Journal Pre-proof

Human experts vs. machines in taxa recognition

Johanna Ärje, Jenni Raitoharju, Alexandros Iosifidis, Ville Tirronen,
Kristian Meissner, Moncef Gabbouj, Serkan Kiranyaz,
Salme Kärkkäinen

PII: S0923-5965(20)30113-2
DOI: https://doi.org/10.1016/j.image.2020.115917
Reference: IMAGE 115917

To appear in: *Signal Processing: Image Communication*

Received date : 16 May 2019
Revised date : 2 April 2020
Accepted date : 11 June 2020

Please cite this article as: J. Ärje, J. Raitoharju, A. Iosifidis et al., Human experts vs. machines in taxa recognition, *Signal Processing: Image Communication* (2020), doi: https://doi.org/10.1016/j.image.2020.115917.

Highlights

- Multiple image data of real specimens of benthic macroinvertebrates
- Taxonomic resolution added to the image data
- Comparing classification results for human experts vs. machines
- Comparing those results taking into account the taxonomic resolution
- Hierarchical classification of the image data with inherent hierarchical structure

Human experts vs. machines in taxa recognition

Johanna Ärje[a,b](✉), Jenni Raitoharju[b], Alexandros Iosifidis[c], Ville Tirronen[d], Kristian Meissner[e], Moncef Gabbouj[b], Serkan Kiranyaz[f], Salme Kärkkäinen[a]

[a] *Department of Mathematics and Statistics, University of Jyvaskyla, P.O. Box 35 (MaD), FI-40014 University of Jyväskylä, Finland, johanna.arje@gmail.com*

[b] *Unit of Computing Sciences, Tampere University, Korkeakoulunkatu 1, FI-33720 Tampere, Finland*

[c] *Department of Engineering, Aarhus University, Inge Lehmanns Gade 10, DK-8000, Aarhus C, Denmark*

[d] *Faculty of Information Technology, University of Jyvaskyla, P.O. Box 35, FI-40014 University of Jyväskylä, Finland*

[e] *Programme for Environmental Information, Finnish Environment Institute, Survontie 9A, 40500 Jyväskylä, Finland*

[f] *Department of Electrical Engineering, Qatar University, Doha, Qatar*

Abstract

The step of expert taxa recognition currently slows down the response time of many bioassessments. Shifting to quicker and cheaper state-of-the-art machine learning approaches is still met with expert scepticism towards the ability and logic of machines. In our study, we investigate both the differences in accuracy and in the identification logic of taxonomic experts and machines. We propose a systematic approach utilizing deep Convolutional Neural Nets and extensively evaluate it over a multi-pose taxonomic dataset with hierarchical labels specifically created for this comparison. We also study the prediction accuracy on different ranks of taxonomic hierarchy in detail. We compare the results of Convolutional Neural Networks to human experts and support vector machines. Our results revealed that human experts using actual specimens yield the lowest classification error ($\overline{CE} = 6.1\%$). However, a much faster, automated approach using deep Convolutional Neural Nets comes close to human accuracy ($\overline{CE} = 11.4\%$) when a typical flat classification approach is used. Contrary to previous findings in the literature, we find that for machines following a typical flat classification approach commonly used in machine learning performs better than forcing machines to adopt a hierarchical, local per parent node approach used by human taxonomic experts ($\overline{CE} = 13.8\%$). Finally, we publicly share our unique dataset to serve as a public benchmark dataset in this field.

Keywords: hierarchical classification; taxonomy; convolutional neural networks; taxonomic expert; multi-image data; biomonitoring

1. Introduction

Due to its inherent slowness, traditional manual identification has long been a bottleneck in bioassessments (Fig. 1). The growing demand for biological monitoring and the declining funding and number of taxonomic experts is forcing ecologists to search for alternatives for the cost intensive and time consuming manual identification of monitoring samples [6, 28]. Identification of taxonomic groups in biomonitoring of, e.g., aquatic environments often involves a large number of samples, specimens in a sample, and the number of taxonomic groups to identify. For example, even in relatively species-poor regions like Finland, the calculation of the EU Water Framework Directive related indices often involves hundreds of individual specimens from 118-349 lotic diatom taxa and 44-113 lotic benthic macroinvertebrate taxa [1].

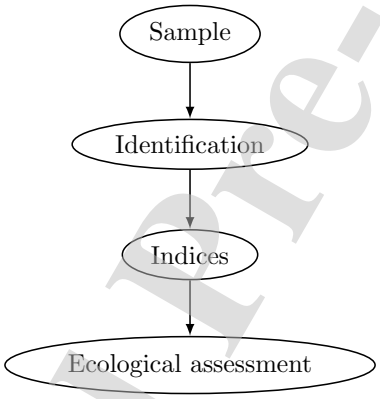

Figure 1: A schematic of the biomonitoring process.

While a growing body of work has used different genetic tools [e.g. 12, 39] for species identification, these methods are not yet standardized or capable of producing reliable abundance data currently required in, e.g., Water Framework Directive. While we have also worked on genetic approaches and acknowledge the great promise that genetic taxa identification methods hold [e.g. 14], we will not explore them here but alternatively examine the suitability of machine learning techniques on image data for routine taxa identification.

Many studies on automatic classification of biological image data have been published during the past decade. Yousef Kalafi et al. [38] have done an extensive review on automatic species identification and automated imaging systems. Classification methods for aquatic macroinvertebrates have been proposed in several studies [e.g. 11, 25, 20, 4, 17, 31]. The most popular classification methods used for identification of biological image data, such as insects, are deep neural networks and support vector machines [19] which are also applied in this work.

Despite the potential of computational, as well as DNA methods for taxa identification, some taxonomists continue to object the shift from manual to

novel identification methods [18, 23]. Often biologists that take a cursory look at automated identification tend to mistrust computational methods because they observe that a classifier is unable to separate two specimens which to them are clearly different to the human eye. Similarly, experts are baffled when the same classifier is able to discriminate between two specimens from low-resolution images while they as taxonomic experts cannot. This mismatch in the ability of computers to identify taxa observed for single cases is often mistakenly extrapolated into an overall unreliability of algorithms. But how different truly is both the logic used and the overall accuracy of taxonomic experts and algorithms?

Only few studies assess the accuracy of human experts and automatic classifiers, and their consequences on aquatic biomonitoring. In a study on human accuracy, Haase et al. [13] reported on the audit of macroinvertebrate samples from an EU Water Framework Directive monitoring program. They found a great discrepancy between the experts determining the true taxonomic classes and the audited laboratory workers. Contrastingly, in a study on the effect of mistakes made in automated taxa identification on biological indices, Ärje et al. [2] found a relatively small impact. Literature on direct human versus machine comparisons in classification tasks in an aquatic biomonitoring context is equally scant and ambiguous. Culverhouse et al. [10] compared human and machine identification of six phytoplankton species using images and noted a similar average performance for both the experts and a computer algorithm. In Lytle et al. [25], automatic classifiers outperformed 26 humans (a mix of experts and amateurs) when distinguishing between two stonefly taxa. Given these contrasting results, we feel it is necessary to simultaneously examine the effect of taxonomic hierarchy and of using human logical pathways for human and computer-based identification.

Taxonomic experts identify specimens based on a predefined taxonomic resolution while automatic classifiers operate on the information of taxonomic rank used in the training data. There are different ways for accounting for data hierarchy, such as taxonomy, in classification. Hierarchical classification is widely investigated in the current literature. Silla and Freitas [34] sought to describe and unify the concepts of methods used in hierarchical classification problems from different domains. Using the existing literature, they categorized the classification approaches into: 1) flat classification, where the classification is performed at the most specific (deepest) rank of the taxonomy which may not always be species level, 2) local classification per level, per node or per parent node, and 3) global classification, where the whole hierarchical structure of taxonomy is taken into account at once. They found that the existing literature suggested any local or global hierarchical classifier performed better than a flat classifier, if the performance measure was specifically designed for a hierarchical structure.

