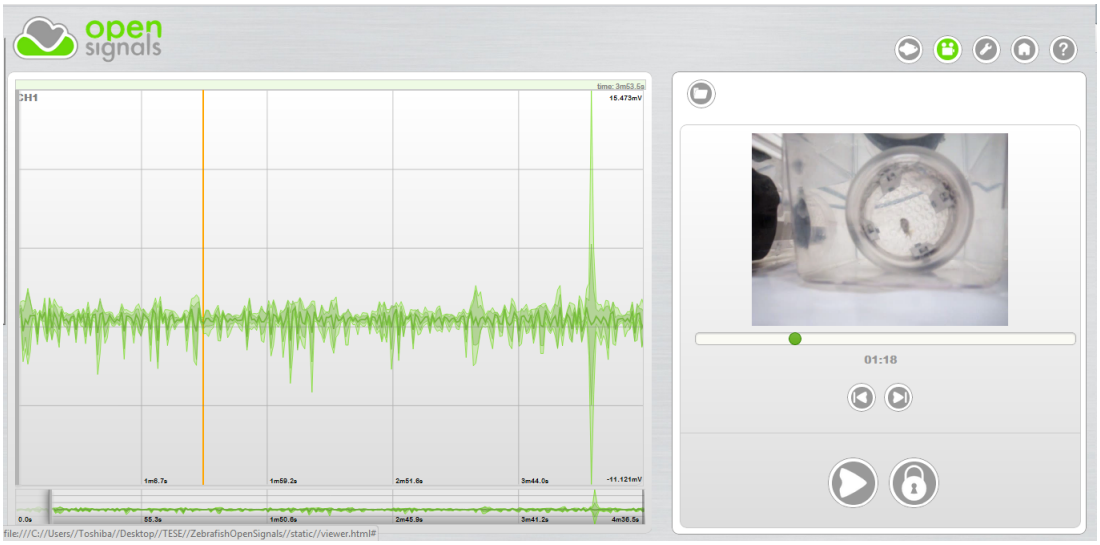


Figure 3.5: Platform Open Signals for synchronism between signal and video.

minutes with the empty chamber submersed in water. In this 30 minutes several stimuli were made in the chamber and recorded in video. After synchronism it was verified that each stimulus in the signal corresponded at the same moment in the video (variation of 0.13 ± 0.05 seconds between the stimulus identified in the signal and video).

Hence, it is possible to make behavioural tests for 30 minutes efficiently since for at least this length of time we know that the main device is precise.

3.2.3 Experimental Design

This subsection presents the experimental design performed with zebrafish. These tests will allow the study of their behaviour using the synchronism between video and signal. Since the drug that simulates PD leads to a decrease in the fish activity [17, 18], it is also intended to analyse by eye the tail-flip movements when the fish are submitted to the drug 6-OHDA.

3.2.3.1 Test Animals and 6-OHDA

The zebrafish (*D. rerio* Hamilton 1822) strain used for this work was the AB line (Zebrafish Facility, IMM, Portugal). Animals were maintained under standard conditions and experiments were approved by the Institutional Animal Care and Use Committee. A master stock solution of 6-hydroxydopamine hydrochloride (6-OHDA, Sigma-Aldrich, USA) was prepared in 0.2% ascorbic acid solution (analytical grade, Sigma) and stored at -20°C. This stock solution was used to prepare all working solutions in experiments with zebrafish.

3.2.3.2 Behaviour Assay

Before the experiments, small groups of female fish (24 animals, body weight 0.5 ± 0.05 g) were acclimatized to the experimental testing conditions (temperature 22 °C $\pm$ 1 °C, 10 h:12 h light-dark cycle) in 17 litre glass aquaria under static conditions and for a minimum of one week. Food was not provided 24 h before or during the experiments. The behaviour analysis was divided into two groups: non-treated (12 fish) and for that considered as "healthy" fish in which no injection was administered, and treated (12 fish) also considered as "ill" or less active where $5\mu L$ of 6-OHDA was injected via intramuscular. During the injection they were in a medium-to deep-plane level of anaesthesia (tricaine $50mg/L$) and had lost their reflex responses and muscular control. Afterwards they returned to their original test chambers and allowed 30 min to recover from the anaesthesia.

On the day of experiments, either the treated or non-treated groups of fish were placed individually in the test chambers supplied with oxygenated tap water (22 °C $\pm$ 1 °C). Fish were acclimated to the test chambers for 30 min and then individual baseline responses were monitored using MOBS and video recording (at 25 frames per second) for five minutes between 10 and 12 a.m.

After behavioural recording, treated fish were sacrificed with tricaine. The behavioural experiments were always performed by the same experimenter.

3.2.3.3 Behaviour Detection

Using video recording it is possible to distinguish tail-flip movements. This behaviour is characterized by abrupt and fast changes of fish direction which imply strong burst in the fish tail (Figure 3.6).

Figure 3.6: Abrupt tail-flip movement.

3.2.4 Visual Analysis

A visual and detailed analysis was made with the Open Signals platform using video frame by frame with both signals synchronised taken in consideration the behaviour tail-flip.

To simplify the analysis, it was created a function that received the signal and the instant where the behaviour was detected with a time precision of 0.01 seconds. After all

the detections, this information was saved in a file in the following order: time; signal; behaviour detection. The result is presented in Figure 3.7.

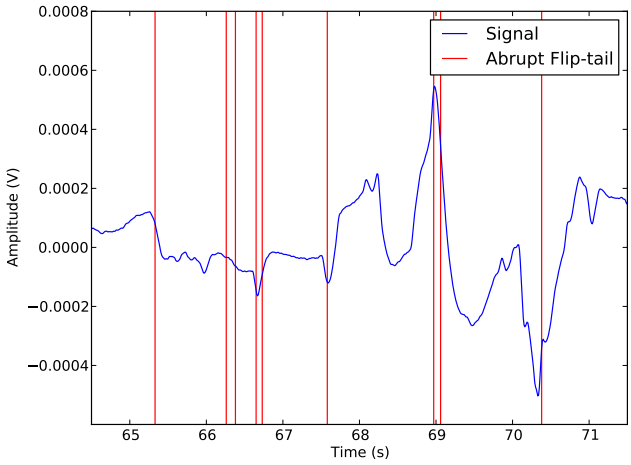


Figure 3.7: Visual analysis example. The signal is represented in blue and the behaviour tail-flip detection in red.

Since the actual algorithm output returns the number of abrupt tail-flips per minute, we can now compare it with the visual analysis. The process is as simple as count the number of abrupt tail-flips visually detected in the created file and divide it by the total signal time in minutes. Then compare it with the value of the algorithm output. This may bring an idea of how far we are from reality.

3.2.5 User Test/Visual Analysis Validation

Since visual analysis depends on the user that is interpreting the data, it is important to test other users and compare the results. Therefore, a visual test using a different user was made, providing only the description and images explained in section 3.2.3.3. Figure 3.8 shows the detection for both users.

The test consisted in a precise analysis frame by frame using a signal of 30 seconds, and for this time both users detected 46 abrupt tail-flips. After User 1 had detected the abrupt tail-flip it was considered an interval of 0.25 seconds in which the User 2 had also to detect the same abrupt tail-flip to be a valid success. Given that, in 46 detections, 44 were accepted, leading to an error of 4.35% between both users.

The agreement between both users classifying the behaviour, implies that the visual result may be a valid information to be compared with the actual algorithm or to be used in future works.

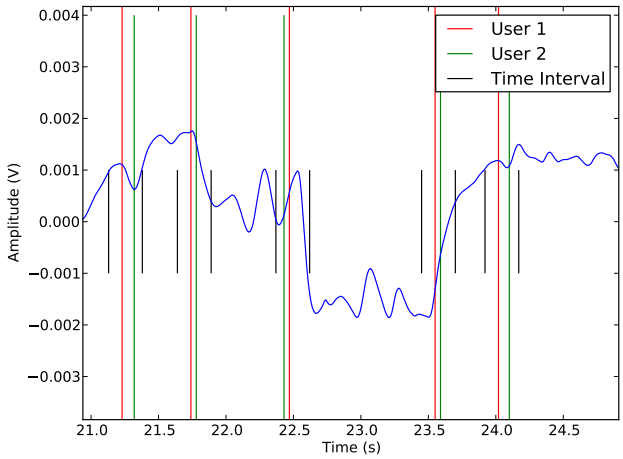


Figure 3.8: User test. The signal is represented in blue, User 1 is represented in red and User 2 in green. The time interval accepted is in black.

3.3 Thresholds

This section will allow an improvement in the thresholds already implemented in the current algorithm, specifically in the maximum and minimum amplitude accepted for the fish activity. The multiplicative factor is analysed in the next section. Several tests were performed and based on the results, new considerations were made, as following:

- **Minimum Amplitude** – The threshold used to limit the minimum amplitude for the fish activity and therefore the maximum amplitude for the noise is 0.5 mV. Tests without fish and with the chambers submersed in the water were performed. Afterwards the maximum amplitude for each test was measured. The maximum amplitude encountered was 0.6 mV, leading to a variation of 0.1 mV from the previous threshold.

- **Maximum Amplitude** – The threshold used to limit the maximum amplitude of the fish activity is 0.01 V. Tests performed with fish, showed that the maximum amplitude measured from all chambers was the same. Given that, no change was made.

3.4 Algorithm Evaluation

This section intends to compare the visual analysis with the algorithm output. The result is shown in Figure 3.9 where linear regression was applied for each group (treated and non-treated).

Figure 3.9 shows that there is no direct relation between the visual analysis and the algorithm output as it would be expected both for treated and non-treated fish. The next

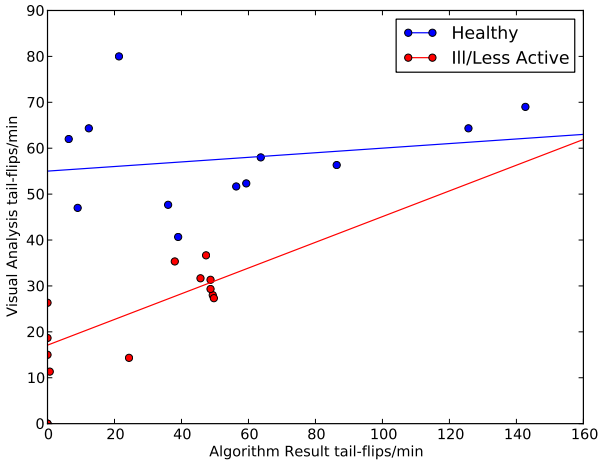


Figure 3.9: Comparison between the visual analysis and the algorithm output both in number of tail-flips per minute. Linear regression is presented for each group and relative error was estimated with the *leave one out* method.

subsections demonstrate the validation for each group and the error associated will show the need for improvement in the current algorithm, concretely in the multiplicative factor.

3.4.1 Validation for healthy fish

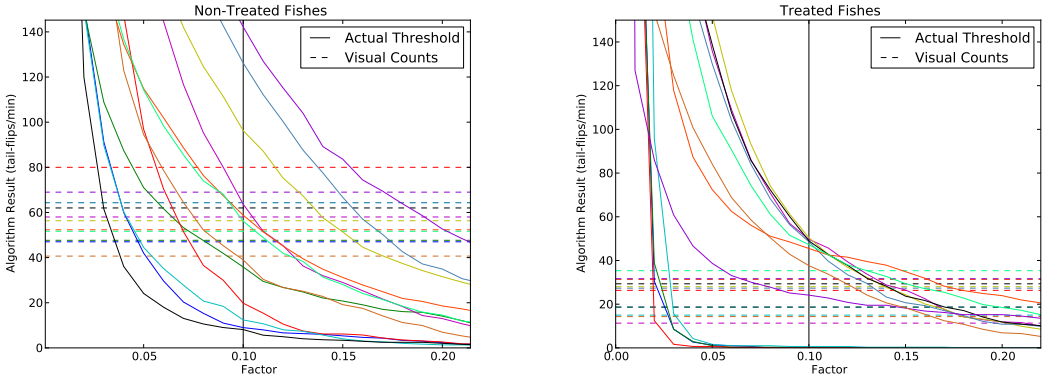
For validation it the statistic method *leave one out* was used. This was chosen because the number of points analysed is small ($n = 12$). The process was: take one point out, obtain the linear regression with all the others points, and measure the expected tail-flips of the point that was excluded using the calculated linear regression. The relative error of the respective point consists in the difference of its real value (the tail-flips obtained visually) with the expected value divided by the real value. Then it is necessary to repeat this process to all points, meaning, there will be as much relative errors as the numbers of points used. In the end, all relative errors are averaged. The non-treated group has a relative error of 17.29% using a window of 180 seconds (Figure 3.9) and a correlation coefficient of 0.015. More points can be provided with the usage of a smaller window, and this was accomplished using windows of 60 seconds which resulted in an error of 19.34% and a correlation coefficient of 0.014. Given that the relative error is higher, the validation will use the analysis for a window of 180 seconds.

