Benchmark Database for Fine-Grained Image Classification of Benthic Macroinvertebrates

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

PII:	S0262-8856(18)30101-X
DOI:	doi:10.1016/j.imavis.2018.06.005
Reference:	IMAVIS 3698
To appear in:	*Image and Vision Computing*
Received date:	22 June 2017
Revised date:	24 April 2018
Accepted date:	21 June 2018

Please cite this article as: Jenni Raitoharju, Ekaterina Riabchenko, Iftikhar Ahmad, Alexandros Iosifidis, Moncef Gabbouj, Serkan Kiranyaz, Ville Tirronen, Johanna Ärje, Salme Kärkkäinen, Kristian Meissner , Benchmark Database for Fine-Grained Image Classification of Benthic Macroinvertebrates. Imavis (2018), doi:10.1016/j.imavis.2018.06.005

Benchmark Database for Fine-Grained Image Classification of Benthic Macroinvertebrates

Jenni Raitoharju[a], Ekaterina Riabchenko[a], Iftikhar Ahmad[a], Alexandros Iosifidis[a], Moncef Gabbouj[a], Serkan Kiranyaz[b], Ville Tirronen[c], Johanna Ärje[d], Salme Kärkkäinen[d], Kristian Meissner[e]

[a]*Laboratory of Signal Processing, Tampere University of Technology, Tampere, Finland*
[b]*Department of Electrical Engineering, Qatar University, Doha, Qatar*
[c]*Faculty of Information Technology, University of Jyvskyl, Jyvskyl, Finland*
[d]*Department of Mathematics and Statistics, University of Jyväskylä, Jyväskylä, Finland*
[e]*Freshwater Centre, Finnish Environment Institute, Jyväskylä, Finland*

Abstract

Managing the water quality of freshwaters is a crucial task worldwide. One of the most used methods to biomonitor water quality is to sample benthic macroinvertebrate communities, in particular to examine the presence and proportion of certain species. This paper presents a benchmark database for automatic visual classification methods to evaluate their ability for distinguishing visually similar categories of aquatic macroinvertebrate taxa. We make publicly available a new database, containing 64 types of freshwater macroinvertebrates, ranging in number of images per category from 7 to 577. The database is divided into three datasets, varying in number of categories (64, 29, and 9 categories). Furthermore, in order to accomplish a baseline evaluation performance, we present the classification results of Convolutional Neural Networks (CNNs) that are widely used for deep learning tasks in large databases. Besides CNNs, we experimented with several other well-known classification methods using deep features extracted from the data.

Keywords: Biomonitoring, Fine-grained Classification, Benthic Macroinvertebrates, Deep Learning, Convolutional Neural Networks

1. Introduction

All ecosystems and ultimately human societies depend on biodiversity and ecosystem functioning [18]. Freshwater ecosystems are among the most threatened ecosystems worldwide (see, e.g., [39, 13]) as the loss of aquatic biodiversity and associated ecosystem services is estimated to surpass the loss of biodiversity in rainforests (e.g., [37, 48]). The importance of monitoring aquatic ecosystems and biodiversity is acknowledged in environmental legislation, such as the EU Water Framework Directive (WFD) [1], the EU Marine Strategy Framework Directive [2], and the US Clean Water Act [3]. The EU WFD legislation requires monitoring of several biological indicator groups for freshwater ecological status assessment. Knowledge obtained from these biomonitoring programs is used to assess the status of ecosystems, preserve and assure good future water quality. Even for the species living in species poor freshwaters of northern Europe, this legal requirement brings about a need to track hundreds of macroinvertebrate and thousands of microscopic periphyton and phytoplankton taxa.

Taxonomic identification of biomonitoring samples involving microscopy is cost intensive, as the identification of indicator species is usually done by human experts. While there are also DNA-based methods of identification, they currently do not cope well with the WFD requirements for information on indicator taxa abundance. Furthermore, genetic methods are still at least as cost intensive as traditional ones, although the price of these methods is decreasing quickly [14]. A recent study showed that manual identification of freshwater macroinvertebrate taxa done by human experts is more prone to errors (i.e., 30%) than previously assumed [16], and this may extend to other microscopic indicator groups as well [11, 12]. Thus, human-made taxonomic identification errors can affect the results and reliability of ecological aquatic research and managerial decisions regarding ecosystems services and resources. Put into a management context and recalling that the number of highly trained taxonomic experts is decreasing, this suggests that large amounts of resources may be ineffectively allocated in restoration efforts [16].