Several subsequent studies have compared flat classifiers to hierarchical classifiers. Rodrigues et al. [33] did not find a significant difference between flat and hierarchical approaches in classification of points-of-interest for land-use analysis whereas Levatic et al. [24] found that the use of hierarchy and multi-label structure improved classification results when compared to single-label cases. Babbar et al. [5] performed a theoretical study on the difference between flat and hierar-

chical classification and found that for well-balanced data flat classifiers should be preferred, whereas hierarchical classifiers are a better for unbalanced data.

Automatic classification of benthic macroinvertebrates, as well as plankton, has received increasing attention in recent years. However, most of the previous studies have focused on single-image data [see e.g. 3, 20, 4, 17, 35, 22, 2] and have not taken the inherent hierarchical structure of the data into account. In single-image data studies, the posture of the specimens can have substantial impact on the classification. Besides Lytle et al. [25], an imaging system producing multiple-image data is presented in Raitoharju et al. [31]. In this paper, we present a comparison of taxonomic experts and automatic classification methods on a benthic macroinvertebrate data that incorporates information on the taxonomic resolution. We test flat classifiers, local per level classifiers, and hierarchical top-down classification, i.e., local classification per parent node, and perform the automatic classification using convolutional neural networks (CNNs) and support vector machines (SVMs). The results are compared with the results of a proficiency test organized for human taxonomic experts and with a test where taxonomic experts used the same images as the automatic classifiers. The comparisons evaluate traditional single level accuracy and additionally use a novel variant of an accuracy measure that accounts for the hierarchical structure of the data.

2. Theory

2.1. Hierarchy in classification

Silla and Freitas [34] unified the concepts of methods used in hierarchical classification problems, and in this section we follow their terminology.

Human experts base visual identification of, e.g., invertebrate taxa on rules defined in the International commission on zoological nomenclature [15]. Therefore, human experts can be thought of as hierarchical, local per parent node classifiers (see Fig. 2c) that first identify the order of the specimen, then the family, genus, and species. The classification task is not necessarily a single level problem as some taxa need to be identified to different taxonomic levels (see Fig. 2a) either because of predefined rules, such as minimal taxonomic requirements, or as a function of necessity when specimens lack characteristics needed to allow for better resolution. While for some taxa, genus or family might be enough, others might require species level identification depending on what the taxa information is later used for.

Usually, automatic classification methods have no information on the possible hierarchical nature of the data. The classifiers simply aspire to identify the specimens to the class labels provided in the training data. In the case of benthic macroinvertebrate data, the class labels represent a mix of families, genera, and species. An algorithm working this way is called a flat classifier as it is not aware that species A and B belong to the same genus A, but uses the same approach to distinguish them from each other as when separating species A from genus B. Flat classification produces a single label prediction for each

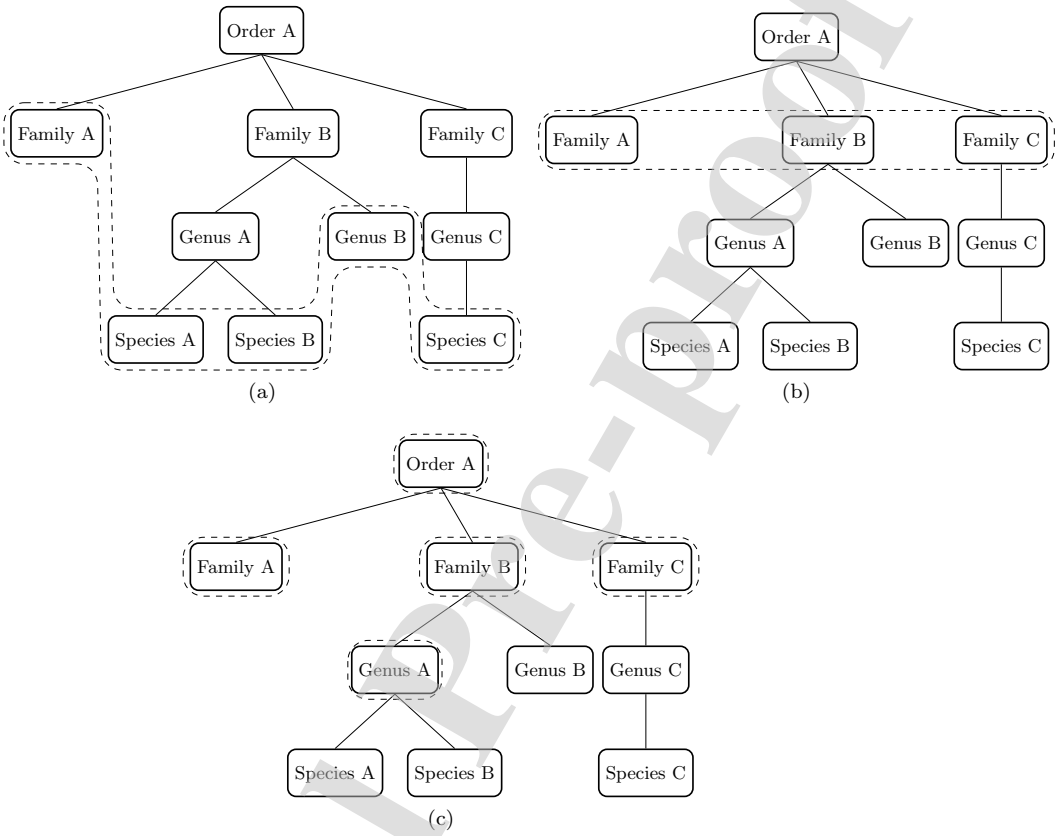

Figure 2: Different types of classifiers for hierarchical data: (a) Flat classification, (b) Local classification per level, (c) Local classification per parent node. The dashed boxes represent a single trained classifier.

specimen but the hierarchical level of that label may vary depending on the data (Fig. 2a).

Depending on what the taxa information is later used for, it could be beneficial to build a classifier that identifies a certain taxonomic rank well. For example, a common biological index used in river macroinvertebrate biomonitoring is the number of typical EPT families (*Ephemeroptera*, *Plecoptera*, *Trichoptera*). For the purpose of evaluating this index, it would be reasonable to train a classifier to identify the family level with high accuracy. However, such a classifier trained with the family level labels would have no intrinsic information on certain families descending from the same order. This type of a classification scheme is known as local classification per level (see Fig. 2b). One could build a classification system with local level classifiers for each level of the hierarchy. While such a system would predict multiple labels for each specimen there would

be no guarantee that the predictions for the different levels are taxonomically coherent.

It is also possible to build a hierarchical classification system that accounts for the hierarchical nature of the data and force it to operate in the same manner as human experts. This requires to build a sequence of several classifiers: i) an order level classifier to predict the order of each specimen, ii) multiple family level classifiers, one for each possible order present in the data, iii) multiple genus level classifiers, one for each family present in the data, and finally, iv) multiple species level classifiers to predict the species within each genus. This type of a hierarchical classification scheme is known as local classification per parent node and it predicts the labels for each rank of taxonomic resolution for all the specimens in the data (see Fig. 2c). While a human-like hierarchical classifier is guaranteed to logically follow taxonomy all classification errors made on higher levels of hierarchy will propagate to the lower level predictions.

The focus of this work is on the comparison of identification results obtained by taxonomic expert logic and machine logic. As traditional machine logic uses flat classification and taxonomic expert logic can be thought of as local classification per parent node, we will not consider global hierarchical classifiers.

2.2. Performance measures

Traditionally, classification methods are compared based on their accuracy, which is the proportion of correct predictions, or classification error (CE),

$$CE = \frac{1}{n}\sum_{i=1}^{n} L(\hat{y}_i, y_i),$$

where $L(\cdot,\cdot)$ is a 0-1 loss function and n is the total number of observations. Other measures of performance such as false positive rate, false negative rate, sensitivity, and specificity can also be calculated from the confusion matrix and take single label predictions into account. These performance measures can be calculated for both flat classification (Fig. 2a) or for each level of local classification (Fig. 2b, 2c).