3.4.2 Validation for ill fish

Again, for the treated group it was used the statistic method *leave one out*, which resulted in an error of 25.31% for a window of 180 seconds and a correlation coefficient of 0.76. The elevated error values and the poor correlation coefficient implies that the algorithm should be improved. The next subsection presents a more detailed study of the multiplicative factor.

3.4.3 Multiplicative factor

The multiplicative factor in the algorithm is used so that the derivative can be comparable to the standard deviation thus allowing the behaviour tail-flip to be detected. Given that, to improve the algorithm, the multiplicative factor should be analysed. Also, after the studies made in the previous sections, it was said that this factor needed verification (see section 3.1). The value used so far has been 0.1. To facilitate we vary the factor according to the algorithm output as shown in figures 3.10 and compare it with the visual result. The factor is analysed from 0 to 0.25 with a variation of 0.01.



(a) Non-treated fish group.

(b) Treated fish group.

Figure 3.10: Multiplicative factor effect over the algorithm output. Visual analysis is applied for each case in dotted lines to understand which multiplicative factor is the most suited.

Focusing on a particular case (red analysis in Figure 3.10(a)) it is visible that the actual threshold used (0.1) leaded to a result that was different from the visual analysis, indicating that in this case, the factor that should be used is not 0.1 but in fact 0.08 approximately. With the analysis of more cases, it was expected to find an approximate factor value for all cases or a direct association. Unfortunately this did not happen either for non-treated and treated groups, in that, there are different factors that suit the actual algorithm according to each case. However it is visible that if there is an ideal multiplicative factor, the one should probably be between 0 and 0.25.

To reinforce this study, table 3.1 demonstrates the specific values obtained for each group. In these tables the visual results obtained are shown as well as the algorithm output using the current multiplicative factor (0.1). These tables demonstrate that there are substantial differences between the visual analysis and the algorithm output.

The intention of the next analysis is to be able to understand which multiplicative factor is the most suited to be used for the detection of the behaviour tail-flip and its respective relative error. The process was to subtract each value of the curves in figures 3.10 by its respective visual result and divide it by the visual result to provide a relative

Table 3.1: Specific values from figures 3.9, namely the visual analysis result and the algorithm output using the actual multiplicative factor (0.1).

(a) Non-treated fish group

Visual Result (tail-flip/min)	Algorithm Output (tail-flip/min)
40.67	38.980
47	8.993
47.667	35.97
51.67	56.133
52.333	58.768
56.333	86.044
58	63.653
62	6.295
64.333	126.144
64.333	12.332
69	142.191
80	21.324

(b) Treated fish group

Visual Result (tail-flip/min)	Algorithm Output (tail-flip/min)
11.333	0.654
14.333	24.143
15	0
18.667	0
26.333	0
27.333	50.401
28	49.853
29.333	48.559
31.333	48.976
31.667	45.497
35.333	37.639
36.667	47.252

error. In the end all curves analysed are averaged and the result is shown in Figure 3.11 for each group. Here is presented the minimum error accepted as well as the error used with the actual factor for each group.

The error using the actual factor is 55.26% and 68.79% for non-treated and treated groups respectively, and even improving the factor, the minimum error accepted would be 53.20% for non-treated group which leads to a best factor of 0.11 and 44.53% for treated group with a best factor of 0.13. To be able to choose the best factor these obtained errors should be as close to zero as possible which indicates that even with these improvements the best multiplicative factor cannot be certain to characterize the behaviour as close to reality as it is pretended.

Because the user analysis has already been tested, and thus, considering that the visual analysis is a valid measure, there are two possible reasons to explain these high errors: the algorithm or the biosensor MOBS.

3.4.3.1 Algorithm Insight

The algorithm output consists in the peaks detection of the derivative using a given threshold so that the behaviour tail-flip can be detected. This threshold is represented by the standard deviation with a multiplicative factor so that the standard deviation may be comparable with the derivative.

The main problem verified is that the abrupt tail-flips detected visually do not always show the same characteristic in the signal, and consequently, an abrupt tail-flip detected visually not always imply a representative peak in the derivative. Figure 3.12(a) shows that case.

Also there are peaks from the derivative that were detected as abrupt tail-flips by the algorithm but visually were not verified - Figure 3.12(b). This justifies more clearly the disagreement between the algorithm behaviour detection and the visual analysis.

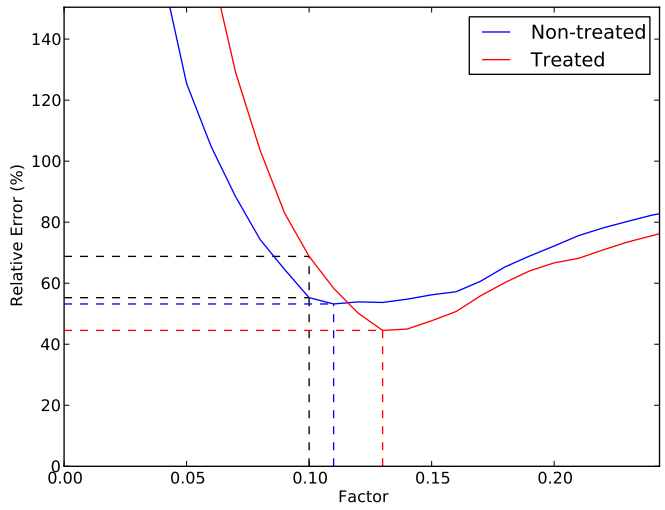
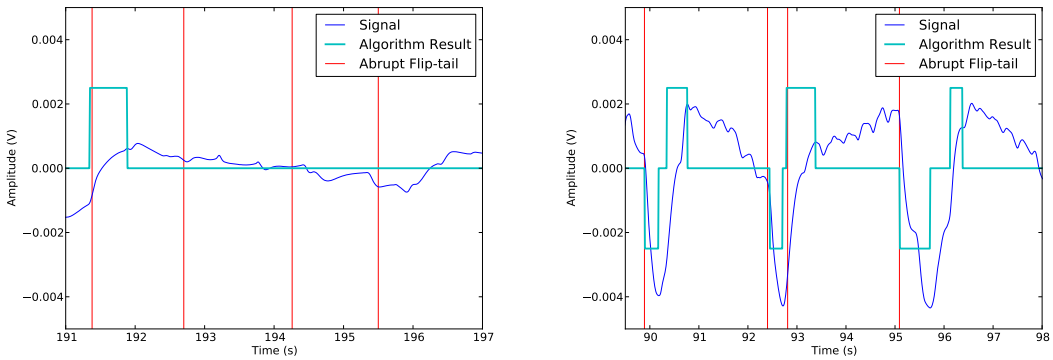


Figure 3.11: Relative error in percentage of the visual analysis and the algorithm output to understand which multiplicative factor is most suited for each group by minimizing its relative error. The black dotted lines represent the actual multiplicative factor (0.1), the red dotted lines the best multiplicative factor for treated fish and the blue dotted lines the best multiplicative factor for non-treated fish.



(a) Behaviour detection visually identified but not from the algorithm.

(b) Behaviour detection from the algorithm but not visually identified.

Figure 3.12: Relation between signal, visual analysis, and algorithm effect. The signal is represented in blue, the algorithm in cyan and the visual marks in red.

Therefore it is suggested the development of a new algorithm that can characterize the behaviour as close to reality as possible.

3.4.3.2 Biosensor MOBS

If a new algorithm cannot be implemented to provide better results in the behaviour characterization, then it is suggested that the problem is in the biosensor MOBS. Thus, it is proposed an improvement in this equipment before the implementation of new studies.

4

Proposed Algorithm

In this chapter new parameters are discussed to characterize the abrupt tail-flip movements. With the visual analysis obtained from the previous chapter it will be possible to study new parameters using supervised learning methods, more precisely, regression models. Thus our visual analysis will be considered as the output variable, and the new parameters the input variables. It is also shown the need for classification between "healthy" and "ill" fish. Finally, a new algorithm is proposed as well as its integration in the Open Signals platform.

4.1 Behaviour Characterization

To be able to characterize the behaviour in number of tail-flips per minute, the parameter zero crossing rate proved to be useful. This parameter is defined as the number of time-domain zero crossings within a defined region of signal, divided by the number of samples of that region [39]. The zero crossing process consists in counting the number of times that the signal changes sign, meaning, it counts when the signal passes from negative to positive and from positive to negative. Each data was divided by its standard deviation, so that, all data is at the same scale to be comparable and because the signal is centred at zero, it was not necessary to subtract its average. Also the signal was smoothed using a Hanning window with a length of 0.05 seconds. The comparison between the visual analysis and the zero crossing rate for each group is shown in Figure 4.1 with their respective linear regressions.

This parameter presents a direct relation with the visual analysis both for treated and non-treated groups. The next subsections will validate this parameter using the statistic method *leave one out*.

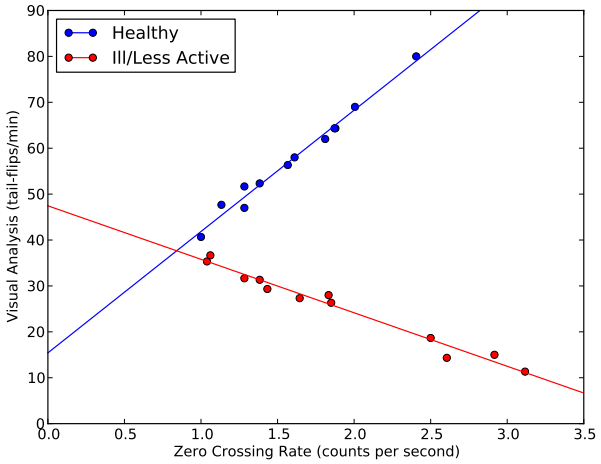


Figure 4.1: Comparison between the visual analysis and the zero crossing rate parameter. Linear regression is presented for each group and relative error was estimated with the *leave one out* method.

4.1.1 Validation for healthy fish

To validate this parameter the statistic analysis *leave one out* was used. This was chosen because the number of points analysed is small ($n = 12$). For the non-treated group in Figure 4.1, the relative error obtained was 2.55% for a window of 180 seconds and 12.08% for a window of 60 seconds.

Again, the idea to use smaller windows is to provide more points for validation, however the relative error increases. Hence it will be considered the window of 180 seconds. The relative error of 2.55% compared with the 17.29% from the previous algorithm can be considered as an excellent improvement.

The user test from the previous chapter (see subsection 3.2.5) showed an error of 4.35%. Given that, the reason why this parameter shows a smaller error (2.55%) it is because it suits the user that performed this analysis. If User 2 had also performed these analyses, a bigger error should be expected.

