Proceedings of the 21st International Conference on Global Research and Education (Inter-Academia 2024)

Improving Automated Behaviour Analysis in Zebrafish Laboratory Trials

V.C. Craciun[a,*], A.V. Panaintescu[a], R.M. Lefter[b] and M. Luca[c]

[a]*Faculty of Computer Science, "Alexandru Ioan Cuza" University of Iaşi, General Berthelot 16, 700483 Iaşi, Romania*
[b]*Center of Biomedical Research, Romanian Academy, Iaşi Branch, Carol I 8, 700506 Iaşi, Romania*
[c]*Institute of Computer Science, Romanian Academy, Iaşi Branch, Codrescu 2, 700481 Iaşi, Romania*

Doi: 10.12693/APhysPolA.146.483 *e-mail: vlad.craciun@info.uaic.ro

Studies on zebrafish are carried out to learn how specific health conditions might be resolved. The fact that the genomes of this species and humans are so similar makes this possible. These trials allow behaviour comparisons between healthy, sick, and cured subjects. When the behaviours of those healthy and healed match or show a high degree of similarity, researchers frequently find a new medicine. However, there are obstacles that researchers must overcome, such as pricey equipment, time-consuming analysis techniques, and unsuitable settings for analysis devices. To enhance our comprehension of the activities of the fish subjects, we created the *Framework for Activity Real-Time Monitoring*. This application makes real-time processing of video records possible, as well as simple calibration and reconfiguration. It also attempts to characterize the behaviour to facilitate comparisons across a broad spectrum of behaviours. In this study, we provide prospective methods for using k-means clustering applied to a set of 85 features from motion tracking routes to categorize fish behaviours. To assess our solution, we additionally create a public dataset (based on 95 zebrafish and 80 goldfish video recordings) and compare the behaviour of the two fish species.

topics: behaviour analysis, zebrafish, laboratory trials, clustering

1. Introduction

Behavioural analysis of fish has developed into a powerful tool for biomedical research on various behaviour paradigms, including social interaction, anxiety, learning, and memory. Fish species studied include zebrafish (Danio rerio) and goldfish (Carassius auratus) [1, 2]. Zebrafish are thought to be better laboratory animals than goldfish, mainly because of their fully sequenced genome and genetic tractability. Their quick development, which involves transparent embryos, enables real-time monitoring of developmental processes as well as an establishing database and protocol network [3, 4]. Conversely, research on toxicity and learning and memory, which involve a variety of intricate cognitive tasks, also frequently employ goldfish [5, 6]. Zebrafish have often been utilized more successfully in behavioural tests because of their size, convenience of handling, and the availability of automated tracking devices [7].

Software systems like ZebraLab, CleverSys, and EthoVision XT offer an integrated platform for video tracking and quantifying different behavioural parameters like time spent in particular zones, movement patterns, speed, and travelled distance. They are pricey, nevertheless, especially when it comes to additional hardware requirements, usage training, and restrictions on the target animal species [7]. In biomedical research, the state-of-the-art automated tracking currently used involves tracking behaviours such as erratic swimming for anxiety and depression tasks, preference for specific areas of the test apparatus, or behaviours reflecting conditioned avoidance and learning or spatial memory for cognitive tasks.

The monitoring tools are divided into two categories: behaviour analysis applied to the tracking [6, 8, 9] and actual motion tracking [10–14]. Many readily available tools created specifically for tracking depend on heavy software platforms or languages such as MATLAB, Java, and LabVIEW, as stated in [15]. These platforms could require extra setups and expenses. A software program called idtracker.ai is suggested in research published in 2019 [11] for the simultaneous automated tracking of up to 100 animal subjects. The video segmentation settings must first be defined manually in the application in order for it to monitor a large number of comparable recordings. Many challenges still exist, though, including the need to automatically adjust parameters for dynamic light, track the intricacy of fish movements during interrupted swimming [8], track targets in three dimensions [10], and — possibly the most challenging of all — integrate the disparate behavioural patterns in a comprehensive way to enable the standardization of particular fish phenotypes.

