# Applications of Artificial Intelligence in Environmental Research: Automated Macroinvertebrate Identification

Nicole Jefferies



June 16, 2021

# Applications of Artificial Intelligence in Environmental Research
## Automated Macroinvertebrate Identification

**Nicole Jefferies**
*Fundamentals of Computational Intelligence*
Flinders University

jeff0194@flinders.edu.au

**Abstract**

Developments in artificial intelligence have influenced many scientific disciplines, including environmental research. A recent subject for this type of automation has been the identification of macroinvertebrates—which is considered a valuable, albeit inefficient aspect of bioassessment. This paper explores the development and limitations of automated macroinvertebrate identification technology, in addition to suggesting an alternate approach involving clear-carapaced organisms.

## 1 Introduction

As complex communities of living organisms and abiotic factors, ecosystems have a wide range of human-centred implications and are changing "more rapidly and extensively than in any comparable period of time in human history" (WHO, 2012). Thus, the development of more technologically advanced and resource efficient bioassessment methods is essential for both the conservation and more sustainable use of ecosystems. Macroinvertebrates—"macro" referring to visibility with the naked eye and "invertebrate" referring to the lack of backbone—are considered bioindicator organisms with the potential to facilitate this process (Dharan et al., 2010). Reasons for this classification of macroinvertebrates include their diversity in both species and pollution tolerance, wide-spread global abundance, ease of sampling, and ability to recover from sampling events (Ibid).

Historically employed macroinvertebrate identification tools such as dichotomous keys have strongly relied on small distinctions in external morphology, especially at a species level. A considerable quantity of time, costs, and taxonomic expertise is required for this process—making macroinvertebrate identification by humans too inefficient for rapid and reliable bioassessment (Milošević et al., 2020). This paper will explore the applications of artificial intelligence (AI) in the automation of macroinvertebrate identification in addition to the current limitations and possible directions of this AI-driven technology.

## 2 History and Development of AI-Driven Macroinvertebrate Identification

Tirronen et al (2009) were the first to propose that image-based macroinvertebrate identification could be automated through AI. They utilised a multi-class Support Vector Machine (SVM) classifier (developed through the combination of multiple binary class SVMs) to test this proposal on eight taxonomic groups of macroinvertebrates (Ibid). The 1529 images tested included the macroinvertebrate samples at random orientations, some of which were distorted, mutilated, or overlapping with another macroinvertebrate—which are all common issues that stem from macroinvertebrate fragility (Ibid). Despite these challenges regarding morphological preservation, the multi-class SVM correctly identified 88.17% of its training data and 75.31% of its validation data (Ibid). Thus, Tirrounen at al (2009) demonstrated that automated macroinvertebrate identification was a feasible possibility, and substantiated their claims with a morphologically diverse, albeit taxonomically limited dataset.

Subsequent research in this area has involved the application of alternate machine learning models to larger macroinvertebrate datasets. For example, Joutsijoki and Juhola (2012) compared the performance of a Directed Acyclic Graph Support Vector Machine (DAGSVM) model and a Directed Acyclic Graph k-Nearest Neighbour (DAGKNN) model on a 50 species dataset, finding that the DAGSVM model was slightly more accurate whereas the DAGKNN model was more user-friendly. Further research by Joutsijuko and Juhula (2015) using the DAGSVM model replicated the high classification accuracy results (approximately 80%) that were obtained in the Tirronen et al (2009) study. Although a larger sample size of 2156 images was used, this study was also limited to eight taxonomic groups—decreasing its applicability to more taxonomically-diverse macroinvertebrate datasets.

An alternate approach to this field of research was undertaken by Milošević et al. (2020), where a dataset containing only Chironomidae was used. Chironomidae (non-biting midges) are considered ‘dark taxa’, which are biotic groups that are exceptionally costly, time-consuming, or otherwise difficult to identify—resulting in them being omitted from most bioassessments and conservation studies (Ibid). The Milošević et al. (2020) study’s specialization of the training dataset to a single taxonomic family (Chironomidae) allowed for a species-level classification accuracy of 99.5%. The success of this study calls into question whether the automation of broader or more specialised macroinvertebrate identification has better implications for environmental research (Ibid).