The correlation coefficient obtained in this case was 0.99, indicating that there is a very good positive relation between the zero crossing rate and the visual analysis. Finally using all points for a window of 180 seconds, linear regression can be applied to define our hypothesis:

$$h_\theta(x) = 15.42 + 26.43x \tag{4.1}$$

where x represents the signal zero crossing rate in counts per second, and $h_\theta(x)$ the expected output of the fish activity in number of tail-flips per minute. This means that 15.42 tail-flips per minute is the minimum activity that this parameter can detect for a "healthy" fish. If no more changes had to be done, the new algorithm would provide the behaviour characterization of a new signal in number of tail-flips per minute by simply

measuring its zero crossing rate and applying it on the equation 4.1.

4.1.2 Validation for ill fish

For the treated group – Figure 4.1, it is visible that using the parameter zero crossing rate the "ill" fish do not follow the same tendency as the "healthy" fish, meaning, if we apply the hypothesis already defined in equation 4.1, the fish that were exposed to the drug would not show a decrease in their activity as seen visually. In fact, the ones that present lower levels of activity visually would provide higher values of activity after using the hypothesis 4.1. Thus, it is necessary to have a classifier that can distinguish between a "healthy" fish from one that is "ill".

After a successful classification it is relevant to characterize the behaviour for "ill" fish to provide the number of tail-flips per minute as made with the "healthy" fish. Figure 4.1 shows that the "ill" fish present an inverse linear tendency between the zero crossing rate and visual analysis, which means that the higher the number of counts per second from the zero crossing rate parameter, the less active the fish is.

Again it was used the *leave one out* method to validate this parameter. The relative error obtained was 5.75% which can be a good estimative even though it is higher than the error obtained to characterize "healthy" fish (2.55%). This error compared to the 25.31% from the previous algorithm can also be considered as an excellent improvement. The correlation coefficient was -0.99, meaning there is a very good inverse relation between the visual analyses and the zero crossing rate.

Using all points for a window of 180 seconds, linear regression can be applied to define our hypothesis:

$$h_\theta(x) = 47.45 - 11.65x \tag{4.2}$$

where x represents the signal zero crossing rate in counts per second, and $h_\theta(x)$ the expected output of the fish activity in number of tail-flips per minute. The negative slope represents the inverse relation between the visual analysis and the zero crossing rate. The value of 47.45 tail-flips per minute limits the fish activity, which means that "ill" fish will not show a higher value of activity than 47.45 tail-flips per minute. Also for a fish that does not present any activity (0 tail-flips per minute) it should be expected a value of 4.07 counts per second.

Given this analysis it should be understood the signal physiology for different groups of fish with the same value of the zero crossing rate, for example, considering two fish from different groups with a zero crossing rate of 4.07 counts per second (therefore the "ill" fish does not show any activity). The assumption for the signal interpretation is that the zero crossing rate for a "ill" fish is only considering ventilation which presents a high frequency and a low amplitude [2]. As for a "healthy" fish, using the hypothesis from equation 4.1 it is expected an activity of 122.99 tail-flips per minute. This is a very active

fish and the signal can be explained with the consecutive bursts that may bring a high frequency and a high amplitude to the signal. To confirm this hypothesis it would be necessary to separate in this specie, ventilation from locomotion; however that cannot be possible due to the high activity of the zebrafish.

It is also convenient to justify the intersection between both curves. This intersection is verified for an activity of 37.65 tail-flips per minute and a zero crossing rate of 0.84 counts per second. Given that, from 0 to 0.84 counts per second, there remains the possibility that a "ill" fish may present higher activity than a "healthy" one. The assumption is that there might be fish that react differently to the drugs, and therefore those fish continue to present high activity, even though it is not expected. The other way around may also be justified: a "healthy" fish may present itself as less active even though is not submitted to any drug. This is the reason why the curves are not cut at this intersection and only when they present an activity of 0 tail-flips per minute.

Besides the zero crossing rate, other parameters were tested, however they presented higher relative errors when submitted for validation. Still, there was the possibility to merge other parameters with the zero crossing rate. The idea was to find a parameter that, besides having an elevated relative error, when merged with the zero crossing rate, could complement areas of the zero crossing that presented higher variations. Therefore the final relative error could be minimized. This study was taken in consideration, however not successfully achieved, because a parameter that could fit this need was not found.

4.2 Classification

The previous section showed the need to create a classifier that could distinguish between "healthy" and "ill" fish. Now our output is defined by two classes: "healthy" and "ill" (less active) fish. The *Orange* is a comprehensive, component-based software suitable for machine learning and data mining. It is a free software and open source. It allows to use data mining through visual programming or *Python* scripting [8].

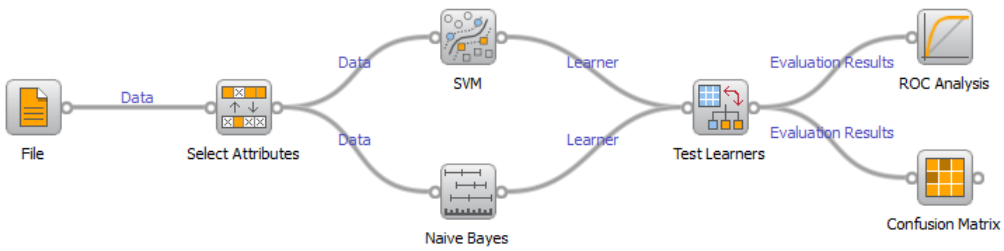


Figure 4.2: Classifier scheme in the Orange Software.

Figure 4.2 shows the classifier design using the Orange software. First it is necessary to organize the file (in a .tab format) according to the Orange specifications. In the file

we need to provide the parameters results as well as the class that they belong (if it is treated or non-treated). The parameters used in this analysis were the zero crossing rate, the standard deviation, the maximum number of occurrences using the histogram, the maximum power spectral density using the periodogram and the previous algorithm output (see section 2.5.3). Then we give the possibility to choose the parameters with which we want to construct the classifier (Select attributes in Figure 4.2). Afterwards we build the classifier with the chance to use different methods. The ones used were SVM and Naïve Bayes. The Test learner widget will then provide the accuracy, sensitivity and specificity for each method used (SVM and Naïve Bayes). Thus, varying the number of parameters available we choose the ones that give higher accuracy for the respective method. The confusion matrix gives the number/proportion of examples from one class classified in to another (or same) class. Besides, selecting elements of the matrix feeds the corresponding examples onto the output signal. This way, one can observe which specific examples were misclassified in a certain way [8]. It is also analysed the ROC curve to reinforce the study in choosing the best classifier.

Since the classifier does not require the visual analysis as output, which is a long process, instead of using the data obtained so far (24 case studies), it was used data from a previous work to provide more points to the classifier (108 case studies with equal number for each class). This work developed at IMM provides data with non-treated and treated fish (submitted to the drug 6-OHDA).

4.2.1 Validation

The parameters used that leaded to a higher accuracy for the SVM were the zero crossing rate, the standard deviation, the maximum power spectral density using the periodogram, the maximum number of occurrences using the histogram, and the previous algorithm output. The learning options used were the Sigmoid kernel function ($\tanh(8 * x.y)$), a Cost of 2.0 (Model Complexity - penalty parameter) and a numeric precision of 0.001.

For validation it was used the *leave one out* which holds out one example at a time, inducing the model from all others and then classifying the held out. This method is obviously very stable and reliable but very slow [8].

The accuracy obtained using *leave one out* for the SVM method was 100% (with sensitivity and specificity of 100%), meaning that all cases analysed were classified correctly. The confusion matrix is presented in table 4.1 for the SVM (table 4.1(a)) and Naïve Bayes (table 4.1(b)) methods.

On the other hand, the Naïve Bayes method based on the relative frequency presents a maximum accuracy of 67.59% (with sensitivity of 70.37% and specificity of 61.11% - target class non-treated group) using the parameters standard deviation, algorithm output and maximum power spectral density with the periodogram. As presented in the confusion matrix, 35 fish were misclassified. 12 that are "ill" but the classifier predicted as "healthy",

Table 4.1: Confusion Matrix for each method used. Allows the comparison between the predicted values and the correct class.

(a) SVM

Correct Class	Predictions: Healthy	Predictions: Ill	Predictions: Sum
Healthy	54	0	54
Ill	0	54	54
Sum	54	54	108

(b) Naïve Byes

Correct Class	Predictions: Healthy	Predictions: Ill	Predictions: Sum
Healthy	31	23	54
Ill	12	42	54
Sum	43	65	108

and 23 that are "healthy" but were classified as "ill".

The ROC curve is presented in Figure 4.3 for each method as well its convex curves. It emphasizes that the SVM method is a suitable classifier to choose because its curve passes through the upper left corner (100% sensitivity, 100% specificity). The diagonal black line represents the behaviour of a random classifier. The Naïve Bayes is not a random classifier, but is not also as good as the SVM method. There could even be an area where the Naïve Bayes would behave better than the SVM, however this was not verified. Therefore, the SVM method is the most indicated classifier to choose for the construction of the algorithm.

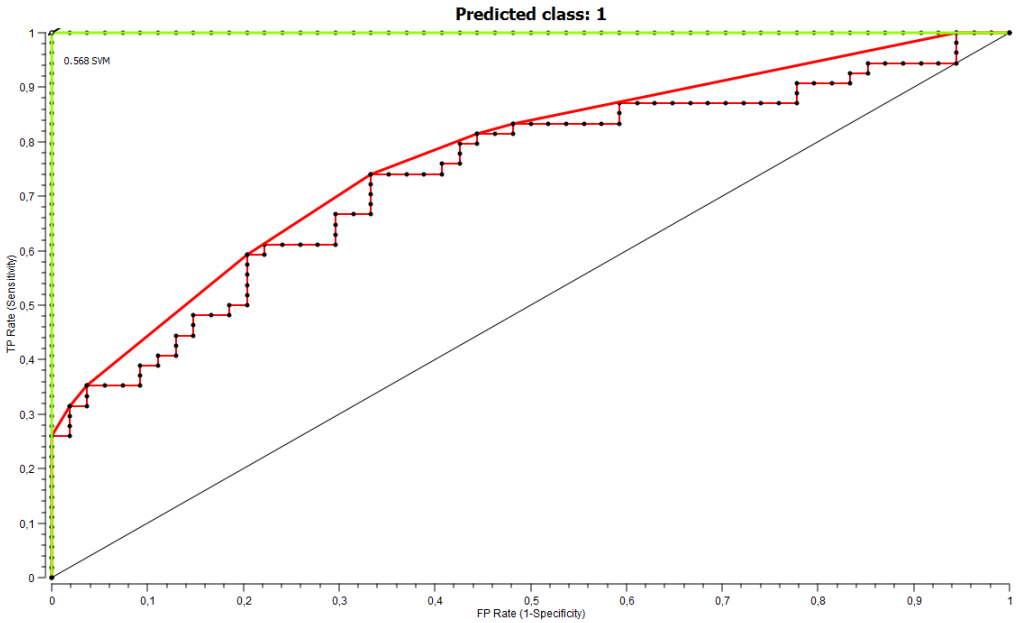


Figure 4.3: ROC curves and its convex curves for SVM (Green) and Naïve Bayes (Red) methods. Predicted class – "Healthy"

Because the *Orange* program is open source, with the access to the functions that build

the classifier SVM we can use them to construct the final algorithm in *Python*.

4.3 Final Algorithm

Now it is possible to build the final algorithm. Figure 4.4 exemplifies the process. First we prepared the data with the removal of the initial peak from the main device, the application of the filter, the normalization of the data and the signal smoothing using a Hanning window of 0.05 seconds. Then, we used the classifier to predict if the fish is "healthy" or "ill" (less active). According to the classification, it is possible to characterize the behaviour in terms of number of tail-flips per minute using the corresponding hypothesis. Each hypothesis consists in the use of the parameter zero crossing rate.