In this paper, we focus on automatic methods of identification of benthic macroinvertebrates and present a benchmark database to evaluate and test automatic identification methods. Alongside aquatic macrophytes, which generally do not require microscopic identification, benthic macroinvertebrates are the most commonly used biological indicators in the WFD implementation [7]. In Finland, 280 taxa are currently used as indicators for WFD index calculations. In official Finnish aquatic monitoring, four to six samples, each containing ca. 50-1000 benthic macroinvertebrates belonging to 2-70 taxa, are taken from each

monitoring site using a kick-net. Samples are individually preserved and transported to a laboratory where the specimens are picked from the debris before microscopic inspection. A technician then takes about 1-4 hours to pick all animals from the collected material in a single sample and a highly trained expert needs another 1-4 hours to identify the sorted specimens using microscopy. In national proficiency tests human experts classify test samples with a 87-100 percent accuracy [36]. As funding for environmental protection and thus monitoring is nowadays decreasing in Europe [43], alternative ways to achieve the requirements set forth in the WFD have to be considered. Here, we propose to employ computer vision and machine learning techniques for the identification of macroinvertebrate taxa in order to provide cost-effective and more accurate solution for this crucial problem. We propose a semi-automatic taxa identification method, which significantly reduces the expert time and, thus, provides a cost-effective solution. To our best knowledge, this is the first time such a semi-automatic approach is adopted in this domain. Using the proposed imaging approach, we present a benchmark dataset containing more than 15000 images and, in order to serve as a baseline performance evaluation, our initial classification results obtained using Convolutional Neural Networks (CNNs) applied on the original images and several other classification methods applied on the deep features extracted from the images.

The rest of the paper is organized as follows. Section 2 presents related work, while Section 3 provides an extensive description of the proposed database and procedures used to produce it. Section 4 shows preliminary results obtained on the proposed database using deep CNNs and a combination of deep features and well-known classifiers. Finally, Section 5 concludes the paper and suggests topics for future research.

2. Related Work

Introduction of publicly available benchmarks has been a driving force in the development and improvement of computer vision algorithms. With time, computer vision databases transformed from having low resolution grayscale images with only one object per image, e.g., UIUC car database [4], to ones with more complex image composition (multiple objects, occlusion, and truncation) and containing multiple visual classes in the database, as in Pascal VOC [15]. The introduction of the large-scale ImageNet database [44], which includes 1000 image classes with approximately 1 million images, has stimulated the development and success of the modern state-of-the-art deep learning techniques, i.e., CNNs, that are now applied in all areas of computer vision (e.g., object classification, detection, and segmentation).

There are also several public databases available for fine-grained classification depicting, e.g., cars [26, 31], flowers [38], birds [6, 45], dogs [32, 23], aircraft [35, 47], plant leaves [28], or plankton [17]. However, even in the aforementioned databases, different categories are generally easily distinguished —even if not recognized —also by non-experts. For macroinvertebrates, this is not the case. Non-experts typically cannot detect the subtle morphological differences between the taxa and, thus, the classification problem is even more fine-grained than in the datasets typically used for fine-grained classification. The macroinvertebrate identification is further complicated by the fact that often specimens from two different taxa can appear more similar than two specimens from a single taxon.

The area of automatic identification of freshwater macroinvertebrates is relatively unexplored. Most of the previous works have been using only small datasets containing only 8-9 image categories [22, 25, 24, 33]. Recent works have presented results for more than 30 categories of macroinvertebrates (35 in [5], 50 in [21], and 54 in [30]). However, the datasets used in these experiments are not publicly available.

The goal of this work is to introduce and publish a new and significantly larger benchmark database for automatic fine-grained classification of benthic macroinvertebrates containing 64 categories and more than 15000 images. We have used a part of this database (Dataset 2) for experiments in two earlier papers [42, 41]. In [42], we compared the classification performance obtained using engineered features (e.g. SIFT and HOG) and features learned by a CNN. Our experiments showed that the CNN features clearly outperform the engineered ones. In [41], we applied simple data augmentation to improve the classification results. In this work, we focus on the database itself and provide baseline results using CNN features with different classifiers.