With hierarchical data, each observation has multiple labels and we need to measure the performance as a whole accounting for all the labels. Verma et al. [37] presented context sensitivite loss (CSL) function which takes the top-down success into account. They used this loss function to define context-sensitive error (CSE),

$$CSE = \frac{1}{nH}\sum_{i=1}^{n} L(\hat{y}_i, y_i),$$

where

$$L(\hat{y}_i, y_i) = \begin{cases} h, \text{ where } h \text{ is the height of the deepest} \\ \quad \text{common ancestor of pair } (\hat{y}_i, y_i) \\ 0, \text{ if } \hat{y}_i = y_i \end{cases}$$

and H is the total number of levels in the hierarchy.

Because the deepest available level of hierarchy can vary in taxonomic data, we propose to modify the measure to a level-aware context-sensitive error (LCSE),

$$LCSE = \frac{1}{n}\sum_{i=1}^{n}\frac{1}{H_i}L(\hat{y}_i, y_i),$$

where $L(\hat{y}_i, y_i)$ is as above and H_i is the number of available levels in the hierarchy for observation i.

3. Materials and methods

3.1. Proficiency test for human experts

In order to compare automatic and manual classification, we needed classification results on the same set of taxa for both. The Finnish Environment Institute (SYKE), an appointed National Reference Laboratory in the environmental sector in Finland, organized a proficiency test on taxonomic identification of boreal freshwater lotic, lentic, profundal, and North-Eastern Baltic benthic macroinvertebrates in 2016. The aim of the test was to assess the reliability of professional and semi-professional identification of macroinvertebrate taxa routinely encountered during North-Eastern Baltic coastal or boreal lake and stream monitoring [26]. A part of the proficiency test included 10 participants who all identified a different set of 50 specimens of lotic freshwater macroinvertebrates belonging to a total of 46 taxonomic groups, of which 39 are in common with the multiple-image data introduced in the following Section 3.2 (see taxa list in Table .3). The samples sent out to the participants included 0–4 specimens of each taxa. The class labels of the 39 overlapping taxa consisted of 26 species, 12 genera, and one family. The chosen taxonomic resolution is based on the requirements for the Finnish national freshwater monitoring program for macroinvertebrates [16]. The 'true' labels of the specimens were predetermined by an expert panel and the specimens were shipped to the participants. Participants were provided with the list of the almost 300 possible taxa labels [26].

3.2. Image data

We produced all images with a new imaging system described in Raitoharju et al. [31] that allows for multiple images per specimen. The system is illustrated in Fig. 3. It consists of two Basler ACA1920-155UC cameras (frame rate of 150 fps) with Megapixel Macro Lens (f=75mm, F:3.5-CWD<535mm) placed at a 90 degree angle to each other, a high power LED light and a cuvette (i.e. a rectangular test tube) in a metal container. The device is sealed with a lid to block any extra light. The imaging system has a software that builds a model of the background of the cuvette filled with alcohol and sets off the cameras when a significant change in the view of the camera is detected. When a macroinvertebrate specimen is put into the cuvette, it sinks and both cameras

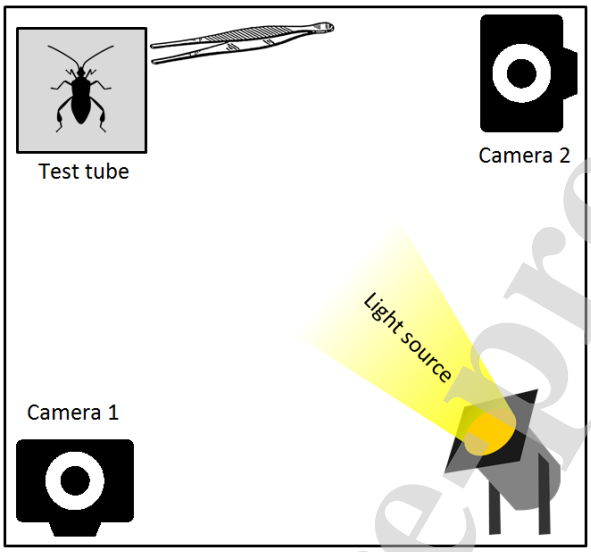

Figure 3: Schematic of the imaging system for macroinvertebrates pictured from above.

take multiple shots of it (Fig. 4). The number of images per specimen depends on the size and weight of each specimen: Heavier specimens sink faster, leading to a smaller number of images. Compared to the system and data described in Raitoharju et al. [31], we have improved the system to handle more than two images per specimen.

In Finland, the national reference taxa list determines the taxonomic ranks to which human experts are required to identify specimens from monitoring samples [16]. In the human proficiency test only a subset of the taxa from the national reference taxa list is used. The choice of the specific taxa and specimens used in the proficiency test is determined both by relevance of the taxa in national assessment indices, the availability of adequate testing material and to a lesser degree the inclusion of easily misidentified taxa. Human participants were required to key 50 specimens in total for the river benthic subtest [26]. Using the described imaging device, the Finnish Environment Institute compiled a new image database of 126 lotic freshwater macroinvertebrate taxa and over 2.6 million images. This data has 39 taxa overlapping with those present in the human proficiency test which are therefore used in the current work (Table .3). We restricted the number of images per specimen to a maximum of 50 images for computational reasons. If a specimen had more images from both cameras combined, we randomly selected 50 of them. The final data comprises 9631 observations and a total of 460004 images belonging to 39 taxa at the deepest available taxonomic rank. In total, considering one taxonomic rank at a time, the data consists of 7 orders, 23 families, 30 genera, and 26 species (see Fig. 5). The number of specimens for each taxa and the taxonomic resolution are shown in Table .3. The image resolution for this data varies from 32×20 pixels to 468×540

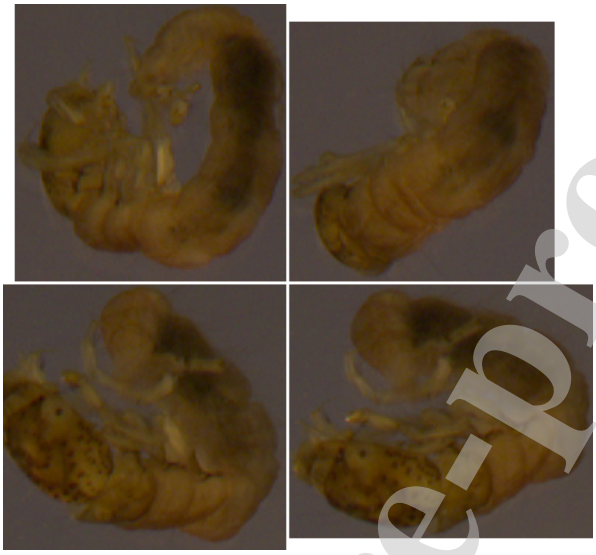

Figure 4: Example images of a *Polycentropus flavomaculatus* specimen from two cameras. The top row images are from camera 1 and the bottom row images from camera 2.

pixels. The 'true' labels for the specimens were defined by a group of taxonomic experts. While we acknowledge that there might be some mislabeled specimens, combining the knowledge of multiple taxonomic experts should improve the accuracy [7]. We provide the data for public use as FIN-Benthic2 in `https://etsin.fairdata.fi/dataset/a11cdc26-b9d0-4af1-9285-803d65a696a3`.

3.3. Classification set-up

To have classification results comparable to the proficiency test, we compiled a set of data divisions for the image data with the exact same number of test specimens per taxa as in the proficiency test. As the proficiency test had 10 participants identifying lotic freshwater macroinvertebrates, we created 10 data divisions. The test sets comprise randomly selected 45–46 specimens belonging to the 39 taxonomic groups present in both the physical data and the image data. The test sets have an approximately equal number of specimens from each class. We divided the rest of the specimens of each data split for training (80 %) and validation (20 %). Due to the nature of the collected data, the training and validation data are unbalanced. In the following sections, these data sets are referred to as the "comparison data". The number of specimens per test set in the comparison data is lower than in the proficiency test because 4–5 specimens sent to each participant belonged to taxonomic groups not present in the image data.