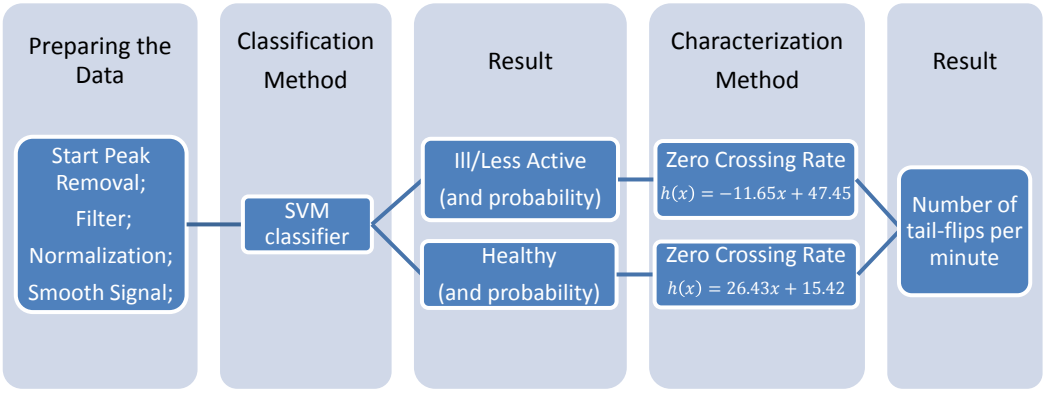


Figure 4.4: Final algorithm process.

The classification is made using the method SVM with the parameters zero crossing rate, standard deviation, maximum power spectral density using the periodogram, maximum number of occurrences using the histogram and the previous algorithm output. This classifier presents an accuracy of 100%. If the fish is classified as "ill" the parameter zero crossing rate is used to characterize the behaviour with a relative error of 5.75%. If the fish is classified as "healthy", it is also used the parameter zero crossing rate but with a different hypothesis to characterize the behaviour. This one presents a relative error of 2.92%. The final result will present the classification, the probability for that classification and the number of tail-flips per minute.

4.4 Open Signals integration

This section intends to integrate the final algorithm in the Open Signals platform in order to provide a more user-friendly method for behavioural analysis. This requires *Javascript* and *HTML* programming knowledge. Besides the algorithm, the user can also benefit

from the synchronism already implemented in this platform, hence to understand what is happening in the signal according to the fish behaviour in the video.

Figure 4.5: Open Signals with algorithm integration.

The complexity of this new algorithm may constitute a disadvantage in terms of the time spending in the evaluation of a new signal. Given that, it was taken in consideration parallel programming in the algorithm that could reduce the time from 35 seconds to 6 seconds approximately (with a Intel(R) Core(TM) i7 CPU and 8 GB RAM) for a signal of 15 minutes. The idea of parallel computing is to carry out many calculations simultaneously, operating on the principle that large problems can often be divided into smaller ones, which are then solved concurrently ("in parallel") [47]. The idea in the algorithm implementation was to programme the output for one chamber, and execute this action in parallel for all chambers used.

The final result is presented in Figure 4.5. The signal acquisition uses the *MATLAB* software, which provides a unique .txt file where the signals from all chambers are presented. The Open Signals platform was programmed to process all signals from that file. Given that, and as shown in Figure 4.5, it is possible to identify the four signals from each chamber. The algorithm output is shown in form of a table, where the first column identifies the chamber, the second column provides the classification and its respective probability and the third column the behaviour characterization in number of tail-flips per minute. To obtain the algorithm output, the user simply has to press the respective button.

The use of this platform does not require the individual installation of *Python* or the *Orange* software, only a setup to access the Open Signals functionalities. This integration allows the usage of the algorithm without requiring any knowledge in programming. Given that, any researcher is able to use this algorithm without difficulties.

5

Applications

This chapter intends to apply the new algorithm in a new case study related with PD to verify if the results are in agreement with the biological responses. Therefore, we can understand the improvements that the algorithm may need and its importance for further studies.

5.1 Parkinson's Disease

PD has no cure, but medications can help control the symptoms, often dramatically. Medications can help manage problems with walking, movement and tremor by increasing brain's supply of dopamine. The patient may have significant improvement of symptoms after beginning PD treatment. Over time, however, the benefits of drugs frequently decrease or become less consistent, although symptoms usually can continue to be fairly well controlled [48].

There is no way to measure directly neuronal loss *in vivo*, and it is unclear how clinical symptoms correlate with neuronal death [49]. Recently, Correia et. al (2012) [18] had demonstrated that the neurotoxin - 6-OHDA, induced cell loss and behavioural deficits in dopaminergic neurons of a zebrafish transgenic line Tg(-2.5th:EGFP). The behavioural alterations seen in the transgenic zebrafish were detected by using the electric biosensor (MOBS). However, the component of the MOBS that relates to the signal processing still need to be improved for a better distinction between different phenotypes. Given that, a new algorithm was developed and its application is shown in the next subsections.

5.1.1 Experimental Design

The animals used and the neurotoxin 6-OHDA are equivalent to the description in subsection 3.2.3.1, with exception that we have used a transgenic zebrafish – Tg(-2.5th:EGFP). Effects of 6-OHDA on the motor activity of adult zebrafish were examined by 9-day behavioural tests. Adult fish, Tg(-2.5th:EGFP) were treated intramuscularly with $5\mu L$ of 6-OHDA (33 mg/kg) – Figure 5.1, and the individual fish swimming responses ($n = 6$) recorded at various time-points after injections using MOBS. The dose, was selected on the bases of literature data and from our pilot experiments. We also tested a control group ($n = 6$) of fish treated with saline solution (the vehicle solution). The intramuscularly injections were administrated into anaesthetized fish in a total volume of $4.0\mu L$ per $0.3g$ fish using a gastight syringe and a 30-gauge needle (Hamilton, USA). Fish used for the studies had an average body weight of $0.5 \pm 0.05g$. Before the behavioural tests, small groups of female fish ($12 - 14$ animals) were acclimatized to the testing conditions (temperature $22°C \pm 1°C$, 10 h: 12 h light-dark cycle) in 17 liter glass aquaria under static conditions and for a minimum of three days. Fish were fed (Sera Vera, Germany) 1% of body weight per day throughout the tests. On the day of experiments (day 0), either the treated or control groups of fish were individually placed in test chambers for 30 minutes and then individual baseline responses (pre-treatment) were monitored for one hour between 10 a.m to 15 p.m. Fish were then individually anaesthetized with tricaine ($50mg/l$) and were injected with the neurotoxin or the vehicle solution. After injections, fish were kept in extensively aerated water tank until they recovered from the anaesthesia. Behaviour responses were then monitored at day 1. At the end of day 3 (after a new monitoring), fish received a second re-injection with similar volume and dose of neurotoxin. Individual fish swimming responses were recorded again at day 6 and 9. In each session of analysis the individual responses were evaluated every 15 minutes intervals for a total period of 60 minutes.

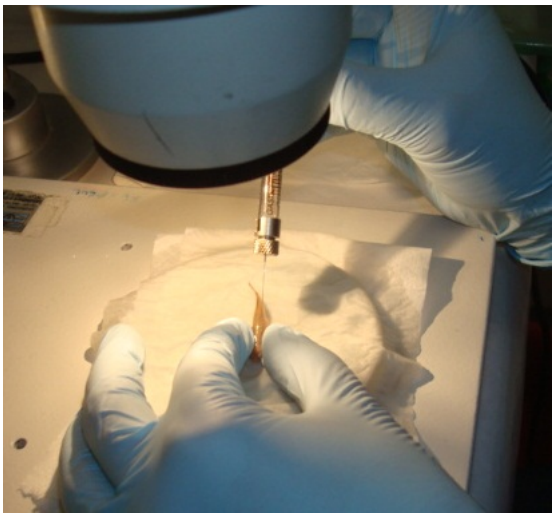


Figure 5.1: Intramuscular injection with 6-OHDA.

The analysis with MOBS system contained in total four independent experiments, each including two controls and two treated fish with 6-OHDA and the data was then polled for statistical analyses using the new developed algorithm. After behavioural recording, fish were sacrificed with tricaine. The behavioural experiments were always performed by the same experimenter [18].

5.1.2 Statistical Analysis

The effect of 6-OHDA on the changes of zebrafish swimming activity across the recording sessions was analysed between both groups using the one-tailed Mann-Whitney U-Test. The level of statistical significance was set to $p < 0.05$ and $p < 0.02$. All analysis were performed in IBM ®SPSS ®Statistics 20.0.

5.1.3 Results and Discussion

After recording swimming activity, the new algorithm was applied in each data using the procedure already described in Figure 4.4. The outcome provided the number of tail-flips per minute, and considering each group (control and treated group) the average was measured for all fish according to each day. The results are shown in Figure 5.2.

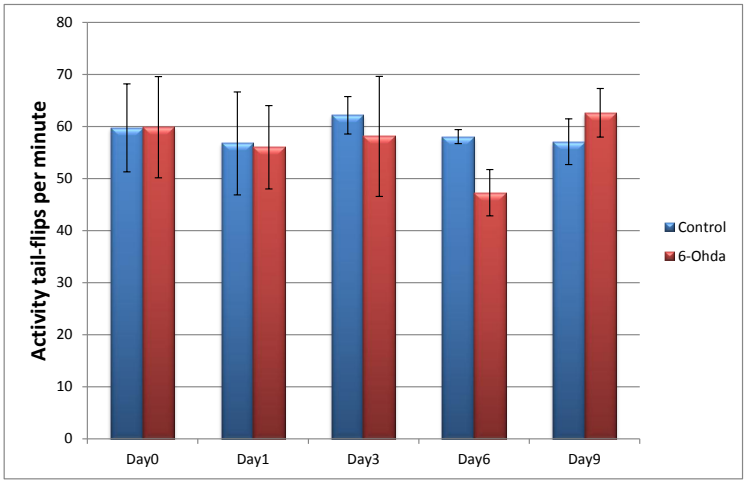


Figure 5.2: Behaviour results over the effect of 6-OHDA. The black bars represent mean±standard deviation.

To follow the biological responses according to [12, 17, 18], at day 0 both groups should be similar in their level of activity which can be verified in Figure 5.2: activity in tail-flips per minute of 59.72 ± 8.45 for control and 59.84 ± 9.72 for treated ($p > 0.05$). At the end of day 0, the injection of 6-OHDA was applied, and as a consequence, at day 1 this group should show a decrease in their activity in relation to the control group: activity of 56.74 ± 9.88 for control and 55.99 ± 8.01 for treated. This is not shown for a significant level of 5% ($p > 0.05$). At day 3 it is expected an increase in the level of activity for both groups, which justify their high capacity for regeneration: activity of 62.15 ± 3.60

for control and 58.08 ± 11.53 for treated .This situation is also verified ($p > 0.05$). At the end of day 3 a new re-injection of 6-OHDA was administered and as shown at day 6, both groups have significant differences between them, meaning that the re-injection caused a higher decrease in the zebrafish level of activity ($p < 0.05$): activity of 58.05 ± 1.34 for control and 47.28 ± 4.45 for treated. At day 9 it is expected again a increase in the level of activity for both groups, which is also verified in Figure 5.2 ($p > 0.05$): activity of 56.67 ± 4.44 for control and 62.62 ± 4.66 for treated. The activity of control fish was maintained constant throughout the experiment in comparison to day 0 ($p > 0.05$). To refer that for a significant level of 2% day 1 ($p > 0.02$) and 6 ($p > 0.02$) do not present differences between both groups.

To reinforce this study, and to understand where to improve in the algorithm, a confusion matrix was built - table 5.1. This may tell us how the classifier is behaving. The accuracy obtained was 80.80%, the sensitivity of 95.56% and specificity of 20.45% (target class control group).

Table 5.1: Confusion Matrix applied in the behavioural analysis.