1.1. Problem

Studying the behaviour of animal subjects in laboratory trials is a fascinating area of research in the field of biomedicine. Researchers can utilize existing technologies to construct and compare statistical models for these subjects, encompassing three different groups: healthy, sick, and cured. They can then assess the extent to which the statistics of the cured group align with those of the healthy group. For comparing behaviours, one-way ANOVA is commonly utilized in various applications such as GraphPad Prism, IBM SPSS, or MATLAB, which typically require annual subscriptions costing around $1000. Working with new datasets, as well as considering past ones, can present challenges in managing and storing statistical models for future use. Moreover, categorizing behaviours based on subject movements and distinct tracking characteristics can be challenging. Although advancements in this area have incorporated machine learning to track and analyse behaviours, these methods are not yet suitable for widespread use due to their high cost and resource-intensive nature. For example, ZebraZoom [9] is freely accessible and utilizes an interesting combination of technologies for tracking and analysis. However, it does require MATLAB for operation. Additionally, although ZebraZoom utilizes machine learning techniques such as *support vector machine* (SVM) to categorize behaviours, it is unlikely that the model can be effectively applied to different scenarios.

1.2. Contributions

In this paper, we will be discussing our contribution in three different areas. Firstly, we successfully extracted the motion traces from 95 zebrafish and 80 goldfish video recordings. The traces are saved in JSON format and can be found on GitHub at vcraciun/Fish_Behaviour_Datasets. The repository also includes some of the plots and Python scripts to reproduce the results. The repository also contains scripts for simulating the original recordings, generating heatmaps, and analysing performance and behaviour. Additionally, our experiments involve behaviour extraction and classification using FARM (*Framework for Activity Real-Time Monitoring*) [16]. The application is briefly introduced in

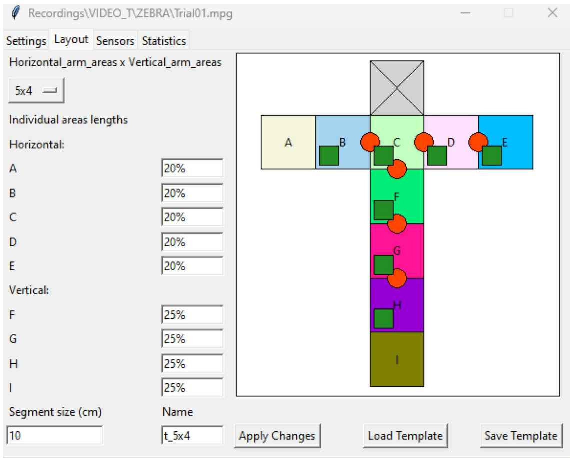

Fig. 1. FARM application layout tab for a zebrafish water tank (T shape); user can design custom layouts or load existing ones; green squares are sensors for time spent inside a box (A, B, C, etc.), the red circles are sensors to monitor box transitions.

Sect. 2. Next, we explored a k-means approach using a set of 85 features to categorize behaviour classes. Additionally, we compared these classes across the two datasets in Sect. 3. Our classification aims to analyse fish behaviours solely based on their tracking features. This enables researchers to accurately label and associate behaviours with specific health states.

2. Framework for Activity Real-Time Monitoring

FARM [16] is a Python framework developed to allow researchers to merge video processing for subject tracking with the functionalities of behaviour analysis. Although it shares similarities with idtracker.ai [11], it does not offer the same tracking capabilities, such as multiple species and CUDA optimization. However, it does provide the flexibility to dynamically adjust video processing parameters. With this application, researchers can easily customize various video processing parameters such as rotation, polygonal tracking area, threshold for noise reduction, experimental layouts, and sensors. Additionally, the application provides a set of scripts for behaviour analysis. Figure 1 represents a screenshot of the application featuring the layout tab, which is populated with sensors defined by the user. Once the video processing parameters are configured, the video processing can be seamlessly and automatically executed on a large set of video recordings. Some parameters may shift during the silent processing to enhance the level of detail in subject tracking.

2.1. Application design

FARM consists of two components, i.e., one for motion tracking and another for behavioural analysis. While the first component has the ability to handle minimal motion interaction, the second can conduct various analyses solely based on movement trace. Additionally, Fig. 2 represents a FARM screenshot from a video processing. in panel (a), we can observe the presence of the fish, while in panel (b), we can see the movement trace. The fish tank in this case has a shape and layout that closely resembles the one shown in Fig. 1.