## 3 Current Challenges and Opportunities

Although current research into AI-based macroinvertebrate identification has yielded promising results, it is important to consider the limitations of this technology—many of which can be attributed to the novelty of the field (relative to subsets of other scientific disciplines). Firstly, image-based macroinvertebrate identification through machine learning has yet to be conducted on a large-scale, taxonomically diverse dataset. The creation of such a model could require thousands to hundreds of thousands of macroinvertebrate images. In the case of SVM-based models, overfitting—the creation of a model that too closely fits its training data—becomes a significant hurdle to overcome as this weakens its ability to generalise, and subsequently its performance when identifying macroinvertebrates (Joutsijoki and Juhola, 2012). Additionally, the creation of a more broadly applicable model would substantially increase the computational power required to run it. There is literature to suggest that training large AI models, such as those of natural language processing, can emit over 200,000 kg of carbon dioxide—the equivalent carbon lifetime of five cars (Nishant, Kennedy and Corbett, 2020). Hence, the carbon footprint of this hypothetical macroinvertebrate identification model becomes a potential area of concern, especially in the context of environmental research. Further consideration of computational logistics and required resources would be necessary to determine both the feasibility and the scientific need of such a model.

Another significant limitation to consider is the value of taxonomic identification alone for bioassessment purposes. For example, the Saprobian System—which utilises species-level identifications to determine environmental oxygen deficits—is unable to identify specific forms of organic pollution or illustrate their impact on a linear scale (Bonada et al., 2006). The examination of multiple biological traits, on the other hand, is capable of providing a more statistically grounded understanding of specific human impacts on ecosystems (Ibid). Thus, a possible direction for automated macroinvertebrate identification could be to focus on clear-carapaced macroinvertebrates such as Cladocera. These macroinvertebrates, often called ‘water fleas’, are crustaceans whose bodies are enclosed by a thin, uncalcified, and subsequently see-through outer shell (Ebert, 2014). As their internal organs are readily visible from photographs, AI-driven models could be trained to generate biological trait data regarding size, reproductive behaviours, and the presence of specific parasites.

## 4 Conclusion

The use of AI in the field of environmental research for macroinvertebrate identification and bioassessment purposes is a promising, albeit recent development. Hence, it is important not to overestimate the capabilities of AI, and appropriately consider the computational logistics and resources required to create a particular model. The creation of a taxonomically all-encompassing model, for example, would be a costly, time-consuming endeavour with a considerable carbon footprint. Thus, taxonomically specified models such as that of Milošević et al. (2020) are likely to be the direction of future research undertaken in this field.

## References

Bonada, N., Prat, N., Resh, V.H. and Statzner, B. (2006). Developments in Aquatic Insect Biomonitoring: A Comparative Analysis of Recent Approaches. Annual Review of Entomology, 51(1), pp.495–523.

Dharan, M., Chellappa, S., Sundar, S. and Raja, M. (2010). Macroinvertebrates as Potential Indicators of Environmental Quality. Journal of Environmental Quality, pp.23–28.

Ebert, D. (2014). Introduction to Daphnia Biology. [online] NCBI. Available at: https://www.ncbi.nlm.nih.gov/books/NBK2042/.

Joutsijoki, H. and Juhola, M. (2012). DAGSVM vs. DAGKNN: An Experimental Case Study with Benthic Macroinvertebrate Dataset. Machine Learning and Data Mining in Pattern Recognition, pp.439–453.

Joutsijoki, H., Siermala, M. and Juhola, M. (2014). Directed acyclic graph support vector machines in classification of benthic macroinvertebrate samples. Artificial Intelligence Review, 44(2), pp.215–233.

Milošević, D., Milosavljević, A., Predić, B., Medeiros, A.S., Savić-Zdravković, D., Stojković Piperac, M., Kostić, T., Spasić, F. and Leese, F. (2020). Application of deep learning in aquatic bioassessment: Towards automated identification of non-biting midges. Science of The Total Environment, 711, pp.135–160.

Nishant, R., Kennedy, M. and Corbett, J. (2020). Artificial intelligence for sustainability: Challenges, opportunities, and a research agenda. International Journal of Information Management, 53, pp.102–104.

Tirronen, V., Caponio, A., Haanpää, T. and Meissner, K. (2009). Multiple Order Gradient Feature for Macro-Invertebrate Identification Using Support Vector Machines. Adaptive and Natural Computing Algorithms, 9, pp.489–497.

WHO (2012). Ecosystem Goods and Services for Health. [online] World Health Organisation. Available at: https://www.who.int/globalchange/ecosystems/en/ [Accessed 19 Mar. 2021].