		Predictions		
		Healthy	Ill	Sum
Correct Class	Healthy	172	8	180
	Ill	35	9	44
	Sum	207	17	224

There was a total of 224 analysis of 15 minutes each, where it should be expected to have 180 analysis classified as "healthy" and 44 classified as "ill". The classifier predicted 207 cases as "healthy", and the other 17 as "ill". This means that the classifier is showing difficulties classifying "ill" fish, which as presented in table 5.1, 35 cases were classified as "healthy" when in fact they were "ill" (also there were 8 cases that were misclassified as "ill"). This justifies the low value of specificity. To confirm that improvement needs to be done in the classifier, the algorithm was applied again in all data, but providing classification. The result is presented in Figure 5.3.

It is visible that the activity for the control groups are maintained over the days (comparison to day 0 $p > 0.02$ and $p > 0.05$). Activity for control groups of 59.72 ± 8.45 at day 0; 56.90 ± 13.38 at day 1; 53.59 ± 6.65 at day 3; 58.44 ± 5.18 at day 6 and 59.93 ± 7.33 at day 9. Also the treated groups are maintained at days 0 ($p > 0.02$ and $p > 0.05$), 3 ($p > 0.02$ and $p > 0.05$) and 9 ($p > 0.02$ and $p > 0.05$). Activity for treated groups of 62.87 ± 8.13 at day 0; 59.88 ± 13.88 at day 3 and 66.70 ± 9.16 at day 9. But most importantly, day 1 ($p < 0.02$) and 6 ($p < 0.02$) present differences between groups with a lower level of significance ($\alpha = 2\%$), which shows that the behaviour characterization is well suited for this behaviour analysis: activity for treated groups of 25.8 ± 6.39 at day 1 and 21.36 ± 11.29

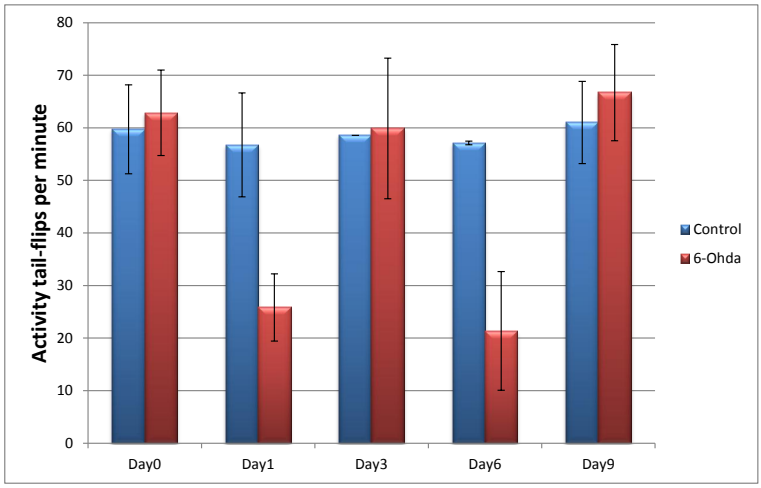


Figure 5.3: Behaviour results over the effect of 6-OHDA without using the SVM classifier. The black bars represent mean±standard deviation.

at day 6. Nevertheless if there are improvements to be done, these should be done in the SVM classifier.

There is still the possibility that the fish did not react to the drugs effect as expected (see subsection 4.1.2). However from all classified points, there was not one case (both with or without classifier) whose characterization had shown a zero crossing rate before the intersection of the curves from Figure 4.1. Therefore is assumed that the problem can be from the classifier, or from the data used.

Several tests were performed aiming to improve the classifier. For example provide more "ill" data than "healthy" to see if the classifier is more likely to predict this class. Also increase the parameter Cost from the SVM properties with the intuit to penalise more "healthy" fish. A higher Cost value provides a solution with less points misclassified, however is less tolerable to outliers [50]. These two hypothesis were analysed but since the confusion matrix shown in table 5.1 did not improve with these changes (accuracy 44.20%, specificity 77.27% and sensitivity 36.11%), the previous classifier properties were maintained and there still remains the need for improvement in this matter.

There is also the need to assume that the problem can be from the data that was used for the construction of the classifier, or even that there is not sufficient data to make a better distinction for new cases. The last assumption is that the signal for "ill" and "healthy" fish cannot be distinguished, even though the SVM allowed a perfect separation for this data.

5.2 Other Applications

This algorithm was built with the intuit to study the zebrafish behaviour when submitted to drugs that decrease their level of activity. Nevertheless, this algorithm can be used in other applications.

5.2.1 Test and Assess new Drugs

Besides using 6-OHDA to simulate PD there are other drugs that can be tested in zebrafish to study other diseases including acute and chronic pain.

Pain is a major symptom in many medical conditions, and often interferes significantly with a person's quality of life. Although a priority topic in medical research for many years, there are still few analgesic drugs approved for clinical use. One reason is the lack of appropriate animal models that faithfully represent relevant hallmarks associated with human pain. The work performed by Correia et. al (2011) [23], proposes zebrafish as a model to study nociception. Their results suggests that changes in zebrafish behavioural responses to acetic acid measured with the biosensor MOBS is a reasonable model to test analgesics. Thus the developed algorithm can also be a contribution to this work. More precisely, an algorithm that can distinguish different behavioural phenotypes of zebrafish to allow to test and assess new analgesics.

5.2.2 Water Quality/Pollution Detection

Nowadays coastal zones are confronted with intense human activities. Given the social-economic and ecological relevance of these areas, much effort has been directed towards new technologies that can rapidly detect the harmful presence of toxic chemicals in the water. A quick and effective monitoring still define a high priority in environmental research . Automated on-line biomonitor systems with living organisms reveal a promising solution. Ideally, these systems should detect environmental pollution situations as early stress responses of sensitive test organisms by automated recording [51]. Using organisms as biological sensors has the general advantage that changes in their behaviour (e.g., avoidance responses, swimming patterns and breathing) can be measured directly as responses to environmental changes. Indeed, behaviour has been used as an integral parameter of physiological activity and as a robust biological warning indicator of water quality supplies and effluents [52]. Although many aquatic organisms can be considered as relevant for behavioural studies, fish is the most used as a test specie [53]. The MFB for example, has been used to detect pollution based on behavioural stress responses [25, 26, 28, 54].

Given that, the developed algorithm in this research may also contribute favourably to this field allowing the detection of water pollution contaminants.

5.2.3 Regeneration

Regeneration is the process by which damaged or lost structures are perfectly or near-perfectly replaced. Mammals contain several organ systems capable of regeneration, such as blood and liver, but the majority of organs heal by scarring [55]. Today, investigation of regeneration in lower vertebrate model systems complements the modern field of stem cell research. That is, if we understand how regeneration occurs naturally in these organisms, we can learn how to optimize regenerative medicine in humans. Zebrafish is

known for its ability to regenerate multiple structures (fins, optic nerve, scales, heart, and spinal cord [56, 57, 58, 59]). For example, zebrafish caudal fin is an organ that is easily accessed for surgery and its injury does not compromise survival [60].

Hence, assuming that the surgery will cause variations in the behaviour without compromising its survival, our algorithm may be a valuable mean to characterize the behaviour and allow a different view over regeneration.

6

Conclusions

A new algorithm is proposed to classify and characterize behaviour in zebrafish specimens. The characterization provides the number of tail-flips per minute, and with the injection of the neurotoxin 6-OHDA to simulate PD it was noticed that the behaviour characterization to the less active fish operates differently from the "healthy" ones. Therefore a classifier was needed in this development.

The first intention would be to improve the current algorithm, however a detailed analysis using video frame by frame synchronised with the signal to detect the behaviour, proved that the algorithm was apart from reality with significant errors. The relative error obtained was 17.29% for "healthy" fish and 25.31% for "ill" fish, and even with the possibility to improve the algorithm, more specifically in the multiplicative factor, it was noticed that the best factor for both groups was far from being ideal (relative error of 53.20% for "healthy" and 44.53% for "ill" fish). Given that, a new algorithm was implemented.

The behaviour characterization required visual analysis. The functionality that allowed synchronism between video and signal was built in the Open Signals platform. The result from this analysis showed that the behaviour tail-flip could be characterized using the parameter zero crossing rate both for "healthy" fish with a relative error of 2.55% and "ill" fish with a relative error of 5.75% using different hypothesis. Given that, a classifier was needed to separate "healthy" and "ill" fish. This one was built using the software *Orange* that allowed the study of different methods, the SVM and Naïve Bayes. In the end it was chosen the classifier more accurate - the SVM with an accuracy of 100%. The final output of the algorithm presents the classification ("healthy" or "ill") with its respective probability and the behaviour characterization using the respective hypothesis to provide the number of tail-flips per minute – equations 4.1 and 4.2 for "healthy" and

"ill" groups respectively.

The final algorithm was integrated in the Open Signals platform to facilitate its use, and to allow any researcher to use it without requiring any knowledge in programming. The user can too benefit from the synchronism developed during this dissertation. This integration took also in consideration parallel programming to allow a faster result from the algorithm.

The final step of this thesis was to apply the algorithm in a new case study related with PD to confirm if the responses of the algorithm were in agreement with the biology and literature, and to understand the improvements that should be taken in the algorithm. The results showed that the fish activity were in agreement to the biology and literature for a significant level of 5% with exception at day 1. Yet the classifier needed to be improved to allow more significant differences between both groups ("healthy" and "ill"). More specifically, it was noticed that the classifier had difficulties in classifying "ill" fish, therefore it was provided more "ill" data than "healthy" to see if the classifier had a tendency to classify "ill" fish. Also, the Cost parameter from the SVM properties was increased to decrease misclassification. These changes did not improve the classifier output, which means that there still remains the need for improvement in this matter.

The fact that this algorithm uses classification can be an advantage as it may bring an efficient separation between a "healthy" fish from one that has been genetically modified to have PD. Also with the visual analysis it is known that the new algorithm is closer to reality which will allow the study and test of new drugs that uses zebrafish behaviour. This algorithm may be useful for further studies not only related with PD, but any other that uses zebrafish behaviour as an end point to study human diseases.

The MOBS device also proved to be an important system to characterize the behaviour, since it is non-invasive and provides fast and sensitive results that allowed the development of the new algorithm.

This research also led to a publication available in appendix A that presents the development of the new algorithm.

To conclude, in this dissertation, a new algorithm was developed to characterize motor behaviour of zebrafish. This algorithm is more realistic to simulate zebrafish behaviour, even though still requires a better distinction between "healthy" and "ill" groups. However, is a valuable contribution to the PD research area, in particular, to test and assess new drugs.

6.1 Future Work

In this dissertation there are still improvements to be done. In the following list those needs are presented.

- **Improve the classifier**: Besides having an accuracy of 100% the classifier proved to have difficulties at separating efficiently "healthy" and "ill" fish to new cases. It is proposed for future work an improvement in this classifier to allow the separation between groups for a significant level of $\alpha = 2\%$. Furthermore, analyse different methods that may possibly provide better results for classification, as for example Logistic Regression, K Nearest Neighbours, Majority etc. Also understand if there are other features that provide better results than the ones used in this research.

- **Visual analysis**: If possible increase the number of visual analysis to strengthen the zero crossing rate parameter as a valuable mean to characterize the behaviour of zebrafish.

- **Study new behaviour**: In this research it was only studied the abrupt tail-flip, however it would also be important to include other types of behaviour, for example swimming and ventilation. To analyse ventilation the suggestion would be to confine the fish in smaller chambers. Hence, if ventilation could be studied separately from locomotion it would be possible to confirm the signal physiology with the zero crossing rate parameter.

- **Fish position in the chamber**: According to Cunha et. al (2008) [2], the smaller the distance between the electrodes and the organism is, the better the corresponding electric field can be identified and quantified. Therefore it would also be relevant to evaluate the position of the fish in the chamber and understand if the new algorithm is influenced by this situation.

- **Apply algorithm in other works**: Use this algorithm in other areas, namely to test the influence of new drugs on the behaviour of zebrafish, understand if this algorithm is a valuable mean for water pollution detection using MOBS and assess regeneration. Also judge if this algorithm can be used in other species besides zebrafish. If this is proven then we can assume that this algorithm is an general one to be used in future works.