The tracking data in Fig. 2b, is saved as a list of numeric pairs (x, y), representing the fish position. The fish position is captured in each frame, enabling the behaviour analysis component to engage in subsequent computations, ultimately determining the real elapsed time (based on the original video FPS — frames per second).

2.2. Behaviour analysis

The behaviour analysis component accepts the trace generated by FARM in the previous stage, together with a layout (which may be the same or different from the one used for video processing). It then either proceeds to simulate the trace or uses specialized behaviour categorization methods. Figure 3 represents a heatmap in panel (a) and the locomotion trace in panel (b) for a completed simulation. The heatmap serves as a graphical representation of the zones where fish is most frequently observed. One may reconstruct this image from the trace by considering both the fish's location and the time spent at specific coordinates.

While the layout shown in Fig. 1 is an experimental design, the one shown in Fig. 3 is derived from a series of observations and can be regarded as customary for our specific set of experiments, as well as typical for this type of T-shaped water-tank. As the fish exhibits a preference for the corners of the T shape, we have divided the former B box into B1 (30%) and B2 (70%), the former D and E boxes into D1 and D2, the former F and G into E1, and the former H into E2. In essence, the behaviour analysis considers B as the region characterized by both high and low levels of social activity, C as a transitory pathway, D as a remote contact or retreat zone and E as an exploratory area.

An investigation of behaviour was conducted by applying Time Series k-means clustering on a collection of 85 features. The features set was formed by combining a Cartesian product with an additional set consisting of the overall mean and maximum speed [cm/s] along with the fish's trajectory through the six walls (B1–B2, B2–C, C–D1, D1–D2, C–E1, E1–E2) in both directions. The Cartesian

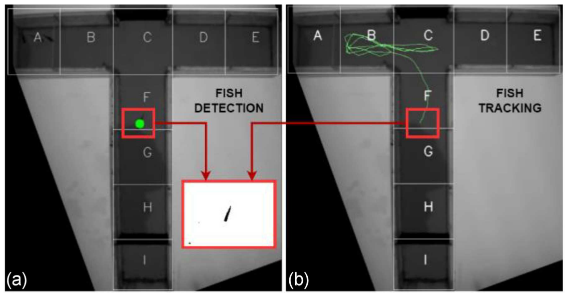

Fig. 2. (a) Fish detection, (b) fish tracking.

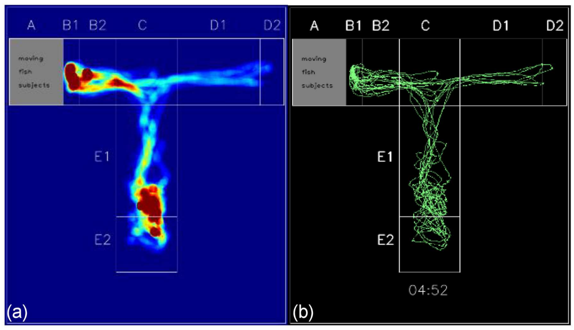

Fig. 3. (a) Heatmap plot, (b) trace simulation and total time spent in the original video recording.

product is computed between the set of boxes {B1, B2, C, D1, D2, E1, E2} and a set of properties {distance, time, sharp-turns, geo-directions}. We performed a silhouette analysis to automatically determine the optimal number of clusters for k-means, ranging from 2 to 30 in both zebrafish and goldfish datasets. This technique selects the first highest value for a given number of desired clusters. Figure 4 demonstrates that the k-Shape [17] distance identifies 6 clusters for zebrafish and 8 clusters for goldfish. We also conducted some experiments with other distance formulas such as Euclidean, *dynamic bandwidth allocation* (DBA), and *soft dynamic time warping* (soft-DTW). These three distances identify 3 clusters for zebrafish and between 11 and 25 clusters for goldfish. From the 4 distance computation formulas, we selected the k-Shape.