3. Database Creation and Description

3.1. Sample Preparation and Manual Classification

When creating the database, we prepared the samples of benthic macroinvertebrates and manually classified them following the typical steps used in manual identification. The samples were collected from Finnish rivers and transported to the laboratory for further processing. First, larger debris was manually removed from the samples. Then each sample was placed onto a sieve of equal or smaller mesh size than used in the kick-net and rinsed clean of smaller particles under water. After careful rinsing, the sample material was spread evenly onto a tray and benthic macroinvertebrates were picked by a technician using a strong light. Next, specimens were placed into vials filled with preservative for later inspection under a microscope. Prior to microscopic identification, some of the specimens were presorted into smaller groups based on morphotypes to speed up the process of manual identification. The process of cleaning a sample and its manual classification is illustrated in Figure 1.

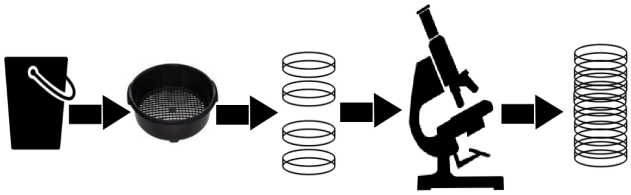


Figure 1: A simplified scheme of the cleaning and classification processes

Depending on the difficulty of classification (i.e., taxa identification), specimens were classified with different taxonomic resolution. In Finland, identification at the species level is commonly performed, but for some taxa only the coarser family or genus level (marked as *"sp."* in Table 1) can be established with certainty. For some taxa, specimens at different developmental stages of the same species have to be split into separate groups during identification due to absence of common visual features between those stages, see Figure 2: larval (marked as "larva" in Table 1) and adult (marked as "adult"). Examples of coleopteran (represented by two developmental stages) taxa in the proposed database are *Elmis aenea* and *Oulimnius tuberculatus*.

Figure 2: Difference of the larval (left) and adult (right) stages of development of *Elmis aenea*

3.2. Proposed Imaging Setup

Taking kick-net samples and extracting macroinvertebrates from them is a very laborious task and, at the moment, there are no cost-effective alternatives to using manual labor. However, a semi- or fully automatized imaging stage, facilitating automated identification, can easily be assembled nowadays. In this subsection, we present a semi-automatic imaging setup developed for aquatic macroinvertebrates, which allows imaging of approximately 1000 specimens in only 3 hours.

The imaging setup is presented in Figure 3. It consists of two Guppy PRO F-125B/C cameras (frame rate of 30 fps) with Megapixel Macro Lens (f=75mm, F:3.5-CWD<535mm) viewing the test tube from perpendicular angles and a LED light of 1040 lumen. A rotating mechanism, allowing the change of the test-tube without interruption of the imaging session, is shown in the top left corner of the image. During the imaging process, the software builds a model of a background and controls/sets off shutters of the cameras when a significant change is detected, for example, due to the appearance of a specimen into the field of view of the cameras. This process is controlled by computer with Intel Core 2 Duo processor (at 2.66 GHz) and 4GB of RAM. The material cost of this imaging system is approximately 4-5K €, which is much less than an average cost of a high quality stereo microscopes traditionally used for fine-grained classification of macroinvertebrates [33].

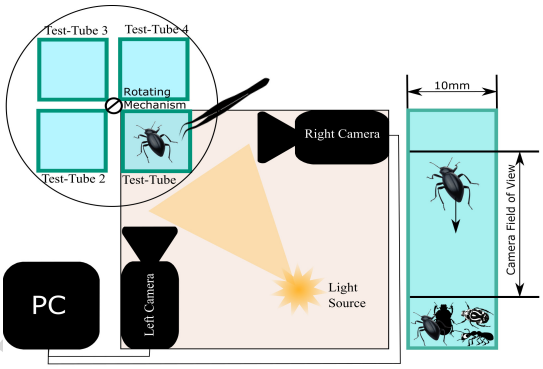


Figure 3: Schematic of the imaging setup

During imaging, the operator takes the samples stored in separate vials and processes them one-by-one. (When creating the database, the samples were previously classified and each vial contained species of the same category.) Using tweezers, the operator drops each specimen into a high-grade glass cuvette (test tube in Figure 3) filled with alcohol, where it is photographed from two viewpoints by cameras positioned at a 90 degree angle, while sinking to the bottom of the cuvette. When the number of already processed specimens grows on the bottom of the cuvette and they start appearing in the field of view of the cameras, the operator changes the cuvette. If a specimen is too big or too heavy (sinks too fast), the resulting frames are stitched in order to obtain a full view of an object in one image. While this process sometimes causes blur, thus degrading quality of the final image, image stitching is required for the following processing.

