

Deep Learning and feature selection in the bioinformatic modelling of functionally annotated microbial communities in aquaculture

J. Szyda^{1,2*}, M. Sztuka², M. Jakimowicz², K. Sidorczuk², D. Słomian³, and Ł. Napora-Rutkowski⁴

¹ Wroclaw University of Science and Technology, Department of Biomedical Engineering, Wroclaw, Poland; ² Wroclaw University of Environmental and Life Sciences, Department of Genetics, The Biostatistic Group, Wroclaw, Poland; ³ National Research Institute of Animal Production, Balice, Poland; ⁴ Polish Academy of Sciences, Institute of Ichthyobiology and Aquaculture, Zaborze, Poland; *Presenting and corresponding author: joanna.szyda@upwr.edu.pl

Introduction

To meet rising demand for aquaculture products and maximise production, feed additives such as growth promoters, vitamins, and antibiotics are widely used in aquatic feeds to combat pathogens and improve disease resistance (Banerjee and Ray, 2017). The common carp (*Cyprinus carpio*) is a key aquaculture species, particularly in Eastern and Central Europe. Although research on effective microorganism supplementation in common carp has been extensive for nearly two decades (Dawood and Koshio, 2016), little work has addressed practical on-farm use of effective microorganism products, especially mixed microbial communities rather than single bacterial species. Most existing studies used tank-based laboratory conditions. To our knowledge, the only reported field trial in common carp is by Mohammadian *et al.* (2022). While tank experiments are valuable and cost-effective, they omit many biological and environmental factors present in farms. Field experiments therefore yield more representative results and are more relevant for practical aquaculture applications.

Our study was conducted as a field experiment in earthen ponds to reflect practical fish rearing conditions while investigating how supplementation with effective microorganisms affects microbial diversity of fish intestine. Since microbial communities represent a set of genes and their products rather than a list of individual species (or higher taxonomic units), the biological objective of the study was to express the abundance of Amplicon Sequence Variants (ASVs) identified in fish intestines by ASVs metabolic functions, expressed by KEGG metabolic pathways (Kanehisa, 2004). The methodical objectives were twofold: (i) to assess the ability of Deep Learning (DL) based architectures to analyse microbiome abundance, in particular to assess whether the functional patterns underlying gut microbial communities differ between experimental groups, by jointly considering the metabolic landscape resulting from abundance of bacterial families, (ii) to evaluate the impact of particular KEGG pathways on model classification.

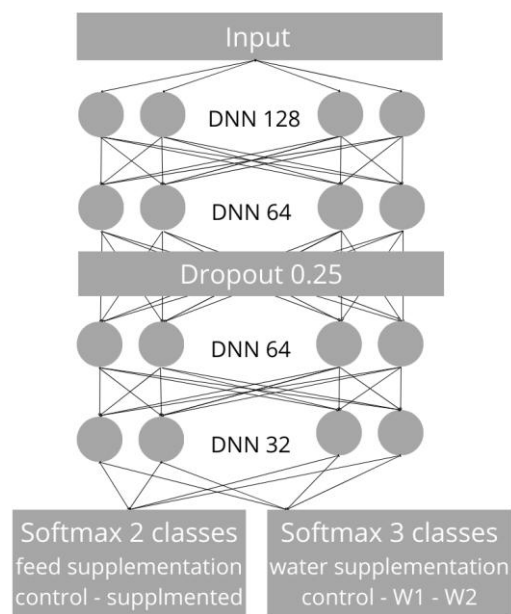
Materials & Methods

Experimental design. The data originated from an experiment set up to investigate the effects of feed and water supplementation of common carp (*Cyprinus carpio*) with commercially available effective microorganism supplements during fish growth period. The experiment spanned a full fish production season and was carried out in real-production conditions in 25 ponds. The ponds were divided into experimental groups with five ponds per group: **G0** - control conditions without supplementation, **G1** - water supplemented with W1 supplement, **G2** - water supplemented with W1 and fish feed supplements, **G3** - water supplemented with W2 supplement, **G4** - water supplemented with W2 and fish feed supplements. From each

pond, fish intestinal samples were collected at the end of the experiment, between 101 and 103 days after the first water supplementation.

Functional annotation. Diversity and abundance were assessed based on two hypervariable regions (V3 and V4) of the 16S rRNA gene. Sequence reads were processed by QIIME2 (Bolyen *et al.*, 2019) with the following steps: the assessment of raw read quality, trimming sequences with quality scores below 30 from 5' and 3' ends, removing sequencing adapters and reads shorter than 200 bp, merging paired end reads, additional quality filtering step and denoising. The resulting abundances of ASVs identified for each fish, were functionally annotated to KEGG Orthologs that were further used to predict the abundance of KEGG metabolic pathways using PICRUSt2 (Douglas *et al.*, 2020) with ggpicrust2 (Yang *et al.*, 2023) in R (R Core Team, 2021).