Bibliography

[1] Fish for a healthier future | mending broken hearts | guardian.co.uk. http://www.guardian.co.uk/mending-broken-hearts/zebra-fish-fight-heart-disease, March 2012.

[2] S. R. Cunha, R. Gonçalves, S. R. Silva, and A. D. Correia. An automated marine biomonitoring system for assessing water quality in real-time. *Ecotoxicology*, 17:558–564, 2008.

[3] Machine learning. https://class.coursera.org/ml/lecture/preview, March 2012.

[4] ROC curves. http://www.medcalc.org/manual/roc-curves.php, September 2012.

[5] T. Gasser. Mendelian forms of parkinson's disease. *Biochimica et Biophysica Acta (BBA)-Molecular Basis of Disease*, 1792(7):587–596, 2009.

[6] P. Arsenault. Parkinson's disease in focus. *Canadian Family Physician*, 56(2):85–85, 2010.

[7] LaTeX – a document preparation system. http://www.latex-project.org/, September 2012.

[8] T. Curk, J. Demšar, Q. Xu, G. Leban, U. Petrovič, Bratko. I., G. Shaulsky, and B. Zupan. Microarray data mining with visual programming. *Bioinformatics*, 21(3):396–398, 2005.

[9] Fish for science. http://www.fishforscience.com/, March 2012.

[10] T. Fonseca. Zebrafish: A new model of parkinson's disease. Master's thesis, Universidade de Lisboa, 2010.

[11] T. Becker, M. F. Wullimann, C. G. Becker, R. R. Bernhardt, and M. Schachner. Axonal regrowth after spinal cord transection in adult zebrafish. *The Journal of comparative neurology*, 377(4):577–595, 1998.

[12] K. D. Poss. Getting to the heart of regeneration in zebrafish. *Seminars in Cell & Developmental Biology*, 18(1):36–45, 2007.

[13] E. C. Hirsch. Biochemistry of parkinson's disease with special reference to the dopaminergic systems. *Molecular neurobiology*, 9(1):135–142, 1994.

[14] G. R. Breese, D. J. Knapp, H. E. Criswell, S. S. Moy, S. T. Papadeas, and B. L. Blake. The neonate-6-hydroxydopamine-lesioned rat: a model for clinical neuroscience and neurobiological principles. *Brain research reviews*, 48(1):57–73, 2005.

[15] A. V. Kalueff and J. M. Cachat, editors. *Zebrafish Models in Neurobehavioral Research: 52*. Humana Press, 1st edition. edition, 2010.

[16] P. McGrath. *Zebrafish: Methods for Assessing Drug Safety and Toxicity*. John Wiley & Sons, 2012.

[17] Zebrafish as a new animal model for movement disorders. *Journal of neurochemistry*, 106(5):1991–1997, 2008.

[18] A. D. Correia, R. S. Soares, S. Sousa, T. F. Outeiro, N. Afonso, R. Willemsen, and Herma van der Linde. Green fluorescent protein labeling of dopaminergic neurons in zebrafish for the study of the molecular basis of parkinson's disease (submitted). 2012.

[19] S. Rodriguez-Mozaz, M. J. Lopez de Alda, and D. Barcelo. Biosensors as useful tools for environmental analysis and monitoring. *Analytical and Bioanalytical Chemistry*, 386(4):1025–1041, 2006.

[20] S. E. Lepage and A. E. E. Bruce. Characterization and comparative expression of zebrafish calpain system genes during early development. *Developmental Dynamics*, 237(3):819–829, 2008.

[21] C. Y. Usenko, S. L. Harper, and R. L. Tanguay. Fullerene c_{60} exposure elicits an oxidative stress response in embryonic zebrafish. *Toxicology and applied pharmacology*, 229(1):44–55, 2008.

[22] S. Bretaud, S. Lee, and S. Guo. Sensitivity of zebrafish to environmental toxins implicated in parkinson's disease. *Neurotoxicology and teratology*, 26(6):857–864, 2004.

[23] A. D. Correia, S. R. Cunha, M. Scholze, and E. D. Stevens. A novel behavioral fish model of nociception for testing analgesics. *Pharmaceuticals*, 4(4):665–680, 2011.

[24] W.T. Cochran, J.W. Cooley, D.L. Favin, H.D. Helms, R.A. Kaenel, W.W. Lang, G.C. Maling Jr, D.E. Nelson, C.M. Rader, and P.D. Welch. What is the fast fourier transform? *Proceedings of the IEEE*, 55(10):1664–1674, 1967.

[25] S. Craig and P. Laming. Behaviour of the three-spined stickleback, gasterosteous aculeatus (Gasterosteidae, teleostei) in the multispecies freshwater biomonitor: a validation of automated recordings at three levels of ammonia pollution. *Water Research*, 38(8):2144–2154, 2004.

[26] M. Schriks, M.K. van Hoorn, E.J. Faassen, J.W. van Dam, and A.J. Murk. Real-time automated measurement of xenopus leavis tadpole behavior and behavioral responses following triphenyltin exposure using the multispecies freshwater biomonitor (MFB). *Aquatic toxicology*, 77(3):298–305, 2006.

[27] A. Gerhardt, A. Carlsson, C. Ressemann, and K.P. Stich. New online biomonitoring system for gammarus pulex (L.)(Crustacea): in situ test below a copper effluent in south sweden. *Environmental science & technology*, 32(1):150–156, 1998.

[28] A. Gerhardt, M.K. Ingram, I.J. Kang, and S. Ulitzur. In situ on-line toxicity biomonitoring in water: Recent developments. *Environmental Toxicology and Chemistry*, 25(9):2263–2271, 2006.

[29] H. Muir. *Science in Seconds*. Quercus Books, 2011.

[30] I. H. Witten, E. Frank, and M. A. Hall. *Data Mining: Practical Machine Learning Tools and Techniques*. Elsevier, 2011.

[31] E. Alpaydin. *Introduction to Machine Learning*. MIT Press, October 2004.

[32] A. P. Bradley. The use of the area under the ROC curve in the evaluation of machine learning algorithms. *Pattern recognition*, 30(7):1145–1159, 1997.

[33] J. Davis and M. Goadrich. The relationship between precision-recall and ROC curves. In *Proceedings of the 23rd international conference on Machine learning*, page 233–240, 2006.

[34] M. H. Zweig and G. Campbell. Receiver-operating characteristic (ROC) plots: a fundamental evaluation tool in clinical medicine. *Clinical chemistry*, 39(4):561–577, 1993.

[35] D. Meyer. Support vector machines. *Porting R to Darwin/X11 and Mac OS X*, 2011.

[36] I. Steinwart and A. Christmann. *Support Vector Machines*. Springer, 2008.

[37] C. C. Chang and C. J. Lin. LIBSVM: a library for support vector machines. *ACM Transactions on Intelligent Systems and Technology (TIST)*, 2(3):27, 2011.

[38] H. Zhang. The optimality of naive bayes. *A A*, 1(2):3, 2004.

[39] F. Gouyon, F. Pachet, and O. Delerue. On the use of zero-crossing rate for an application of classification of percussive sounds. In *Proceedings of the COST G-6 conference on Digital Audio Effects (DAFX-00)*, 2000.

[40] R.S.H. Ramos, F. Coito, and M. Ortigureira. *Analise de Sinais em Engenharia Biomédica*. FCT-UNL, 2009.

[41] L. M. Leemis and S. K. Park. *Discrete-event simulation: A first course*. Pearson Prentice Hall, 2006.

[42] P. Stoica and R. L. Moses. *Introduction to spectral analysis*, volume 51. Prentice Hall Upper Saddle River, NJ, 1997.

[43] J. O. Smith, Center for Computer Research in Music, and Calif Acoustics. Stanford. *Spectral audio signal processing*. Stanford University, CCRMA, 2008.

[44] M. T. O'TOOLE. Miller-keane encyclopedia & dictionary of medicine, nursing & allied health-second revised reprint. *Recherche*, 67:02, 2006.

[45] G. D. Ruxton. The unequal variance t-test is an underused alternative to student's t-test and the Mann-Whitney u test. *Behavioral Ecology*, 17(4):688–690, 2006.

[46] R. Shier. Statistics: 2.3 the mann-whitney u test, 2004.

[47] G. S. Almasi and A. Gottlieb. *Highly parallel computing*. 1988.

[48] Parkinson's disease: Treatments and drugs - MayoClinic.com. http://www.mayoclinic.com/health/parkinsons-disease/DS00295/DSECTION=treatments-and-drugs, August 2012.

[49] O. Rascol, C. Goetz, W. Koller, W. Poewe, and C. Sampaio. Treatment interventions for parkinson's disease: an evidence based assessment. *The Lancet*, 359(9317):1589–1598, 2002.

[50] A. Ben-Hur and J. Weston. A user's guide to support vector machines. *Methods in Molecular Biology*, 609:223–239, 2010.

[51] S. Kröger and R. J. Law. Biosensors for marine applications: We all need the sea, but does the sea need biosensors? *Biosensors and Bioelectronics*, 20(10):1903–1913, 2005.

[52] W. H. Van der Schalie, T. R. Shedd, P. L. Knechtges, and M. W. Widder. Using higher organisms in biological early warning systems for real-time toxicity detection. *Biosensors and Bioelectronics*, 16(7):457–465, 2001.

[53] W. H. Van der Schalie, K. L. Dickson, G. F. Westlake, and J. Cairns. Fish bioassay monitoring of waste effluents. *Environmental management*, 3(3):217–235, 1979.

[54] A. Gerhardt, L. Janssens de Bisthoven, Z. Mo, C. Wang, M. Yang, and Z. Wang. Short-term responses of oryzias latipes (Pisces: adrianichthyidae) and macrobrachium nipponense (Crustacea: palaemonidae) to municipal and pharmaceutical waste water in beijing, china: survival, behaviour, biochemical biomarkers. *Chemosphere*, 47(1):35–47, 2002.

[55] C. E. Dinsmore. *A history of regeneration research: milestones in the evolution of a science*. Cambridge Univ Pr, 1991.

[56] T. H. Morgan. Regeneration in teleosts. *Development Genes and Evolution*, 10(1):120–134, 1900.

[57] J. Bereiter-Hahn and L. Zylberberg. Regeneration of teleost fish scale. *Comparative biochemistry and physiology. A. Comparative physiology*, 105(4):625–641, 1993.

[58] T. Becker, M. F. Wullimann, C. G. Becker, R. R. Bernhardt, and M. Schachner. Axonal regrowth after spinal cord transection in adult zebrafish. *The Journal of comparative neurology*, 377(4):577–595, 1997.

[59] R. R. Bernhardt, E. Tongiorgi, P. Anzini, and M. Schachner. Increased expression of specific recognition molecules by retinal ganglion cells and by optic pathway glia accompanies the successful regeneration of retinal axons in adult zebrafish. *The Journal of comparative neurology*, 376(2):253–264, 1996.

[60] K. D. Poss, M. T. Keating, and A. Nechiporuk. Tales of regeneration in zebrafish. *Developmental Dynamics*, 226(2):202–210, 2003.

Publications

In this appendix is presented the publication, *Algorithm for Testing Behavioural Phenotypes in a Zebrafish model of Parkinson's Disease* which demonstrates the algorithm that was developed during this dissertation. This article was accepted for short paper presentation to BIOSIGNALS 2013, which is a conference – 6^th *International Joint Conference on Biomedical Engineering Systems and Technologies* (BIOSTEC 2013), held in Barcelona in February 2013.