3.3. Image Preprocessing

After imaging using the described imaging setup, some images were completely discarded due to their poor quality. The remaining images were in PNG format, their size varied from 640x480 to 1280x960 pixels and they contained a varying amount of background. In order to focus on learning object representations rather than variations in the background, we cropped the objects (macroinvertebrate specimens) from the original images. We used Otsu's threshold [40] for green and red channels, filtering, and analysis of connected components to form binary masks for original images. Then we used the binary masks to define the outer edges of each specimen and an additional margin of at least 20 pixels was added before cropping out

a square region. Finally, we scaled the cropped patches to the size of 256x256 pixels, which is the standard image size used with CNNs. Examples of the original images, cropped images, and corresponding binary masks are presented in Figure 4.

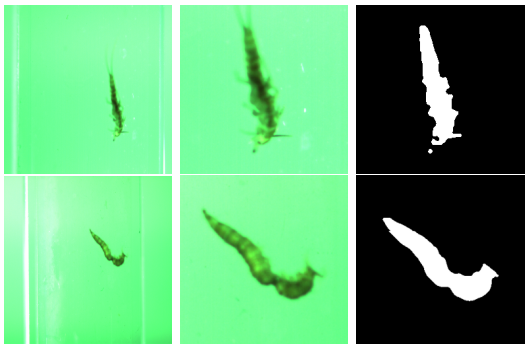


Figure 4: From left to right: original image, cropped image, corresponding binary mask. Top row - *Leuctra sp.*, bottom row - *Atherix ibis*

3.4. Database Description

The proposed database contains 15,074 images of benthic macroinvertebrates from 64 categories (i.e., taxa) with 7 to 577 images per category. The database covers the majority of the taxa used in national river water quality index calculations and represented species can be found in most natural Finnish rivers. The database contains original and cropped images for all specimens along with binary masks corresponding to the cropped image. All images and masks are in PNG-format. Also features extracted by CNNs (i.e., deep features) from the images are provided for public use at `http://urn.fi/urn:nbn:fi:csc-kata20170615175247247938`.

To scale the database for distinct classification and identification purposes, the database is split into 3 datasets:

- Dataset 1 contains the full database with 64 categories;
- Dataset 2 is constructed from 29 most abundant categories, where the smallest category contains 230 images;
- Dataset 3 is the smallest and the most balanced one; it is composed from 9 categories of macroinvertebrates with approximately 350 images in each.

Dataset 1 contains all 64 of the macroinvertebrate categories and is very unbalanced. Therefore, it demonstrates the performance of the tested methods in the most challenging conditions. Dataset 2 has been previously used for experiments in [42, 41]. Dataset 3 is generated in order to compare results of the new methods with the ones applied in the previous works [22, 25, 24, 34].

The images have a green background caused by the operator's choice of the white balance settings during imaging process. However, this does not significantly affect the classification ability of the automated classifier. Most of the macroinvertebrate specimens in the database are represented by a pair of images, but, for some specimens, there is only a single image due to a failure of the imaging system or unsuccessful stitching. The description of the database with the category names, the number of images per category, the number of specimens per category, and dataset memberships is given in Table 1. To give a better impression of the inter-category similarity and also large intra-category variations we show nine example images from each category in Dataset 2 in Figures 5 and 6.

We formed 10 different partitions of each dataset. Each partition contains the full data divided randomly into training (50%), testing (30%), and validation (20%) sets with the following constraints: We maintained these proportions also for each category in each partition (i.e., stratified partitions) and we always placed the two images of the same specimen into the same set (training, testing, or validation) to avoid presenting the same specimen in both training and testing data. We also share these partitions for public use. A separate set of deep features is provided for each partition corresponding to a network trained using the respective training and validation sets.