3. Evaluation

In this section, we assess the performance of the k-means clustering algorithm on a dataset comprising 95 traces for zebrafish and 80 traces for goldfish. The video recordings were captured using the EthoVision XT system, a commonly employed tool for tracking and analysing fish behaviour related to aggression and social interactions. The data collected in the way that EthoVision works, is compared

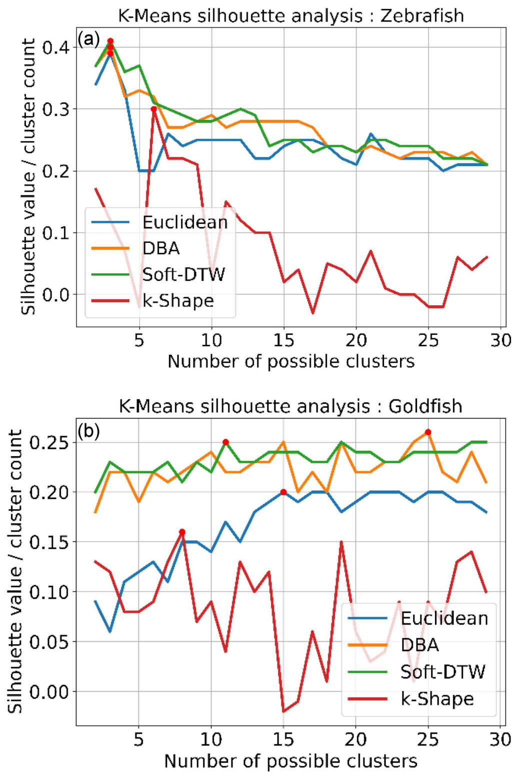

Fig. 4. Number of clusters recommended by silhouette analysis; (a) for zebrafish, the highest values for Euclidean, DBA, and Soft-DTW suggest 3 clusters, while the highest k-Shape value suggest 6 clusters; (b) for goldfish, the highest values for Soft-DTW, Euclidean, and DBA suggest between 11 and 25 clusters, while for k-Shape only 8 clusters seems to be an optimal value.

statistically across different health states of the fish (i.e., healthy, sick, and recovered). The original videos, recorded in grayscale at a resolution of 800×600, span durations of 3 to 5 min. Our primary objective was to analyse fish behaviour independent of their health status, as the recordings include a mixture of potential health conditions. Although EthoVision XT provides its own video processing and tracking data analysis, we conducted independent computations and analyses, utilizing the device solely for its camera; the recordings could have been obtained using a different camera or even a mobile phone. Our analysis involves translating the video data into a set of 85 features, followed by applying the k-means algorithm to group these features. Prior to feature extraction, a spatial layout is superimposed on the water tank (Fig. 3), enabling the computation of features such as time spent, transitions, total tracking distance, speed, sharp turns, and fish geo-direction, all in relation to the defined compartments of the layout. The video-recorded trials were conducted in accordance with established fish social interaction guidelines. In this setup, the top-left compartment included a separate section containing one to four fish subjects, allowing distant interaction with the primary fish. For example, repeated approaches by the main fish toward the invisible barrier on the leftmost side of the tank were indicative of increased social behaviour, whereas movement toward the rightmost arm was linked to fear-driven behaviour.

Although alternative clustering techniques, including statistical and rule-based approaches, were considered, this study primarily focuses on the application of k-means clustering. It is important to note that we do not have information about the initial condition of the fish (e.g., whether they were healthy, ill, or recovered), allowing the clustering to be fully based on the observed data without prior assumptions.

3.1. Behaviour evaluation

In Fig. 5, we emphasize the characteristics of the zebrafish clusters (shown in yellow) and their centroid (shown in red) for the k-Shape distance. We restructured the dataset by allocating 80% of it for training purposes (Fig. 5a) and the remaining 20% for testing purposes (Fig. 5b). The training dataset was shuffled, normalized using ScalerMeanVariance, and reduced to 40 features, which approximates half of the original size. An optimal random state value of 42 was selected for shuffling. Approximately 55% of the training dataset consists of the first cluster for zebrafish, whereas for goldfish, clusters 4 and 8 sum up to around 31% of the dataset. This finding indicates that more than a half of the zebrafish recordings exhibit a highly comparable behaviour. In comparison to the zebrafish, the goldfish shows a more homogeneous distribution (a four times lower standard deviation compared to zebrafish). The remaining test data consists of 19 zebrafish samples and 16 goldfish samples.

The first cluster, which has a higher hit-count in the train set, predicts 13 behaviours from the remaining test samples, as shown in Fig. 5b. Clusters 4, 5, and 2–6 yield predictions of 1 and 2 samples, respectively, however cluster 3 does not provide any predictions. The allocation of the samples within clusters exhibits a comparable pattern to those in the training dataset.