Algorithm for Testing Behavioural Phenotypes in a Zebrafish Model of Parkinson's Disease

Angela Pimentel[1], Hugo Gamboa[1,2], Sérgio Reis Cunha[3] and Ana Dulce Correia[4]
[1], *CEFITEC, Physics Department, FCT-UNL, Lisbon, Portugal*
[2] *PLUX - Wireless Biosignals, Lisbon, Portugal*
[3], *Faculty of Engineering, Porto University, Porto, Portugal*
[4]*Instituto de Medicina Molecular, Faculty of Medicine, University of Lisbon, Lisbon, Portugal*
angela_pimentel9@hotmail.com, hgamboa@plux.info, acorreia@fm.ul.pt, sergio@fe.up.pt

Keywords: Parkinson's Disease (PD), Zebrafish, Behaviour, Biosensor MOBS, Machine Learning.

Abstract: Parkinson's disease (PD) is one of the neurodegenerative diseases with an increased prevalence widely studied by the scientific community. Understanding the behaviour related to the disease is an added value for diagnosis and treatment. Thus the use of an animal model for PD that develops similar symptoms to the human being allows to the clinic a larger vision over the health of a patient. Zebrafish can be used to study some human diseases including PD. This work describes the development of an algorithm for the characterization of behaviour in this specie. The biosensor called Marine On-line Biomonitor System (MOBS) is connected electrically to chambers where the specimen of zebrafish moves freely providing a signal that is related with the fish activity. Using the developed algorithm based on signal processing, statistic analysis and machine learning techniques we present classification of a fish as normal or ill and characterize its behaviour.

1 INTRODUCTION

Biosensors are an essential control and safety tool for our environmental and health quality and commonly used in medicine. Many of today's biosensor applications use living organisms which respond to toxic substances or other stressors at a much lower level than us to warn us of their presence. Under this scope, the MOBS was developed, an automated system for recording behavioural responses of marine and fresh water species. This device has been applied successfully in the environmental field, and the next challenging step is to bring this technology into other research areas. In particular, by sensing behavioural changes in organisms as an indication of stress or disease. A suitable model candidate is the zebrafish, a freshwater specie which has been used in medical research during the past years, e.g in development studies (Lepage and Bruce, 2008), drug toxicity assessments (Usenko et al., 2008) and neurodegenerative diseases (Bretaud et al., 2004).

1.1 PD and Zebrafish

The PD is characterized by tremor, muscle rigidity, a slowing of physical movement, and can also cause cognitive and mood disturbances. It results of the loss of nerve cells in part of the brain known as the substancia nigra. These cells are called dopaminergic (DA) neurons as they produce the neurotransmitter, dopamine, which is used to send messages to the parts of the brain that co-ordinate movement (Fis, 2012). Most insights into human disease are a result of experiments that would be unethical or unfeasible to perform on humans. Instead biomedical research uses models to look at the functions of the genes involved in maintaining healthy organisms in order to obtain vital clues about the causes and progression of human diseases. Zebrafish are an ideal model organism to bridge the gap between too simple (yeast) and too complex (mice or rats). They are vertebrates and have similar body plans (and similar tissues and organs) to humans, and they're much easier and with reduced cost to breed than mice and rats. Zebrafish mutations phenocopy many human disorders and the genome sequence of zebrafish is near completion. The DA nervous system in zebrafish is well characterized in both embryos and adult zebrafish. Some toxins known to induce DA cell loss in other animal models have now also been tested in adult zebrafish, as for example, the *6-hydroxydopamine* (6-OHDA) which is a neurotoxin

that induces death of the DA cells. After injecting the neurotoxin via intramuscular, locomotor activity and dopamine levels of the brain decreases (Kalueff and Cachat, 2010; McGrath, 2012; Breese et al., 2005; Flinn et al., 2008). Thus the evaluation of swimming behaviour can be related with the loss of DA cells, and consequently with the PD. In the work performed by (Correia et al., 2012) a new transgenic line of zebrafish was developed to study the DA neurons, which were validated with the use of the neurotoxin 6-OHDA and with the behaviour analysis using the biosensor MOBS. They verified behavioural changes that were related with the death of the DA neurons. The algorithm to be developed can be a contribution for this work: an algorithm that is sensible in the behaviour characterizations to allow the responses to be comparable with the loss of the DA neurons.

1.2 Current Approach

The current algorithm used to characterize the behaviour of zebrafish consists in the evaluation of a specific locomotion behaviour, with a series of bursts in the domain of MOBS corresponding to the tail-flip activity of zebrafish. Thus the outcome reflects the number of tail-flips per minute per individual fish (Correia et al., 2011). The behaviour detection is based on the derivative peaks resulted from the strong bursts in the signal. However, these peaks require a threshold for the behaviour detection, and this is accomplished using the standard deviation multiplied by a factor so that these two parameter, standard deviation and derivative, may be comparable. It's essential to confirm if the current algorithm is in fact detecting the right behaviour, the tail-flips. The first intention of this research would be to understand and improve the current algorithm, however it will be proved the need to create a new one using supervised learning.

1.3 Supervised Learning

By Arthur Samuel (1959), machine learning is the field of study that gives computers the ability to learn without being explicitly programmed. There are different types of machine learning algorithms, the main two types are: unsupervised and supervised learning.

With supervised learning, the scheme operates under supervision by being provided with the actual outcome for each of the training examples. In this type of machine learning is included regression problems that predicts continuous valued outputs and classification problems which intends to predict discrete valued outputs (mac, 2012). For classification problems, a known method is the Support Vector Machine (SVM) which looks for the optimal hiper-plane between two classes by maximizing the margin. A non-linear separator is possible by projection the data points to higher-dimension space to become linearly separable (projection with kernel techniques) (mac, 2012). Also the method Naïve Bayes which applies Bayes theorem to estimate the probability with the "naïve" assumption of independence between each feature. For validation, a possible statistic test is *leave one out*, which given a dataset of m instances, only one instance is left out as the validation set (instance) and training uses the $m-1$ instances (Witten et al., 2011).

2 METHODS

2.1 MOBS

The main device is controlled via an USB port by external processing software which produces signals in the digital domain (at 48000 samples/s or 48 kHz). These are converted by the main device into analogical electrical signals, power amplified and transmitted to the independent testing units at which they are conducted into the water by a pair of non-invasive stainless steel electrodes. In response to the behavioural signatures of the organisms as a change in impedance of the water, the amplitudes of the electrical signals are modulated and then received by a second pair of electrodes. In the main device they are amplified and converted back to the digital domain at 48000 samples/s, before filtered, demodulated and downsampled at 100 Hz by the external computer software. Upon processing, the system provides a signal in the frequency band of 0.2 Hz to 40 Hz that is correlated with the fish activity (Cunha et al., 2008). With MOBS, locomotion can be presented with a series of bursts in the time domain, and can cover a broad frequency spectrum, at which ventilation is occasionally present. Typically ventilation generates waves of triangular shape with a higher frequency and smaller amplitude than the most of the energy located for locomotion. However ventilation will not be studied with zebrafish given its high level of activity.

2.2 Experimental Design

2.2.1 *Test Animals and 6-OHDA*

The zebrafish (D. rerio Hamilton 1822) strain used for this work was the AB line (Zebrafish Facility, IMM, Portugal). Animals were maintained under standard conditions and experiments were approved by the Institutional Animal Care and Use Committee. A mas-

ter stock solution of 6-hydroxydopamine hydrochloride (6-OHDA, Sigma-Aldrich, USA) was prepared in 0.2% ascorbic acid solution (analytical grade, Sigma) and stored at -20°C.

2.2.2 *Behaviour Assay*

Before the experiments, small groups of female fish (24 animals, body weight 0.5 ± 1 g) were acclimatized to the experimental testing conditions (temperature 22 °C ± 1 °C, 10 h:12 h light-dark cycle) in 17 litre glass aquaria under static conditions and for a minimum of one week. Food was not provided 24 h before or during the experiments. The behaviour analysis was divided in two groups: non-treated (12 fish) and for that considered as normal fish in which no injection was administered, and treated (12 fish) also considered as ill or less active where $5\mu L$ of 6-OHDA (33 mg/kg) was injected via intramuscular. During the injection they were in a medium-to deep-plane level of anaesthesia (tricaine 50 mg/L) and had lost their reflex responses and muscular control. Afterwards they returned to their original test chambers and allowed 30 min to recover from the anaesthesia. On the day of experiments, either the treated or non-treated groups of fish were placed individually in the test chambers (22 °C ± 1 °C) and acclimated for 30 min. Then individual baseline responses were monitored using MOBS and recorded using video (property of 25 frames per second) for five minutes between 10 and 12 a.m. After behavioural recording, treated fish were sacrificed with tricaine.

2.3 Synchronism

The signal in the time domain is delayed in relation to the instant of acquisition start. This delay is caused by the main device, which makes it difficult to compare a video where the fish movements are present, with its respective signal from MOBS. The Open Signals is a platform designed and programmed by *PLUX - Wireless Biosignals*. Using this platform, synchronism is possible with a visible stimulus in the signal and video. This stimulus must be sufficient to not be confused with the fish activity in the signal. A touch in the chamber is a possible stimulus and to not corrupt the signal from the fish activity for further analysis the stimulus should be produced at the end of the recording.

2.3.1 *Visual Analysis*

To verify what the algorithm is detecting a detailed analysis using Open Signals was necessary after synchronism. This analysis using the video frame by frame consisted in the detection of the behaviour tail-flip. The tail-flip is characterized as an abrupt and fast change of direction implying a strong burst in the tail. The visual analysis will consist in counting the number of tail-flips detected and divide it by the total time in minutes. Since the visual analysis is a long process, 24 study cases were made, 12 of them were non-treated and the rest were submitted to the drug 6-OHDA. Each visual analysis consisted in 3 minutes of the video. Since the visual analysis depends of the user that is interpreting the data, it's important to test other user and compare the results. A visual test using a different user was made. The test consisted in a precise analysis frame by frame using a signal with 30 seconds, and for this time both users detected 46 abrupt tail-flips. After the User 1 detect the abrupt tail-flip it was considered an interval of 0.25 seconds in which the User 2 had also to detect the same abrupt tail-flip to be a valid success.

2.4 Current Algorithm Evaluation

In this subsection is intended to compare the visual analysis with the algorithm result using linear regression for each group (treated and non-treated) and estimate the relative error with the *leave one out* method. This was chosen because the number of points analysed is small. Also in consideration is the correlation coefficient which is a numerical value that indicates the degree and direction of relationship between two variables (O'TOOLE, 2006). The relative error obtained will show the need to improve the algorithm.

The multiplicative factor in the current algorithm is used so that the derivative can be comparable to the standard deviation thus allowing to detect the behaviour abrupt tail-flips. Given that, to improve the algorithm the multiplicative factor should be analysed. The value used so far has been 0.1. To understand which is the best factor value, it was decided to vary the factor according to the outcome of the algorithm, and with the visual analysis choose the factor that was closer to reality. A unique study case isn't sufficient to choose the ideal factor, thus with all data analysed for each group it's estimated the average relative error from the current algorithm result with the visual analysis. In the end we chose the factor that minimizes the relative error.

2.5 New Algorithm

2.5.1 *Behaviour Characterization*

To characterize the behaviour in number of tail-flips per minutes it was necessary to use the parameter zero

crossing rate. The zero crossing rate it is defined as the number of time-domain zero crossings within a defined region of signal, divided by the number of samples of that region (Gouyon et al., 2000). Each data was divided by its standard deviation, so that all data is at the same scale to be comparable and because the signal is centred at zero, it wasn't necessary to subtract the average. Also the signal was smoothed using a Hanning window of 0.05 seconds. To validate this parameter it was used the statistic analysis *leave one out*. This was chosen because the number of points analysed is small. This study also considered the correlation coefficient.