3.5. Comparison to Other Databases

In this subsection, we compare the proposed database with other similar databases used for fine-grained classification of freshwater macroinvertebrates (Our previous [24], STONEFLY9 [33], EPT29 [29]). The databases in the comparison differ in the number of images, image quality, availability for public use, supplementary data provided with the images, and many other parameters as described in Table 2.

The proposed database represents a set of taxa typical of Northern Europe, whereas EPT29 [29] contains fauna typical of North America. The proposed database contains images with a 256x256 pixel resolution, while the EPT29 database contains microscopy images with a very high resolution, i.e., 2560x1920 pixels. However, the image size of the proposed database is large enough to capture the most important features of different taxa. Similar images can be obtained in the future for automatic classification with relatively little human effort and low costs as described in subsection 3.2, while microscopy images require a lot more manual labor and are more expensive due to the high cost of the microscope. Furthermore, the proposed database occupies only 1.3 GB. Although the number of images in EPT29 database is less than a quarter of that of the proposed database, its high-resolution images occupy more than 50 times the storage space required by the proposed database. In the proposed database, most of the specimens are represented by a pair of images presenting it from two perpendicular angles. Even though EPT29 presents each specimen with several images (avg. 3), all of them show

Table 1: Database Description

# ID	# Ims/Cat	# Spes/Cat	Category name	Dataset 1	Dataset 2	Dataset 3
1	577	290	*Ephemerella aroni (aurivillii)*	✓	✓	-
2	480	240	*Leptophlebia sp.*	✓	✓	-
3	468	238	*Baetis rhodani*	✓	✓	-
4	468	237	*Elmis aenea* larva	✓	✓	-
5	465	234	*Oulimnius tuberculatus* larva	✓	✓	-
6	460	230	*Isoperla sp.*	✓	✓	-
7	458	230	*Habrophlebia sp.*	✓	✓	-
8	455	228	*Baetis niger*	✓	✓	-
9	447	232	*Asellus aquaticus*	✓	✓	-
10	438	227	*Oxyethira sp.*	✓	✓	-
11	436	223	*Hydraena* adult	✓	✓	-
12	428	219	*Ithytrichia lamellaris*	✓	✓	-
13	418	220	*Simuliidae*	✓	✓	-
14	417	212	*Micrasema gelidum*	✓	✓	-
15	414	208	*Nemoura sp.*	✓	✓	-
16	409	206	*Heptagenia dalecarlica*	✓	✓	-
17	408	204	*Psychodiidae*	✓	✓	-
18	404	205	*Taeniopteryx nebulosa*	✓	✓	-
19	395	202	*Limnius volckmari* adult	✓	✓	✓
20	387	194	*Protonemura sp.*	✓	✓	✓
21	378	204	*Elmis aenea* adult	✓	✓	✓
22	378	189	*Leuctra sp.*	✓	✓	✓
23	372	196	*Micrasema setiferum*	✓	✓	✓
24	367	187	*Dicranota*	✓	✓	✓
25	343	173	*Ameletus inopinatus*	✓	✓	✓
26	330	173	*Philopotamus montanus*	✓	✓	✓
27	322	174	*Ceratopogonidae*	✓	✓	✓
28	280	142	*Hemerodromia*	✓	✓	-
29	230	121	*Atherix ibis*	✓	✓	-
30	199	101	*Plectrocnemia conspersa*	✓	-	-
31	191	96	*Paraleptophlebia sp.*	✓	-	-
32	188	98	*Hydracarina*	✓	-	-
33	179	92	*Oulimnius tuberculatus* adult	✓	-	-
34	170	87	*Hydropsyche pellucidula*	✓	-	-
35	168	84	*Chimarra marginata*	✓	-	-
36	152	76	*Leuctra nigra*	✓	-	-
37	146	75	*Baetis digitatus*	✓	-	-
38	143	72	*Siphonoperla burmeisteri*	✓	-	-
39	132	67	*Polycentropu flavomaculatus*	✓	-	-
40	127	64	*Diura nanseni*	✓	-	-
41	127	66	*Gyraulus sp.*	✓	-	-
42	122	61	*Elodes*	✓	-	-
43	120	63	*Ceraclea excisa*	✓	-	-
44	94	48	*Rhyacophila nubila*	✓	-	-
45	94	48	*Sericostoma personatum*	✓	-	-
46	92	47	*Gammarus lacustris*	✓	-	-
47	83	43	*Eloeophila sp.*	✓	-	-
48	83	56	*Sialis lutaria*	✓	-	-
49	74	37	*Limnephilidae*	✓	-	-
50	69	35	*Centroptilum luteolum*	✓	-	-
51	68	34	*Lepidostoma hirtum*	✓	-	-
52	60	32	*Hydropsyche siltalai*	✓	-	-
53	57	29	*Chelifera*	✓	-	-
54	56	28	*Chironomidae*	✓	-	-
55	56	28	*Rhyacophila fasciata obtilerata*	✓	-	-
56	36	18	*Cyrnus flavidus*	✓	-	-
57	30	15	*Pisidium sp.*	✓	-	-
58	28	14	*Silo pallipes*	✓	-	-
59	25	13	*Polycentropus irroratus*	✓	-	-
60	24	12	*Hydropsyche saxonica*	✓	-	-
61	20	10	*Brachyptera risi*	✓	-	-
62	12	6	*Agrypnia sp.*	✓	-	-
63	10	6	*Neureclipsis bimaculata*	✓	-	-
64	7	6	*Callicorixa wollastoni*	✓	-	-