Figure 6 illustrates certain typical characteristics observed in the goldfish recordings. The reduced size of the whole dataset for this species results in smaller anticipated test clusters compared to zebrafish. Nevertheless, a comparable pattern can also be observed in the train/test sets of the two fish species. Compared to the training data, the distribution of the predicted data is remarkably comparable. The usage of k-Shape allowed us to reduce the number of clusters in goldfish recordings, however, in contrast to zebrafish, all the distances we experimented with, found larger clusters for

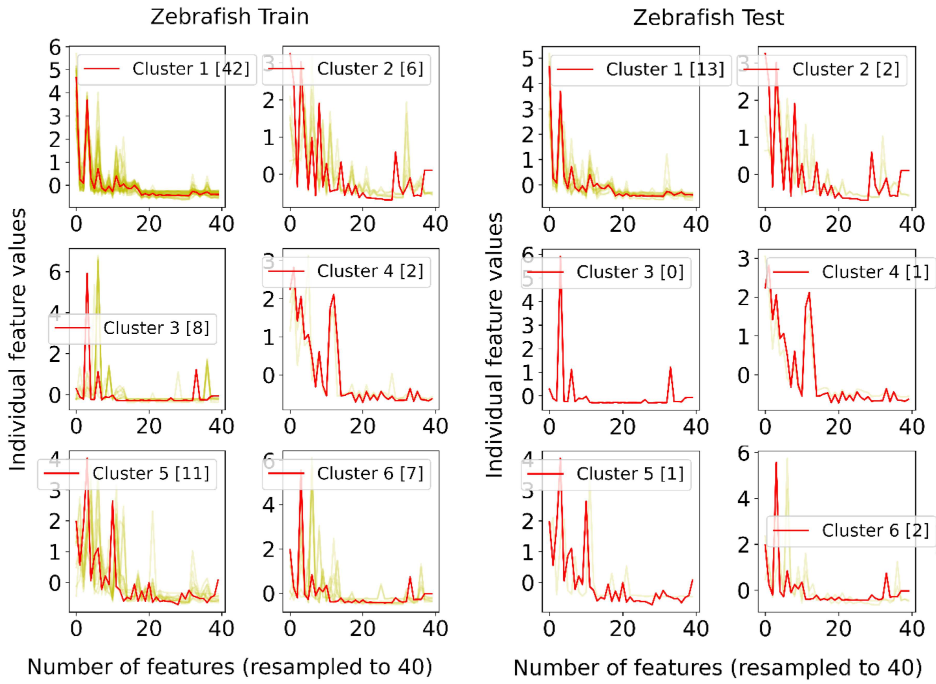

Fig. 5. Zebrafish train/test feature clusters using k-Shape: feature index (x-axis) vs feature value (y-axis); yellow — cluster features; red — cluster centroid.