Since the comparison between professionals and semi-professionals analysing physical data with a laboratory microscope and automatic classifiers using image data is unequal, we asked the proficiency test participants to also try to identify

Figure 5: Taxonomic resolution and distribution of the multilabel image data. The area of the slices represent the relative size of each taxonomic group at the different ranks of taxonomic hierarchy.

taxa from the test images of the comparison data. Each participant received one of the test sets and a list of the 39 possible taxa labels. To avoid fatigue and to encourage more experts to participate, we restricted the number of images per test specimen to 10. The automatic classifiers used exactly the same test data. In addition, because some of the images are fuzzy, the experts were allowed to classify the taxa to a higher taxonomic rank if they were unsure. The automatic classifiers always predicted the classes of the test specimens to the deepest available rank of taxonomic resolution. Of the ten experts participating in the proficiency test, three volunteered to take part in this image classification study.

As the comparison test sets are very small, we also studied the performance of the automatic classifiers on larger test sets. We split the specimens randomly into training (70 %), validation (10 %) and test (20 %) data 10 times. This time the number of specimens in each taxon varied in all training, validation, and test sets depending on the size of the taxa in the dataset. We refer to these sets as the "machine learning data" as the splitting is typical for machine learning, but not suitable for comparisons with humans. For the test sets in the machine learning data, we included all images (max. 50) per specimen.

We considered different approaches to take the hierarchical nature of the data into account: A flat classifier is a single classifier with the 39 taxa as output

labels. Local per level classifiers are built for each taxonomic rank separately: a classifier for the orders and another classifier for the families. We only trained local per level classifiers for the two highest taxonomic ranks as some of the taxa in the data have information only on these ranks. The top-down, local per parent node classifier is a system comprising 17 classifiers: one classifier at the top to identify the order of a specimen, four classifiers at the family level as there are four families with more than one genus within them, five classifiers at the genus level, and seven classifiers at the species level (see Table .3). Some of the specimens get their predictions already at the order level since there are three orders with only one family or genus within them. In the data, there are two genera (*Leuctra sp.* and *Nemoura sp.*) for which only some of the specimens have information on species (*Leuctra nigra* and *Nemoura cinerea*). To separate these groups with the local per parent node classification approach, we temporarily marked the species for the rest of the *Leuctra sp.* and *Nemoura sp.* specimens as '0'. We trained the local species level classifiers and if they predicted the '0' label, we marked the specimen as predicted only to genus level.

3.4. Classification methods

We selected our methods for the automatic classification to be CNN [21] and SVM [9] which are the most popular ones used for biological image data [19]. As our CNN model, we used the MatConvNet [36] implementation of the AlexNet CNN architecture [21]. The architecture has five convolution layers followed by three fully-connected layers. The last fully-connected layer is followed by a softmaxloss(train)/softmax(test) layer. In our tests, we considered also the output of the last fully-connected layer instead of the softmax output, because we observed that this produced better results, when the final class was decided based on the average of the outputs for each image of a specimen [30]. We trained flat and local per level classifiers from scratch using 60 training epochs. For the 17 classifiers of each local per parent node classifier, we took the flat classifier for the corresponding data split as our starting point and fine-tuned the network for 10 epochs (5 epochs only the last fully-connected layer, 3 epochs all fully-connected layers, and 2 epochs all layers). In all cases, we used a batch size of 256 and trained the network using stochastic gradient descent with a momentum of 0.9. When training from scratch, we used a learning rate varying from 0.01 to 0.0001 and for fine-tuning a learning rate varying from 0.005 to 0.0001. We saved the networks after each epoch and selected the final model based on the classification accuracy on the validation set.

While CNNs use the original images as input, we extracted a set of 64 simple geometry and intensity-based features from the images using ImageJ [32] for SVMs. The geometric features extracted include, e.g., area, perimeter, width and height of a bounding rectangle, while the intensity-based features were extracted from gray, red, green, and blue scale channels of the images. The complete set of features is listed in detail in Table .4. As these features are simple and the classification task of identifying such a large number of classes is a complex one, we found that making a principal component transformation on the features improves classification results. Therefore, we performed a principal

component transformation, as well as standardization, on the features before using them for classification.

We built our SVM model [8] using R [29] package e1071 [27] and used a Gaussian kernel. For flat classification and local per level classification, we performed a grid search for the parameters over $c = \{2^8, 2^9, 2^{10}, 2^{11}\}$ and $\gamma = \{2^{-11}, 2^{-10}, 2^{-9}, 2^{-8}\}$. For the local per parent node hierarchical classification system, we explored a larger grid as the classification problems can be very different from another at different nodes of the hierarchical system. Due to the amount of data and time consumed by evaluating just a single parameter combination, we did the following: we randomly selected one image per specimen and used this data to perform the grid search for the parameters over $c = \{2^1, 2^2, \ldots, 2^{15}\}$ and $\gamma = \{2^{-15}, 2^{-14}, \ldots, 2^{-1}\}$. After determining the optimal parameter values with this smaller data, we did a small, 3×3, grid search around those values with all the images (max. 50 images per specimen).

For both, the comparison and the machine learning data, we did the following: With each data split, we used the training data to train the model and the validation data to either select the best epoch to stop training (CNNs) or select optimal parameter values (SVMs) based on the classification accuracy of the validation specimens. With SVMs, we combined the training and validation data to train the final model after fixing the parameters. At the end, we classified each test image and selected the final class for each specimen using either average output (CNNs) or majority vote over all the images of the specimen (CNNs, SVMs).

4. Analysis and inference

4.1. Comparison data

Classification results for the comparison test sets of the image data as well as results of the proficiency test on physical data are presented in Table 1. The first row of results shows the average CE on the deepest available rank of taxonomy. These are the results traditionally examined with flat classifiers. Taxonomic experts using physical data and microscopes to identify the taxa still outperform the automatic approaches. This result by taxonomic experts can be considered as a gold-standard to compare to. However, taxonomic experts predicting taxa from the images make the most classification errors. This is understandable as the image quality can be sub-par for some specimens and the experts have not studied identification from these types of images. For the automatic classifiers, CNN using the flat classification approach and the average output for deciding the final class has the lowest CE and is in the range of taxonomic experts with physical data. The average output clearly outperforms the majority vote as a decision rule for the final class even though the number of images per specimen is relatively high.

While flat classification gives only a single level and single label predictions, it is still possible to make comparisons on different ranks of taxonomic resolution. We simply take the predictions from the deepest rank of taxonomy of the data

		CNN flat, aver.	**CNN** flat, vote	**CNN** local/ level	**CNN** hier.	**SVM** flat	**SVM** local/ level	**SVM** hier.	**Experts** images	**Experts** physical data
Deepest level	$\overline{CE}$	0.114	0.131		0.138	0.243		0.28	0.553	0.061
	$sd(CE)$	0.036	0.054		0.055	0.081		0.074	0.153	0.053
	$\overline{\text{LCSE}}$	0.052	0.070		0.070	0.173		0.191	0.353	0.028
	sd(LCSE)	0.023	0.034		0.036	0.061		0.053	0.162	0.024
Order	$\overline{CE}$	0.004	0.018	0.011	0.011	0.085	0.075	0.075	0.210	0.007
	$sd(CE)$	0.009	0.02	0.012	0.012	0.041	0.026	0.026	0.190	0.015
Family	$\overline{CE}$	0.039	0.059	0.150	0.059	0.173	0.181	0.193	0.291	0.020
	$sd(CE)$	0.029	0.037	0.259	0.044	0.070	0.062	0.069	0.151	0.020
Error structure	#ERR(order)	2	8		5	39		34	29	3
	#ERR(family)	16	19		22	40		54	11	6
	#ERR(genus)	12	12		12	16		22	15	6
	#ERR(species)	22	21		24	16		18	21	13

Table 1: Classification results for comparison test data. CE and LCSE are averaged over all 10 experts/data splits (for experts with images, 3 data splits). The number of new classification errors at each taxonomic rank is summed over all 10 data splits, where $n_{total} = 457$ (for experts with images, 3 data splits, $n_{total} = 137$).

and add the ascending taxa labels accordingly. Let us call this a bottom-up examination. Using the bottom-up examination, we can calculate LCSE also for flat classifiers. The LCSE values for all classifiers as well as for taxonomic experts are clearly smaller than the CE values (see Table 1). This means that most of the classification errors occur on deeper ranks of taxonomic resolution while the order and family might be predicted correctly. If all the classification errors were done already on the order level, CE and LCSE would be the same. For taxonomic experts using physical data, LCSE is close to zero as expected since taxonomic experts use a top-down hierarchical logic for the classification task, and identifying the higher ranks of taxonomy should be an easy task for an expert. Also in terms of LCSE, CNNs get close to the taxonomic expert level.