(1) *Ephemerella aroni* (2) *Leptophlebia sp.* (3) *Baetis rhodani* (4) *Elmis aenea* larva (5) *Oulimnius tuberculatus* larva

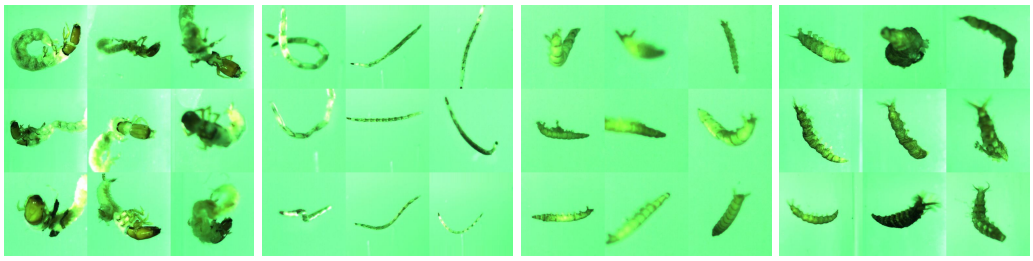
(6) *Isoperla sp.* (7) *Habrophlebia sp.* (8) *Baetis niger* (9) *Asellus aquaticus* (10) *Oxyethira sp.*

(11) *Hydraena* adult (12) *Ithytrichia lamellaris* (13) *Simuliidae* (14) *Micrasema gelidum* (15) *Nemoura sp.*

(16) *Heptagenia dalecarlica* (17) *Psychodiidae* (18) *Taeniopteryx nebulosa* (19) *Limnius volckmari* adult (20) *Protonemura sp.*

(21) *Elmis aenea* adult (22) *Leuctra sp.* (23) *Micrasema setiferum* (24) *Dicranota* (25) *Ameletus inopinatus*

Figure 5: Example images from categories 1-25 of Dataset 2



(26) *Philopotamus montanus* (27) *Ceratopogonidae* (28) *Hemerodromia* (29) *Atherix ibis*

Figure 6: Example images from categories 26-29 of Dataset 2

Table 2: Comparison to other macroinvertebrate databases

	Dataset 1	**Dataset 2**	**Dataset 3**	**Our-previous**	**STONEFLY9**	**EPT29**
Number of categories	64	29	9	8	9	29
Total number of images	15,074	11,832	3,272	1,350	3,826	4,722
Number of views per sample	avg. 2	avg. 2	avg. 2	1	avg. 5	avg. 3
Image resolution	256x256	256x256	256x256	(150-1000)x(150-1000)	2560x1920	2560x1920
Availability for public use	✓	✓	✓	-	✓	✓
Availability of features	✓	✓	✓	-	✓	✓
Availability of binary masks	✓	✓	✓	-	✓	✓
Size of the database	1.3GB	1GB	0.25GB	-	22GB	66GB

Figure 7: Left: A pair of images of a single specimen from the proposed database. Right: Three views of a single specimen from the EPT29 database.

it from a similar, dorsal, point of view. Examples of both are shown in Figure 7.