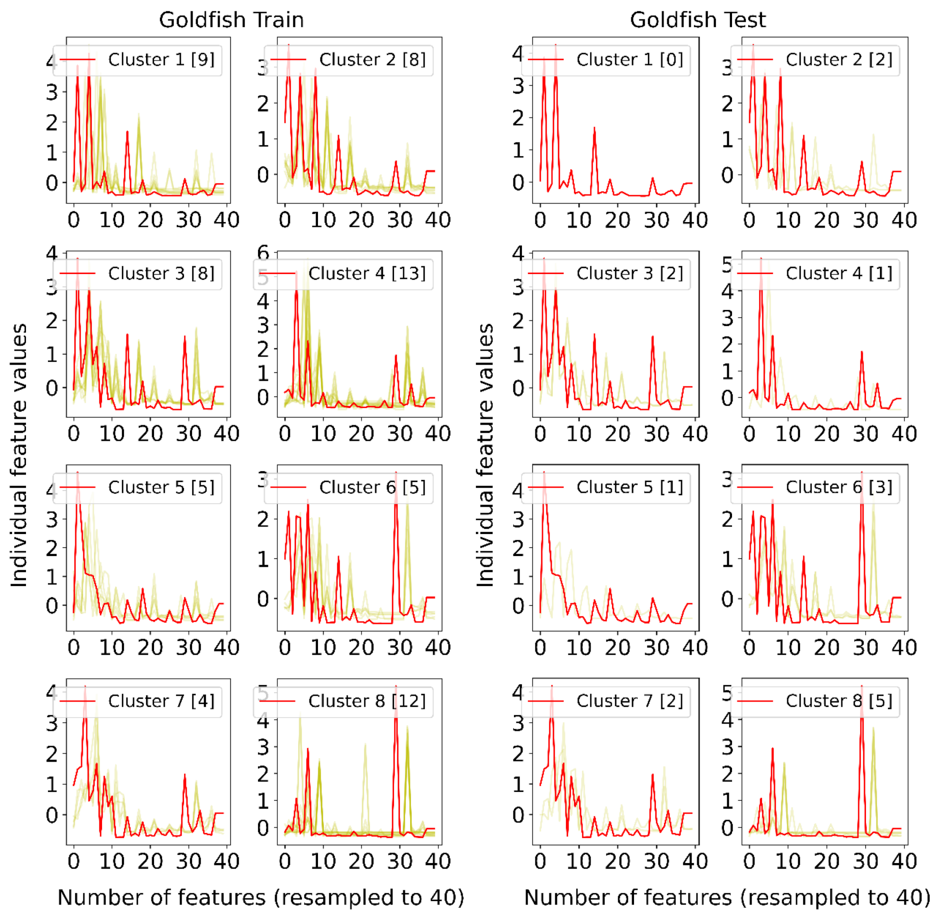

Fig. 6. Goldfish train/test feature clusters based on k-Shape; yellow — cluster features; red — cluster centroid; each image has specified the cluster index and the number of predicted samples.

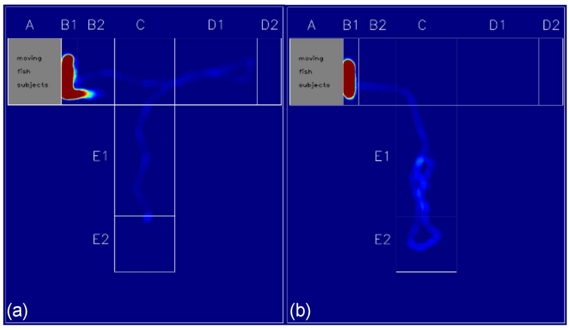

Fig. 7. Zebrafish resembled simulation images in first cluster; (a) train heatmap, (b) test heatmap.

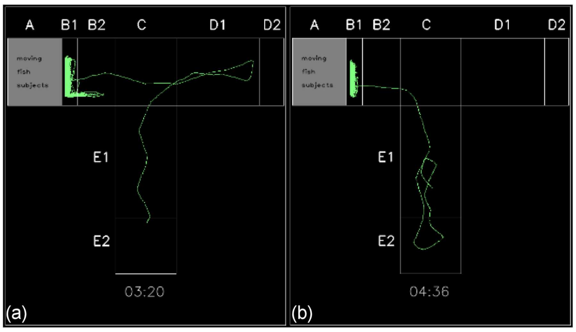

Fig. 8. Zebrafish simulation images for the heatmaps in Fig. 7; (a) train tracking, (b) test tracking.

goldfish (while the dataset was overall 15% smaller). This finding could imply that either goldfish have more complicated behaviours, or there were greater variations in the examined settings and subjects' health conditions.

3.2. Discussion

A random selection of one pair of heatmap and tracking simulation for zebrafish was made to compare the fish tracking heatmaps based on the chosen feature clusters. The initial set of heatmap/tracking images was chosen from the most extensive training clusters (cluster 0 in zebrafish), whereas the second image was taken at random from all the samples in the test cluster with the same label (0). Despite the absence of some of the properties (speed, positions, etc.) employed in k-means clustering, it is obvious that the heatmap and tracking of recordings in Figs. 7 and 8 exhibit numerous shared visual similarities. In these two videos, the zebrafish subject is observed to relocate to the B1 box for social contact and remains stationary until the recording ends. The recording in panel (b) is 1 min longer compared to the one in panel (a), but this does not seem to influence the behaviour.