DL-based classification. The next step was to apply a DL-based classifier to assess whether the functional abundance pattern of KEGG pathways in intestinal samples allows to re-create the underlying experimental design. DL models were constructed, trained and evaluated using TensorFlow (v2.16.1) backend (Abadi *et al.*, 2016) implemented via Keras (v3.3.3) API (Chollet, 2015). Prior to classification, raw abundances were normalised using the Centred Log Ratio (Aitchison, 1982). The full dataset was divided into a training set with 5 validation subsets which contained 80% of the samples and a test set. The split was done separately within each experimental group to retain the original experimental sample proportions. Several classifiers, all based on the set of Dense Neural Networks, but with different sets of hyperparameters were applied. The hyperparameter space comprised the number of layers, the number of nodes per layer, activation functions and loss optimizers was searched manually



(i.e. without the application of a formal hyperparameter tuning scheduler), based on expert knowledge. The selection of hyperparameters was based on the accuracy metric defined as the number of correctly classified samples divided by the total number of class assignments. The model was trained by performing two classification tasks in parallel, but with the common model architecture. Training continued until the loss function failed to improve for 100 consecutive epochs, after which the weights from the best-performing model were saved. The best performing model was then applied for the classification of the test data set. It consisted of four dense layers with 128, 64, 64, and 32 nodes respectively, each using the Rectified Linear Unit (ReLU) activation function and L2 kernel regularizer with the regularization factor set to 0.01.

Figure 1. The best Deep Learning model for fish classification based on KEGG abundance.

A dropout layer with the dropout rate of 0.25 was placed after the second dense layer, however to enhance the model's robustness, the two intermediate layers maintain a consistent representation size of 64 nodes, thereby encouraging the generation of more general representations before reducing the dimensionality to 32 nodes in the subsequent layer (Figure 1). The final layer incorporated two parallel classification tasks – two classes for feed supplementation (control vs. supplemented) and three classes for water supplementation (control, W1, and W2). By sharing knowledge across tasks, the model can potentially learn

more robust features and improve its overall performance, especially when dealing with small datasets with low information content. The categorical cross entropy was used as the loss function which was optimized using the Adaptive Moment estimation (ADAM) optimizer (Kingma and Ba, 2014).

Model explainability. To assess the contribution of individual KEGG pathways to the classification, SHapley Additive exPlanations (SHAP) values (Shapley, 1953) were computed using the SHAP package (Lundberg and Lee, 2017). An explainer was constructed based on the best DL and SHAP values were calculated for each KEGG pathway. The final importance metric was calculated as the mean of the absolute SHAP values estimated for each classification. In the next step, the 1D K-means algorithm was applied to cluster the mean SHAP values into groups representing KEGG pathways that are important and non-important for classification, which is biologically equivalent to pathways that were markedly altered by the EM supplementation.

Results & Discussion

222 KEGG metabolic pathways were identified. The DL architectures exhibited signs of healthy training, expressed by a generally nearly monotonic decrease of the loss function and an increase in accuracy over epochs. For feed supplementation, a high test accuracy of 0.84 was obtained, while the test accuracy for classifying fish based on water supplementation was low, amounting to 0.44 (Figure 2). However, it has to be noted that the accuracy of a two class classification is always higher than multiple class classification.

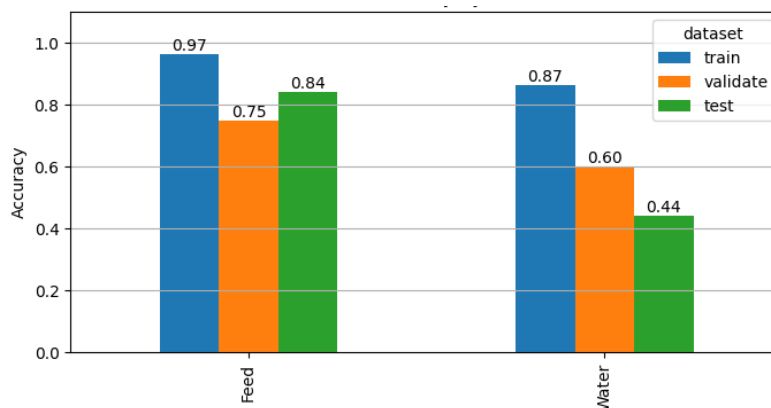


Figure 2. Classification accuracy.

Therefore the classification performance was also expressed in more detailed with confusion matrices. In particular, 3 out of 10 individuals from feed supplemented ponds were misclassified as control individuals, while only 1 out of 15 control individual was misclassified as belonging to the feed supplemented pond. For water supplementation the model exhibited problems in differentiating between W1 and W2 supplements, for which 10 individuals were misclassified (Figure 3).