2.5.2 *Classification*

The Orange is a software suitable for machine learning. It is a free software and open source. It allows to use data mining through visual programming and Python scripting (Curk et al., 2005). The classifier was studied with the methods SVM and Naïve Bayes. The validation used the statistic analysis *leave one out* to provide the accuracy for each method used (SVM or Naive Bayes) which is the proportion of correctly classified examples (Curk et al., 2005). Thus varying the number of parameters obtained from the data we choose the ones that give higher accuracy for the respective method. The parameters extracted from each data were the zero crossing rate, the standard deviation, the maximum power using the periodogram, the maximum number of occurrences using the histogram and the current algorithm output. Also the optimal values for the SVM, namely the Cost parameter and the gamma value for the kernel function were chosen by the Orange software which uses the LIBVSM library. Since the classifier doesn't require the visual analysis as output, which is a long process, instead of using the data obtained so far (24 study cases), it was used data from a previous work to provide more points to the classifier (108 study cases with equal number for each class). This work developed at the Instituto de Medicina Molecular provides data with non-treated and treated fish (submitted to the drug 6-OHDA).

3 RESULTS AND DISCUSSION

3.1 Synchronism

3.1.1 *Visual Analysis*

In 46 detections between both users, 44 were accepted, leading to an error of 4.35%. The agreement between the users characterizing the behaviour, leads that the visual result can be a valid information to be compared with the current algorithm or with future works.

3.2 Current Algorithm Evaluation

We can now compare the algorithm output with the visual analysis. The results are shown in figure 1. It

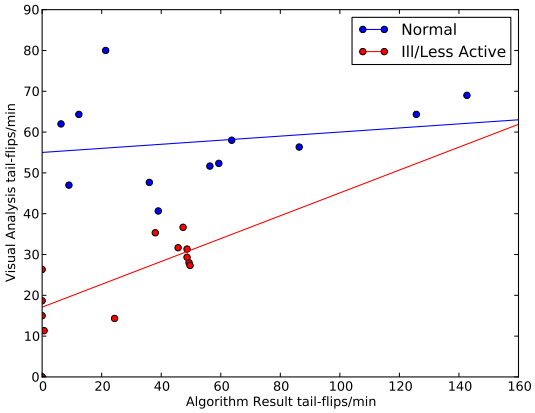


Figure 1: Comparison between the visual analysis and the algorithm result.

is visible that there is no direct relation between the visual analysis and the algorithm output as it would be expected both for treated and non-treated fish. After applying linear regression in each group it was estimated the relative error with the method *leave one out* which resulted in an error of 17.29% for the non-treated and 25.31% for treated. Also the correlation coefficient obtained was 0.20 and 0.76 for the non-treated and treated respectively which can be considered as a poor relation between the visual analysis with the algorithm output. These errors imply an improvement in the algorithm, more specifically in the multiplicative factor. To choose the best factor it was decided to study the error associated with the visual analyse. Figure 2 indicates the minimum error accepted as well as the error used with the actual factor for the treated and non-treated fish. The error using the actual factor is 55.26% and 68.79% for non-treated and treated respectively, and even improving the factor, the minimal error accepted would be 53.20% for non-treated which leads to a best factor of 0.11 and 44.53% for treated with a best factor of 0.13. To be able to choose the best factor these errors obtained should be as close to zero as possible which indicates that even with these improvements the best multiplicative factor cannot be certain. Therefore, and

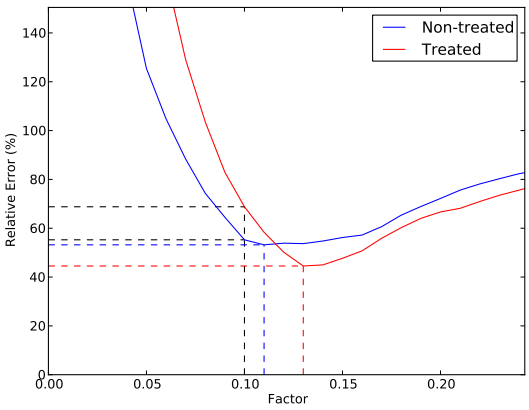


Figure 2: Relative error in percentage. Black dotted lines: Actual multiplicative factor (0.1); Red dotted lines: Best multiplicative factor for treated; Blue dotted lines: Best multiplicative factor for non-treated.

considering that the visual analysis is a valid measure, it is suggested the development of a new algorithm.

3.3 New Algorithm

With the visual analysis it will be possible to study new parameters using supervised learning, more precisely, regression models.

3.3.1 *Behaviour Characterization*

Figure 3 shows visually that there is a linear tendency between the zero crossing rate results with the visual analysis both for treated and non-treated fish. Considering first the normal fish for validation, it was

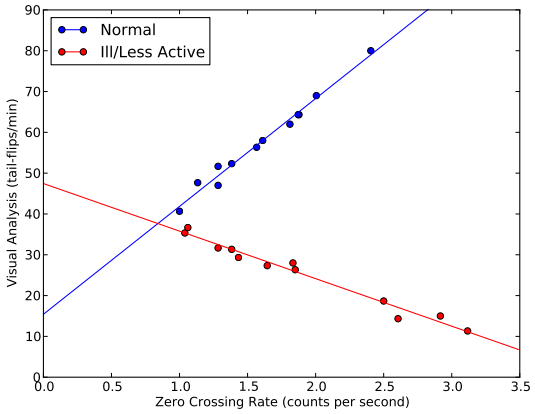


Figure 3: Comparison between the zero crossing rate with the visual analysis for normal and ill fish with a window of 180 seconds.

used the statistic analysis *leave one out*. The result leaded to a error of 2.55%. The relative error of 2.55% compared with the 17.29% from the previous algorithm can be considered as an excellent improvement. The user test from the previous section showed an error of 4.35%. Given that, the reason why this parameter shows a smaller error (2.55%) it's because it suits the user that performed this analysis. If User 2 had also performed this analysis, it should be expected a bigger error. The correlation coefficient obtained in this case was 0.99, indicating that there is a very good positive relation between the zero crossing rate and the visual analysis. Finally using all points for a window of 180 seconds, linear regression can be applied to define our hypothesis:

$$h_\theta(x) = 15.42 + 26.43x \quad (1)$$

To characterize the behaviour for ill fish, Figure 3 shows that this group presents an inverse linear tendency between the zero crossing rate and visual analysis, which means that the higher the number of counts per second the less active the fish is. Again it was used the *leave one out* method to validate this parameter. The relative error obtained was 5.75% which can be a good estimative even thought it's higher than the error obtained to characterize normal fish (2.55%). This error in comparative to the 25.31% from the previous algorithm can also be considered as an excellent improvement. The correlation coefficient was -0.99, meaning there is a very good inverse relation between the visual analysis and zero crossing rate.

Using all points for a window of 180 seconds linear regression can be applied to define our hypothesis:

$$h_\theta(x) = 47.45 - 11.65x \quad (2)$$

The value of 47.45 tail-flips per minute limits the fish activity, which means that ill fish won't show a higher value of activity than 47.45 tail-flips per minute. Also for a fish that doesn't present any activity (0 tail-flips per minute) it should be expected a value of 4.07 counts per second. Since both groups use different equations to characterize the behaviour, to know which equation to use for the development of this algorithm a classifier is needed to distinguish between normal or ill fish.

3.3.2 *Classification*

Now our output is defined by two classes: normal and ill fish. The parameters used that leaded to a higher accuracy for the SVM were the zero crossing rate, the standard deviation, the maximum power using the periodogram, the maximum number of occurrences using the histogram, and the previous algorithm output.

The learning options used were for the kernel function the Sigmoid function ($tanh(8 * x.y)$), a Cost of 2.0 (Model Complexity - penalty parameter) and a numeric precision of 0.001. The accuracy obtained using *leave one out* for the SVM method was 100%, meaning that all cases analysed were correctly classified. On the other hand, the Naive Bayes method based on the relative frequency presented a maximum accuracy of 67.59% using the parameters standard deviation, the maximum power using the periodogram and the previous algorithm output.

As we want to choose the classifier that predicts the classes with a higher accuracy value we choose the method SVM to build our final classifier. Because the Orange program is open source, with the access to the functions that build the classifier SVM we can use them to construct the final algorithm in python.

3.3.3 *Final Algorithm*

Now it's possible to built the final algorithm. First we prepare the data with the removal of the initial peak from the main device, the application of a filter to exclude possible noise, the normalization of the data and the smooth of the signal using a Hanning window of 0.05 seconds. Then we use the classifier to predict if the fish is normal or ill. Consequently, according to the classification it's possible to characterize the behaviour in number of tail-flips per minute using the corresponding hypothesis that consists in the use of the parameter zero crossing rate. The final result will present the classification, the probability for that classification, and the number of tail-flips per minute.

4 CONCLUSIONS

A new algorithm was developed to classify and characterize the behaviour of zebrafish. To facilitate its use, the algorithm should be integrated in the platform Open Signals. The fact that this algorithm uses classification can be an advantage as it may bring an efficient separation between a healthy fish from one that has been genetically modified to have PD. Also, the algorithm should be applied in a case study as executed by (Correia et al., 2012), to verify that the responses are in agreement with the fish behaviour and literature. This algorithm may be useful for further studies not only related with PD, but any other that uses zebrafish behaviour as an end point to study human diseases.

REFERENCES

(2012). Fish for science. http://www.fishforscience.com/.

(2012). Machine learning. https://class.coursera.org/ml/lecture/preview.

Breese, G. R., Knapp, D. J., Criswell, H. E., Moy, S. S., Papadeas, S. T., and Blake, B. L. (2005). The neonate-6-hydroxydopamine-lesioned rat: a model for clinical neuroscience and neurobiological principles. *Brain research reviews*, 48(1):57–73.

Bretaud, S., Lee, S., and Guo, S. (2004). Sensitivity of zebrafish to environmental toxins implicated in parkinson's disease. *Neurotoxicology and teratology*, 26(6):857–864.

Correia, A. D., Cunha, S. R., Scholze, M., and Stevens, E. D. (2011). A novel behavioral fish model of nociception for testing analgesics. *Pharmaceuticals*, 4(4):665–680.

Correia, A. D., Soares, R. S., Sousa, S., Outeiro, T. F., Afonso, N., Willemsen, R., and van der Linde, H. (2012). Green fluorescent protein labeling of dopaminergic neurons in zebrafish for the study of the molecular basis of parkinson's disease (submitted).

Cunha, S. R., Gonçalves, R., Silva, S. R., and Correia, A. D. (2008). An automated marine biomonitoring system for assessing water quality in real-time. *Ecotoxicology*, 17(6):558–564.

Curk, T., Demsar, J., Xu, Q., Leban, G., Petrovic, U., Bratko, I., Shaulsky, G., and Zupan, B. (2005). Microarray data mining with visual programming. *Bioinformatics*, 21(3):396–398.

Flinn, L., Bretaud, S., Lo, C., Ingham, P. W., and Bandmann, O. (2008). Zebrafish as a new animal model for movement disorders. *Journal of Neurochemistry*, 106(5):1991–1997. PMID: 18466340.

Gouyon, F., Pachet, F., and Delerue, O. (2000). On the use of zero-crossing rate for an application of classification of percussive sounds. In *Proceedings of the COST G-6 conference on Digital Audio Effects (DAFX-00), Verona, Italy*.

Kalueff, A. V. and Cachat, J. M., editors (2010). *Zebrafish Models in Neurobehavioral Research: 52*. Humana Press, 1st edition. edition.

Lepage, S. E. and Bruce, A. E. E. (2008). Characterization and comparative expression of zebrafish calpain system genes during early development. *Developmental Dynamics*, 237(3):819–829.

McGrath, P. (2012). *Zebrafish: Methods for Assessing Drug Safety and Toxicity*. John Wiley & Sons.

O'TOOLE, M. T. (2006). Miller-keane encyclopedia & dictionary of medicine, nursing & allied health-second revised reprint. *Recherche*, 67:02.

Usenko, C. Y., Harper, S. L., and Tanguay, R. L. (2008). Fullerene c_{60} exposure elicits an oxidative stress response in embryonic zebrafish. *Toxicology and applied pharmacology*, 229(1):44–55.

Witten, I. H., Frank, E., and Hall, M. A. (2011). *Data Mining: Practical Machine Learning Tools and Techniques*. Elsevier.