4. Conclusions

This study introduces a k-means clustering methodology for conducting laboratory experiments on zebrafish and goldfish. The clustering analysis was performed on two datasets, each containing 85 features, with 95 recordings for zebrafish and 80 recordings for goldfish. We evaluated k-means clustering using various distance metrics, including Euclidean, DBA, Soft-DTW, and k-Shape. Among these, the k-Shape method produced tighter and more consistent clusters across both datasets. For our experiments, we split the original dataset into 80% for training and 20% for testing. Additionally, the features were normalized and resampled to reduce their size by approximately half. The test set predictions were then assigned to the clusters established during training, and the cluster indices were mapped back to the simulated tracking data for visual comparison.

As part of our ongoing development, we want to expand FARM to include the capability to target several species, enhance the performance of video processing, and increase flexibility in altering video processing parameters such as polygonal area shifts and thresholds. In relation to behaviour analysis, we aim to expand the collection of rules and the level of specificity, enabling researchers to select a model that aligns with their necessary criteria from a predetermined list. We anticipate that in the foreseeable future, researchers will allocate time to experiment with various configurations and strive to improve the behaviour tagging and extend the set of features.

References

[1] R.M. Al-Zoubi, H. Abu-Hijleh, A. Zarour, Z.Z. Zakaria, A. Yassin, A.A. Al-Ansari, M. Al-Asmakh, H. Bawadi, *Int. J. Mol. Sci.* **25**, 4895 (2024).

[2] D.T. Cerutti, E.D. Levin, J.J. Buccafusco, *Animal Models of Cognitive Impairment*, CRC Press, Boca Raton 2006, p. 279.

[3] T.Y. Choi, T.I. Choi, Y.R. Lee et al., *Exp. Mol. Med.* **53**, 310 (2021).

[4] T. Teame, Z. Zhang, C. Ran, H. Zhang, Y. Yang, Q. Ding, M. Xie, C. Gao, Y. Ye, M. Duan, Z. Zhou, *Animal Front.* **9**, 68 (2019).

[5] A. Gandar, S. Jean, J. Canal, N. Marty-Gasset, F. Gilbert, P. Laffaille, *Environ. Sci. Pollut. Res.* 23, 3184 (2016).

[6] C. Salas, C. Broglio, E. Durán, A. Gómez, F.M. Ocaña, F. Jiménez-Moya, F. Rodríguez, *Zebrafish* **3**, 157 (2006).

[7] C. Buske, R. Gerlai, *J. Neurosci. Methods* **234**, 66 (2014).

[8] M.O. Barreiros, F.G. Barbosa, D.O. Dantas, D.M.L.D. Santos, S. Ribeiro, G.C.O. Santos, A.K. Barros, *Sci. Rep.* **11**, 9330 (2021).

[9] O. Mirat, J.R. Sternberg, K.E. Severi, C. Wyart, *Front. Neural Circuits* **7**, 107 (2013).

[10] Z. Wu, C. Wang, W. Zhang, G. Sun, W. Ke, Z. Xiong, *Adv. Eng. Inform.* **61**, 102481 (2024).

[11] R.F. Francisco, B.G. Mattia, R.C. Hinz, J.H. Heras Francisco, G.G. de Polavieja, *Nat. Methods* **16**, 179 (2019).

[12] M. Pedersen, J.B. Haurum, S.H. Bengtson, T.B. Moeslund, in: *Proc. of the IEEE/CVF Winter Conf. on Applications of Computer Vision Workshops*, 2020, p. 2423.

[13] C. Xia, L. Fu, B. Fu, H. Li, in: *IEEE 3rd Int. Conf. on Computer Systems (ICCS)*, IEEE, 2023, p. 80.

[14] Q. Al-Jubouri, W. Al-Nuaimy, H.S. AlZu'bi, O. Zahran, J. Buckley, in: *14th UK Workshop on Computational Intelligence (UKCI)*, IEEE, 2014, p. 29.

[15] J.E. Franco-Restrepo, D.A. Forero, R.A. Vargas, *Zebrafish* **16**, 223 (2019).

[16] V.C. Craciun, M. Luca, R.M. Lefter, in: *Advances in Digital Health and Medical Bioengineering (EHB 2023), IFMBE Proc.)*, Vol. 109, Springer, Cham 2024, p. 511.

[17] J. Paparrizos, L. Gravano, in: *Proc. of the 2015 ACM SIGMOD Int. Conf. on Management of Data (SIGMOD '15)*, 2017, p. 1855.