Contrary to the previous findings in hierarchical classification literature [34], the flat classifiers for both CNN and SVM produce better results than the hierarchical classification approach. Babbar et al. [5] stated in their study that if the data is highly unbalanced, hierarchical classifiers are better options even though their empirical error (CE) may be higher due to error propagation. While our test data is balanced, the training data used to train the classifiers is not. However, taking the hierarchical nature of the data into account when building the classifier produces not only a higher CE but also a little higher LCSE. It is worth noting that the optimization of the classifiers is based on CE, not LCSE. The only improvement the hierarchical classification system offers is a slightly lower CE on the order level for SVM. Note that for the order level, the hierarchical classifier and the local per level classifier are the same. Interestingly, the local per level SVM and CNN classifiers for family level perform worse than the flat classifiers with the ascending taxa labels. The notably high CE for local per level CNN for family level is due to data split three, where CNN classifies

all observations to the family *Elmidae*. When leaving this data split out, the average classification error is 7 %.

The bottom part of Table 1 shows the error structure for each classifier and the taxonomic experts. The number of new errors at the different taxonomic ranks sum up to the total amount of misclassifications for the 10 balanced test splits. The difference in taxonomic expert and machine logic is evident through the number of errors on each taxonomic rank. For taxonomic experts using physical data, there are very few misclassifications at the order level and the number of errors increases with the taxonomic resolution. For experts using image data, all the order level errors are due to completely missing predictions for images being too challenging to identify. That is, all the predictions made by the experts were correct at the order level and as with physical data, the number of errors increases as with the taxonomic rank. For the automatic classifiers, most misclassifications are made at either species or family level. There is no such clear hierarchy in the error structure as for the taxonomic experts.

In biomonitoring and ecosystem assessment, not only a low number of classification errors is essential, but also the type of errors made as some misclassifications can have higher cost than others. To examine this, we analysed the confusion matrices of the classifiers and taxonomic experts. Concerning especially demanding taxa, both the taxonomists and automatic classifiers had difficulties identifying *Hydropsyche saxonica*. Human experts easily misclassified them as *Hydropsyche angustipennis* when using physical data and into a mix of other *Hydropsyche* species when using image data. The image data has no *Hydropsyche angustipennis* specimens and the automatic classifiers predicted many of the *Hydropsyche saxonica* to be *Hydropsyche pellucidula* (see Fig. 6). *Hydropsyche saxonica* is also one of the least represented taxa in the image data with only 17 specimens (see Table .3) which is likely to be the reason the automatic classifiers have trouble classifying them. Besides this taxa, the human experts had another challenging taxa in the physical data. Some *Rhyacophila nubila* were misclassified as *Rhyacophila fasciata*. With the more difficult image data, the taxonomic experts classified these individuals to genus level only or left them unidentified, while SVMs mixed them with other taxa as there were no *Rhyacophila fasciata* in the image data. In addition, with the image data, the human experts had trouble identifying the *Coleopteran Elmis aenea* with some of them unidentified completely and some of them misclassified as the *Coleopteran Oulimnius tuberculatus*. The automatic classifiers identified this taxon more easily.

4.2. Machine learning data

The results on the machine learning data with larger test sets are shown in Table 2. Both CE and LCSE for all the classifiers are clearly lower with these data splits. That is due to two factors: these results are more stable, meaning they are not affected by individual difficult specimens, and here the size of each taxa in the test set reflects the size of the taxa in the training/validation sets. The comparison test sets of Section 4.1 had only 0–4 specimens of each taxa and therefore the taxa with only few training specimens had the same weight as

the taxa with hundreds of training specimens. For the machine learning data, taxa with little training data will also have only few test specimens and a small weight on the classification error of the entire test set.

		CNN flat, aver.	**CNN** flat, vote	**CNN** local/ level	**CNN** hier.	**SVM** flat	**SVM** local/ level	**SVM** hier.
Deepest level	$\overline{CE}$	0.078	0.087		0.087	0.17		0.181
	$sd(CE)$	0.009	0.009		0.013	0.008		0.009
	$\overline{\text{LCSE}}$	0.044	0.052		0.048	0.124		0.129
	sd(LCSE)	0.006	0.006		0.005	0.006		0.008
Order	$\overline{CE}$	0.01	0.015	0.011	0.011	0.055	0.053	0.053
	$sd(CE)$	0.002	0.003	0.002	0.002	0.006	0.005	0.005
Family	$\overline{CE}$	0.041	0.05	0.033	0.044	0.129	0.126	0.135
	$sd(CE)$	0.006	0.007	0.003	0.004	0.006	0.008	0.011
Error structure	#ERR(order)	194	287		216	1071		1017
	#ERR(family)	605	685		638	1428		1589
	#ERR(genus)	304	307		319	455		505
	#ERR(species)	410	412		510	344		393

Table 2: Classification results for machine learning test data. CE and LCSE are averaged over all 10 data splits, where each test split has $n = 1937$. The number of new classification errors at each taxonomic rank is summed over all 10 data splits, where $n_{total} = 19370$.

The results are similar to those in Table 1. CNNs produce the best classification results. Again, the flat classification versions of CNN and SVM outperform the hierarchical classifiers contradicting previous findings of hierarchical classification studies [34]. With the machine learning data splits, the local per level classification approach gives slightly lower CE than the flat classifier on both order level (SVM) and family level (SVM and CNN).

When considering individual challenging taxa, the best classifier, CNN, has mostly trouble with the least represented taxa in the data due to lack of adequate training data. The smallest taxa are *Hydropsyche saxonica*, *Nemoura cinerea*, *Capnosis schilleri*, *Sialis sp.*, *Leuctra nigra* and *Sphaerium sp.* with average number of specimens in the training data, $N = \{13, 11, 15, 19, 19, 107\}$ and #images $= \{349, 540, 730, 856, 960, 1239\}$ respectively. With the exceptions of *Sialis sp.* and *Sphaerium sp.*, the average CE for these taxa ranged from 62% to 98% for CNNs and from 61% to 100% for SVMs. On the contrary, all the classifiers performed well on classifying *Sphaerium sp.* ($\overline{CE} \in [0, 8\%]$), and CNNs also relatively well on classifying *Sialis sp.* ($\overline{CE} \in [15, 18\%]$).

One reason why the hierarchical, local per parent node approach performs worse than flat classification could be that the hierarchy in the data is not based on visual aspects. The taxonomic resolution is based on affinity which can be independent of the appearance of the taxa. However, the automatic classifiers base all classification decisions on visual features hence the man-made

Figure 6: Examples of visual differences among taxa belonging to the same family or genus. Top row: *Hydropsyche pellucidula*, *Hydropsyche saxonica*, and *Hydropsyche siltalai* all belong to the genus *Hydropsyche sp.* Bottom row: *Neureclipsis bimaculata*, *Plectronemia*, *Polycentropus flavomaculatus*, and *Polycentropus irroratus* all belong to the family *Polycentropodidae*. In both cases, the taxa are of different sizes and colors.

hierarchy of the data could confuse the classifiers. Fig. 6 gives examples of taxa that belong to the same family or genus but have clear differences in their appearance, e.g., size.