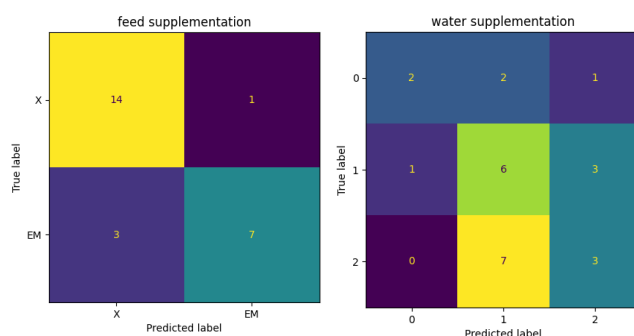
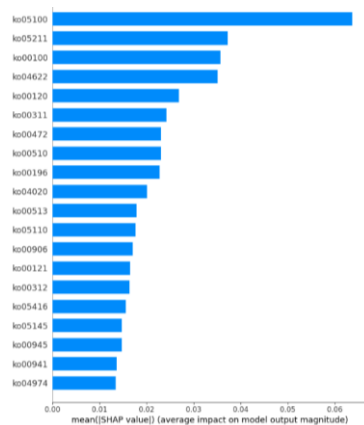


Figure 3. Confusion matrices underlying the classification to feed and water supplementation.



Based on the averaged SHAP values combined for the classification of feed and water supplementation, 20 pathways were selected as important (Figure 4). The most important pathway (ko05100) with the average SHAP value of 0.064 - almost twice as high as the next pathway on the list (0.037) is responsible for bacterial invasion of epithelial cells. The intestinal abundance of this pathway was reported by Zhu *et al.* (2024) as significantly increased after feed supplementation of bullfrogs with rosmarinic acid in bullfrogs. A response of ko05100 on nutrition supplements was also observed in laboratory mammalian species – mice (Ogita *et al.*, 2021) and rats (Lacombe *et al.*, 2013).

Figure 4. KEGG pathways selected as important for classification.

Conclusions

Biological conclusions indicate that the functional landscape of fish intestinal microbiome is mainly influenced by feed supplementation, while supplementing water does not markedly influence fish gut microbial communities. Methodically, a small number of observations (125 fish) impedes accurate statistical inferences. Although originally DL approach was designed to process very large data sets, after extensive model design and hyperparameter tuning, it offers an alternative for small data sets, albeit without P-value-based hypothesis testing opportunity.

References

- Abadi, M., Agarwal, A., Barham, P., Brevdo, E., Chen, Z., et al. (2015). Software available from tensorflow.org.
- Aitchison, J. (1982). *Journal of the Royal Statistical Society: Series B (Methodological)*, 44:139-160. doi:10.1111/j.2517-6161.1982.tb01195.x
- Banerjee, G., and Ray, A.K. (2017). *Research in Veterinary Science* 115:66–77. doi.org/10.1016/j.rvsc.2017.01.016
- Bolyen, E., Rideout, J.R., Dillon, M.R., Bokulich, N.A., Abnet, C.C., et al. (2019). *Nature Biotechnology* 37:852–857. doi.org/10.1038/s41587-019-0209-9
- Chollet, F. (2015). Keras. GitHub, San Francisco.
- Dawood, M.A.O., and Koshio, S. (2016). *Aquaculture* 454:243–251. doi.org/10.1016/j.aquaculture.2015.12.033
- Douglas, G.M., Maffei, V.J., Zaneveld, J.R., Yurgel, S.N., et al. (2020). *Nature Biotechnology* 38:685–688. doi.org/10.1038/s41587-020-0548-6
- Kanehisa, M., Goto, S., Kawashima, S., Okuno, Y., Hattori, M. (2004). *Nucleic Acids Research* 1:D277-80. doi.org/10.1093/nar/gkh063
- Kingma, D.P., and Ba, J.L. (2014). arXiv:1412.6980
- Lacombe, A., Li, R.W., Klimis-Zacas, D., Kristo, A.S., Tadepalli, S., et al. (2013). *PLoS One*. 8:e67497. doi.org/10.1371/journal.pone.0067497
- Lundberg, S.M., and Lee, S.I. (2017). *Proceedings of the 31st International Conference on Neural Information Processing Systems* 4766-4777
- Mohammadian, T., Monjezi, N., Peyghan, R., Mohammadian, B. (2022). *Aquaculture* 549:737787. doi.org/10.1016/j.aquaculture.2021.737787
- Ogita T, Namai F, Mikami A, Ishiguro T, Umezawa K, et al. (2021). *Frontiers in Nutrition* 8:701466. doi: 10.3389/fnut.2021.701466
- R Core Team, 2021. R: A Language and Environment for Statistical Computing.
- Shapley, L.S. (1953). *Annals of Mathematics Studies* 28:307–319
- Yang et al., 2023
- Yang, C., Mai, J., Cao, X., Burberry, A., Cominelli, F., et al. (2023). *Bioinformatics* 39:btad470, doi.org/10.1093/bioinformatics/btad470
- Zhu, B., Xu, S., Zhang, J., Xiang, S., and Hu, Y. (2024). *Fish & Shellfish Immunology* 150:109655. doi.org/10.1016/j.fsi.2024.109655