5. Discussion

The status assessment of ecosystems is often based on the use of biological indicators that are manually identified by human experts. The manual collection and identification of the data by ecological experts is, however, known to be costly and time consuming. While recently a growing number of studies explore the enormous potential of genetic identification methods, these are currently not standardized, and thus currently cannot be used to their full potential for legislative biomonitoring purposes [e.g. 14]. An interim solution could lie in the use of a computer-based identification system that could be used to simply replace the step of human identification in current biomonitoring while preserving all other steps of the existing process chain. To switch to this novel approach, ecologists must start to put trust in the machine logic. In this work, we compared human expert predictions for physical and image data to those of machine learning methods on image data.

To automate the identification process, we have developed a generic imaging system producing multiple images for each specimen. With our imaging system, we collected a large dataset of benthic freshwater macroinvertebrate images and

assigned labels consisting of multiple taxonomic ranks. The classical approach in the computer-based identification has been a flat classification, where the classification is performed at the most specific rank of the taxonomic resolution. In addition to the classical flat approach, we considered also local hierarchical classifiers, namely local per level classifiers and local per parent node classifiers. We selected convolutional neural networks (CNNs) and support vector machines (SVMs) as classification methods. We are not aware of any earlier works applying the local hierarchical classifiers based on the taxonomic resolution of invertebrates. We evaluated both automatic classifiers and taxonomic experts using the classification error (CE) at the most specific level and a novel variant of the context sensitivity error (CSE) taking the top-down success into account. We call this variant level-aware context-sensitive error (LCSE).

We split the image data to produce test sets similar to the ones used in the proficiency test with physical data for taxonomic experts to be able to directly compare machines and human experts. We found that the taxonomic experts obtained the best classification performance when analysing the physical data using a microscope ($\overline{CE} = 6.1\%$ and $\overline{LCSE} = 2.8\%$) and the worst when using the image data ($\overline{CE} = 55.3\%$ and $\overline{LCSE} = 35.3\%$). The best automatic classifier was the CNN using flat classification approach and the average output of all the images for a specimen as the decision rule to decide the final label ($\overline{CE} = 11.4\%$ and $\overline{LCSE} = 5.3\%$). This result is well within the range of human experts taking part in the proficiency test. We observed also that, contrary to earlier observations in the literature, the flat classifiers with both CNN and SVM performed better than the local per parent node hierarchical classifiers. We assume this is because the hierarchy based on the taxonomic resolution does not necessarily correlate with the visual similarity of the taxa. The hierarchical classifiers would be likely more successful if they could first separate the easiest superclasses and then concentrate on more subtle differences within those superclasses. Besides the CE and LCSE measures, we also investigated the main differences in confusion matrices. The most difficult classes were partially overlapping for machines and experts, but there were some differences as well. Human experts using images preferred to stay at higher ranks of taxonomic hierarchy for difficult taxa while machines were forced to predict the deepest possible level, and thus, ended up predicting wrong species. Unsurprisingly, we observed that CNNs had trouble identifying the classes with a low amount of training samples.

The test sets in our comparison data were very small to not burden the human participants too much. This naturally makes the results unstable in the sense that few difficult specimens or bad images may affect the results a lot. Therefore, we evaluated the automatic classifiers also on different data splits, where the test sets were considerably larger and also represented the overall taxa distribution. The ranking of the automatic classifiers with respect to the CE and LCSE measures was similar, while the absolute CE and LCSE values were much smaller for these larger test sets. Again, forcing automatic classifiers to operate with the logic of human experts, i.e., local per parent node approach, did not improve classification results.

6. Conclusion

The main purpose of this paper was to investigate differences in the identification logic of humans and machines. When compared to the existing literature, up to our knowledge this was the first attempt to use a human-like hierarchical classifier for macroinvertebrate image data. With respect to accuracy of identification human taxonomic experts still outperformed the selected automatic methods on the limited set of taxa and specimens used in proficiency tests, but CNNs' performance was close and fell within the range of typical human experts. With respect to speed, human identification is no match to that of machines' as a taxonomic expert uses 2 seconds to several minutes to identify a specimen while a machine spends milliseconds on a single specimen and will be faster still with improved algorithms and increases in computing power. In addition, computers can run during the night and weekends while human experts have limited working hours and also other tasks at work. In future studies, we will apply more advanced machine learning techniques, further boost the identification performance on the most rare classes using, e.g., transfer learning and data augmentation, and consider global hierarchical classifiers.

It is important that ecologists understand and leverage the potential that the high speed and overall good accuracy of automated identification can have on assessments. If applied, these methods will significantly reduce human workload and perform routine identification tasks to a sufficiently accurate degree. Given our results and the fast pace in the field of image recognition, we expect that automatic identification methods can replace human experts in the routine identification of bulk taxa soon, while human experts and genetic methods will still be needed to concentrate on the harder to identify cases. We hope that our results convince doubting ecologists to trust that machine logic can indeed be used to take over a task traditionally done by humans while also increasing their understanding of the main challenges still associated with automatic identification.

Acknowledgements

We thank the Academy of Finland for the grants of Ärje (284513, 289076), Tirronen (289076, 289104) Kärkkäinen (289076), Meissner (289104), and Raitoharju (288584). We would like to thank CSC for computational resources.

References

[1] Ärje, J., Choi, K.-P., Divino, F., Meissner, K., and Kärkkäinen, S. (2016). Understanding the statistical properties of the percent model affinity index can improve biomonitoring related decision making. *Stochastic Environmental Research and Risk Assessment*, 30(7):1981–2008.

[2] Ärje, J., Kärkkäinen, S., Meissner, K., Iosifidis, A., Ince, T., Gabbouj, M., and Kiranyaz, S. (2017). The effect of automated taxa identification errors on biological indices. *Expert Systems with Applications*, 72:108–120.

[3] Ärje, J., Kärkkäinen, S., Meissner, K., and Turpeinen, T. (2010). Statistical classification methods and proportion estimation – an application to a macroinvertebrate image database. Proceedings of the 2010 IEEE Workshop on Machine Learning for Signal Processing (MLSP).

[4] Ärje, J., Kärkkäinen, S., Turpeinen, T., and Meissner, K. (2013). Breaking the curse of dimensionality in quadratic discriminant analysis models with a novel variant of a bayes classifier enhances automated taxa identification of freshwater macroinvertebrates. *Environmetrics*, 24(4):248–259.

[5] Babbar, R., Partalas, I., Gaussier, E., Amini, M.-R., and Amblard, C. (2016). Learning taxonomy adaptation in large scale classification. *Journal of Machine Learning Research*, 17:1–37.

[6] Borja, A. and Elliott, M. (2013). Marine monitoring during an economic crisis: the cure is worse than the disease. *Marine Pollution Bulletin*, 68:1–3.

[7] Caley, M. J., O'Leary, R. A., Fisher, R., Low-Choy, S., Johnson, S., and Mengersen, K. (2014). What is an expert? A systems perspective on expertise. *Ecology and Evolution*, 4(3):231–242.

[8] Chang, C.-C. and Lin, C.-J. (2011). LIBSVM: A library for support vector machines. *ACM Transactions on Intelligent Systems and Technology*, 2(3):27:1–27.

[9] Cortes, C. and Vapnik, V. (1995). Support-vector networks. *Machine Learning*, 20:273–297.

[10] Culverhouse, P., Williams, R., Reguera, B., Herry, V., and González-Gil, S. (2003). Do experts make mistakes? A comparison of human and machine identification of dinoflagellates. *Marine Ecology Progress Series*, 247:17–25.

[11] Culverhouse, P., Williams, R., Reguera, B., Herry, V., and González-Gil, S. (2006). Automatic image analysis of plankton: future perspectives. *Marine Ecology Progress Series*, 312.

[12] Elbrecht, V., Vamos, E. E., Meissner, K., Aroviita, J., and Leese, F. (2017). Assessing strengths and weaknesses of DNA metabarcoding-based macroinvertebrate identification for routine stream monitoring. *Methods in Ecology and Evolution*, 8(10):1265–1275.

[13] Haase, P., Pauls, S. U., Schindehütte, K., and Sunderman, A. (2010). First audit of macroinvertebrate samples from an EU Water Framework Directive monitoring program: human error greatly lowers precision of assessment results. *Journal of the North American Benthological Society*, 29(4):1279–1291.

[14] Hering, D., Borja, A., Jones, J. I., Pont, D., Boets, P., Bouchez, A., Bruce, K., Drakare, S., Hänfling, B., Kahlert, M., Leese, F., Meissner, K., Mergen, P., Reyjol, Y., Segurado, P., Vogler, A., and Kelly, M. (2018). Implementation options for DNA-based identification into ecological status assessment under the European Water Framework Directive. *Water Research*, 138:192–205.

[15] International commission on zoological nomenclature (1999). *International Code of Zoological Nomenclature*. International Trust for Zoological Nomenclature, fourth edition.

[16] Järvinen, M., Aroviita, J., Hellsten, S., Karjalainen, S. M., Kuoppala, M., Meissner, K., Mykrä, H., and Vuori, K.-M. (2019). *Jokien ja järvien biologinen seuranta - näytteenotosta tiedon tallentamiseen. Online guidance.*

[17] Joutsijoki, H., Meissner, K., Gabbouj, M., Kiranyaz, S., Raitoharju, J., Ärje, J., Kärkkäinen, S., Tirronen, V., Turpeinen, T., and Juhola, M. (2014). Evaluating the performance of artificial neural networks for the classification of freshwater benthic macroinvertebrates. *Ecological Informatics*, 20:1–12.

[18] Kelly, M., Schneider, S., and King, L. (2015). Customs, habits, and traditions: the role of nonscientific factors in the development of ecological assessment methods. *WIREs Water*, 2:159–165.

[19] Kho, S. J., Manickam, S., Malek, S., Mosleh, M., and Dhillon, S. K. (2017). Automated plant identification using artificial neural network and support vector machine. *Frontiers in Life Science*, 10(1):98–107.

[20] Kiranyaz, S., Ince, T., Pulkkinen, J., Gabbouj, M., Ärje, J., Kärkkäinen, S., Tirronen, V., Juhola, M., Turpeinen, T., and Meissner, K. (2011). Classification and retrieval on macroinvertebrate image databases. *Computers in Biology and Medicine*, 41(7):463–472.

[21] Krizhevsky, A., Sutskever, I., and Hinton, G. E. (2012). Imagenet classification with deep convolutional neural networks. In *Advances in Neural Information Processing Systems (NIPS)*, pages 1097–1105.

[22] Lee, H., Park, M., and Kim, J. (2016). Plankton classification on imbalanced large scale database via convolutional neural networks with transfer learning. *Image Processing (ICIP), 2016 IEEE International Conference on*, pages 3713–3717.

[23] Leese, F., Bouchez, A., Abarenkov, K., Altermatt, F., Borja, Á.., Bruce, K., Ekrema, T., Čiamporová-Zaťovičová, F., Costa, F. O., Duarte, S., Elbrecht, V., Fontaneto, D., Franc, A., Geiger, M. F., Hering, D., Kahlert, M., Kalamujić Stroil, B., Kelly, M., Keskin, E., Liska, I., Mergen, P., Meissner, K., Pawlowski, J., Penev, L., Reyjol, Y., Rotter, A., Steinke, D., van der Wal, B., Vitecek, S., Zimmermann, J., and Weigand, A. M. (2018). Why we need sustainable networks bridging countries, disciplines, cultures and generations for aquatic biomonitoring 2.0: a perspective derived from the DNAqua-Net COST Action. *Advances in Ecological Research*, 58:63–99.

[24] Levatic, J., Kocev, D., and Dzeroski, S. (2015). The importance of the label hierarchy in hierarchical multi-label classification. *Journal of Intelligent Information Systems*, 45(2):247–271.

[25] Lytle, D. A., Martínez-Muñoz, G., Zhang, W., Larios, N., Shapiro, L., Paasch, R., Moldenke, A., Mortensen, E. N., Todorovic, S., and Dietterich, T. G. (2010). Automated processing and identification of benthic invertebrate samples. *Journal of the North American Benthological Society*, 29(3):867–874.

[26] Meissner, K., Nygård, H., Björklöf, K., Jaale, M., Hasari, M., Laitila, L., Rissanen, J., and Leivuori, M. (2017). Proficiency test 04/2016: Taxonomic identification of boreal freshwater lotic, lentic, profundal and North-Eastern Baltic benthic macroinvertebrates. *Reports of the Finnish Environment Institute*, 2.

[27] Meyer, D., Dimitriadou, E., Hornik, K., Weingessel, A., and Leisch, F. (2018). *e1071: Misc Functions of the Department of Statistics, Probability Theory Group (Formerly: E1071), TU Wien*. R package version 1.7-1.

[28] Nygård, H., Oinonen, S., Lehtiniemi, M., Hällfors, H., Rantajärvi, E., and Uusitalo, L. (2016). Price versus value of marine monitoring. *Fronties in Marine Science*, 3:205.

[29] R Core Team (2016). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.

[30] Raitoharju, J. and Meissner, K. (2019). On confidences and their use in (semi-)automatic multi-image taxa identification. *2019 IEEE Symposium Series on Computational Intelligence (SSCI), Xiamen, China*, pages 1338–1343.

[31] Raitoharju, J., Riabchenko, E., Ahmad, I., Iosifidis, A., Gabbouj, M., Kiranyaz, S., Tirronen, V., Ärje, J., Kärkkäinen, S., and Meissner, K. (2018). Benchmark database for fine-grained image classification of benthic macroinvertebrates. *Image and Vision Computing*, 78:73–83.

[32] Rasband, W. S. (1997-2010). *ImageJ*. U.S. National Institutes of Health, Bethesda, Maryland, USA.

[33] Rodrigues, F., Pereira, F. C., Alves, A., Jiang, S., and Ferreira, J. (2012). Automatic classification of points-of-interest for land-use analysis. *Proceedings of GEOProcessing 2012: The Fourth International Conference on Advanced Geographic Information Systems, Applications, and Services*, pages 41–49.

[34] Silla, C. N. J. and Freitas, A. A. (2011). A survey of hierarchical classification across different application domains. *Data Mining and Knowledge Discovery*, 22(1–2):31–72.

[35] Uusitalo, L., Fernandes, J. A., Bachiller, E., Tasala, S., and Lehtiniemi, M. (2016). Semi-automated classification method addressing marine strategy framework directive (msfd) zooplankton indicators. *Ecological Indicators*, 71:398–405.

[36] Vedaldi, A. and Lenc, K. (2015). MatConvNet: Convolutional neural networks for Matlab. In *Proceedings of International Conference on Multimedia*, pages 689–692.

[37] Verma, N., Mahajan, D., Sellamanickam, S., and Nair, V. (2012). Learning hierarchical similarity metrics. *2012 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2280–2287. Providence, RI USA.

[38] Yousef Kalafi, E., Town, C., and Kaur Dhillon, S. (2018). How automated image analysis techniques help scientists in species identification and classification. *Folia Morphologica*, 77(2):179–193.

[39] Zimmermann, J., Glockner, G., Jahn, R., Enke, N., and Gemeinholzer, B. (2015). Meta-barcoding vs. morpological identification to assess diatom diversity in environmental studies. *Molecular Ecology Resources*, 15:526–542.

Appendix

Taxa	Species	Genus	Family	Order	#specimens	#images
Elmis aenea	*Elmis aenea*	*Elmis*	*Elmidae*	*Coleoptera*	648	32398
Limnius volckmari	*Limnius volckmari*	*Limnius*	*Elmidae*	*Coleoptera*	314	15621
Oulimnius tuberculatus	*Oulimnius tuberculatus*	*Oulimnius*	*Elmidae*	*Coleoptera*	335	16674
Hydraena sp.	-	*Hydraena*	*Hydraenidae*	*Coleoptera*	198	9900
Simuliidae	-	-	*Simuliidae*	*Diptera*	887	44240
Ameletus inopinatus	*Ameletus inopinatus*	*Ameletus*	*Ameletidae*	*Ephemeroptera*	127	6346
Baetis rhodani	*Baetis rhodani*	*Baetis*	*Baetidae*	*Ephemeroptera*	404	19829
Baetis vernus group	*Baetis vernus*	*Baetis*	*Baetidae*	*Ephemeroptera*	176	8588
Ephemerella aurivillii	*Ephemerella aurivillii*	*Ephemerella*	*Ephemerellidae*	*Ephemeroptera*	356	16458
Ephemerella mucronata	*Ephemerella mucronata*	*Ephemerella*	*Ephemerellidae*	*Ephemeroptera*	304	15175
Heptagenia sulphurea	*Heptagenia sulphurea*	*Heptagenia*	*Heptageniidae*	*Ephemeroptera*	438	21502
Kageronia fuscogrisea	*Kageronia fuscogrisea*	*Kageronia*	*Heptageniidae*	*Ephemeroptera*	222	10826
Leptophlebia sp.	-	*Leptophlebia*	*Leptophlebiidae*	*Ephemeroptera*	412	20366
Sialis sp.	-	*Sialis*	*Sialidae*	*Megaloptera*	26	1162
Capnopsis schilleri	*Capnopsis schilleri*	*Capnopsis*	*Capniidae*	*Plecoptera*	21	1050
Leuctra nigra	*Leuctra nigra*	*Leuctra*	*Leuctridae*	*Plecoptera*	27	1350
Leuctra sp.	-	*Leuctra*	*Leuctridae*	*Plecoptera*	298	14899
Amphinemura borealis	*Amphinemura borealis*	*Amphinemura*	*Nemouridae*	*Plecoptera*	322	16100
Nemoura cinerea	*Nemoura cinerea*	*Nemoura*	*Nemouridae*	*Plecoptera*	16	800
Nemoura sp.	-	*Nemoura*	*Nemouridae*	*Plecoptera*	187	9314
Protonemura sp.	-	*Protonemura*	*Nemouridae*	*Plecoptera*	100	4908
Diura sp.	-	*Diura*	*Perlodiae*	*Plecoptera*	98	4427
Isoperla sp.	-	*Isoperla*	*Perlodiae*	*Plecoptera*	243	12148
Taeniopteryx nebulosa	*Taeniopteryx nebulosa*	*Taeniopteryx*	*Taenioptegyridae*	*Plecoptera*	331	16325
Micrasema gelidum	*Micrasema gelidum*	*Micrasema*	*Brachycentridae*	*Trichoptera*	233	11528
Micrasema setiferum	*Micrasema setiferum*	*Micrasema*	*Brachycentridae*	*Trichoptera*	323	13819
Agapetus sp.	-	*Agapetus*	*Glossosomatidae*	*Trichoptera*	290	14387
Silo pallipes	*Silo pallipes*	*Silo*	*Goeridae*	*Trichoptera*	56	2658
Hydropsyche pellucidula	*Hydropsyche pellucidula*	*Hydropsyche*	*Hydropsychidae*	*Trichoptera*	192	6513
Hydropsyche saxonica	*Hydropsyche saxonica*	*Hydropsyche*	*Hydropsychidae*	*Trichoptera*	17	490
Hydropsyche siltalai	*Hydropsyche siltalai*	*Hydropsyche*	*Hydropsychidae*	*Trichoptera*	395	19456
Oxyethira sp.	-	*Oxyethira*	*Hydroptilidae*	*Trichoptera*	218	10381
Lepidostoma hirtum	*Lepidostoma hirtum*	*Lepidostoma*	*Lepidostomatidae*	*Trichoptera*	267	10982
Neureclipsis bimaculata	*Neureclipsis bimaculata*	*Neureclipsis*	*Polycentropodidae*	*Trichoptera*	477	23721
Plectrocnemia sp.	-	*Plectrocnemia*	*Polycentropodidae*	*Trichoptera*	63	3015
Polycentropus flavomaculatus	*Polycentropus flavomaculatus*	*Polycentropus*	*Polycentropodidae*	*Trichoptera*	224	11005
Polycentropus irroratus	*Polycentropus irroratus*	*Polycentropus*	*Polycentropodidae*	*Trichoptera*	59	2917
Rhyacophila nubila	*Rhyacophila nubila*	*Rhycophila*	*Rhyacophilidae*	*Trichoptera*	177	6993
Sphaerium sp.	-	*Sphaerium*	*Sphaeridae*	*Veneroida*	150	1733

Table .3: Taxonomic resolution of the multiple image data and the numbers of specimens and images per taxa. Taxa included in the proficiency test for human experts but not included in the image data were *Brachyptera risi*, *Cloeon sp.*, *Cloeon diptera group*, *Cloeon inscriptum*, *Cloeon simile*, *Helobdella stagnalis*, and *Tinodes waeneri*.

Geometric features	RGB and grey scale features
Area	Mean
Center of mass	Standard deviation
• X and Y coordinates	Mode
Perimeter	Minimum
Bounding rectangle	Maximum
• Width and Height	Center of mass
• X and Y coordinates of the upper left corner	• X and Y coordinates
Fit ellipse	Integrated density
• Major and Minor axis	Median
• Angle	Skewness
• X and Y coordinates of the center	Kurtosis
Circularity	
Aspect ratio	
Roundness	
Solidity	
Feret's diameter	
• Length	
• Angle	
• Minimum caliper length	
• X and Y starting coordinates	

Table .4: Features used for SVM classification.

Author statement

Johanna Ärje: conceptualization, data curation, formal analysis, methodology, visualization, original draft, review & editing

Jenni Raitoharju: conceptualization, data curation, formal analysis, methodology, visualization, original draft, review & editing

Alexandros Iosifidis: conceptualization, methodology, review & editing

Ville Tirronen: data curation, resources, software, visualization, review & editing

Kristian Meissner: conceptualization, data curation, funding acquisition, project administration, resources, supervision, original draft, review & editing

Moncef Gabbouj: conceptualization, funding acquisition, methodology, project administration, supervision, review & editing

Serkan Kiranyaz: conceptualization, funding acquisition, methodology, project administration, review & editing

Salme Kärkkäinen: conceptualization, funding acquisition, methodology, project administration, supervision, original draft, review & editing

AUTHOR DECLARATION TEMPLATE

We wish to confirm that there are no known conflicts of interest associated with this publication and there has been no significant financial support for this work that could have influenced its outcome.

We confirm that the manuscript has been read and approved by all named authors and that there are no other persons who satisfied the criteria for authorship but are not listed. We further confirm that the order of authors listed in the manuscript has been approved by all of us.

We confirm that we have given due consideration to the protection of intellectual property associated with this work and that there are no impediments to publication, including the timing of publication, with respect to intellectual property. In so doing we confirm that we have followed the regulations of our institutions concerning intellectual property.

We understand that the Corresponding Author is the sole contact for the Editorial process (including Editorial Manager and direct communications with the office). She is responsible for communicating with the other authors about progress, submissions of revisions and final approval of proofs. We confirm that we have provided a current, correct email address which is accessible by the Corresponding Author and which has been configured to accept email from johanna.arje@jyu.fi.

Signed by all authors as follows:

23.04.2019
Johanna Ärje

23 Apr 2019
Jenni Raitoharju

23 4 2019
Alexandros Iosifidis

Ville Tirronen

26. APRIL 2019
Kristian Meissner

Moncef Gabbouj

Serkan Kiranyaz

6th May 2019
Salme Kärkkäinen