4. Experiments

4.1. Experimental Setup

4.1.1. Evaluation Protocol

Most specimens in the database are represented by a pair of images. Therefore, we chose specimen-based accuracy as the classification performance metric. Specimen-based accuracy is calculated based on the category labels obtained for both images (whenever two images are available): if both views agree on the category label, the specimen is classified accordingly, if the category labels of different views do not match, the category label with highest confidence is selected. The exact approach to evaluate the label confidence depends on the classifier. For example, for CNNs we compute the average network output and use this for classification. For k-nearest neighbors classifier (k-NN), the category of the closest nearest neighbor is selected. The final classification accuracy is the proportion of correctly classified specimens among all tested specimens. The specimen-based accuracy is a natural choice considering also the future semi-automatic macroinvertebrate monitoring system, where the classification step is fully automatic and its main objective is to classify each specimen correctly using all available images.

4.1.2. Classifiers and Their Parameters

We conducted most experiments using a state-of-the-art deep learning technique, i.e., CNNs. More specifically, we used the MatConvNet [46] implementation of the AlexNet CNN architecture [27]. MatConvNet is a Matlab implementation enabling deep learning either by using one of the state-of-the-art CNN architectures or designing a new one. Also pretrained models are available, but in this work, we trained each network from scratch. The applied AlexNet architecture has five convolution layers followed by three fully-connected Multilayer Perceptron (MLP) layers. The last MLP layer is followed by a softmaxloss(train)/softmax(test) layer, but in our tests, we considered the output of the last MLP layer, because we did not want to suppress the scores for secondary category options (i.e., other high output values besides the winner category) before computing the combined confidence score used in specimen-based accuracy estimation. For training, we used 100 training epochs and saved the model after every epoch. For testing, we then selected the model giving the highest accuracy on the validation set. The batch size of 64 was used and, as the learning rate, we used directly the logarithmic-scale learning schedule used in MatCon-

vNet's ImageNet example.

Besides direct CNN experiments, we extracted the deep features for each dataset and partition using the respective trained AlexNet CNN-models. The features were extracted after the second MLP-layer (layer 'fc7' in the MatConvNet example). According to the AlexNet model, the feature vector dimension was thus 4096. These features are provided as a supplementary material along with the images of the proposed database.

The extracted deep features were then used to train different classifiers, namely k-NN, nearest centroid classifier (NCC), Support Vector Machines (SVMs) [8, 10], Random Forest (RF) [9], Random Bayes Array (RBA) [5], Ridge Regression, regularized Extreme Learning Machine (RELM) [19], and Graph Embedded Extreme Learning Machine (GEELM) [20]. For SVMs, we used linear, polynomial, and Radial Basis Function (RBF) kernels and the one-against-all multi-class classification strategy. We also applied two dimension reduction techniques, Linear Discriminant Analysis (LDA) and Reference Vector Linear Discriminant Analysis (RV-LDA), and used the resulting features for classification with NCC. To optimize the parameters of the applied methods, we defined sets of possible parameter values, trained the classifiers on the training set using each parameter value or value combination, and selected the best parameters for each data partition based on the performance on the validation set. The final classifier was trained using the optimized parameters and both training and validation sets.

The parameters were selected as follows: K from [1, 3, 5, 7, 9] (for k-NN), $C = 10^b$, where $b = -3, ..., 3$ (SVMs, RELM, GEELM), $\sigma = A * 10^b$, where $b = -3, ..., 3$ and A is the mean of the distances between samples (SVM(rbf), RELM, GEELM), p from [1,2,4] (SVM(poly)), the number of hidden layer neurons from [100, 250, 500, 1000, 1500] (RELM, GEELM), *ntree* from [500, 1000, 2000, 3000, 5000] (RF), *mtry* from [10, 30, 64, 90, 200] (RF). For RBA, the number of Bayesian classifiers was selected from [500, 1000, 1500, 2000] and, for each classifier, we used 10, 30, or 64 randomly selected feature vector elements. To improve the RBA results, we did not use feature vector elements, whose variance in the training set was smaller than 0.05 for Dataset 2 or smaller than 0.01 for Dataset 3. Thus, the number of elements exploited with different partitions was varying for RBA. We did not apply RBA for Dataset 1, because for many partitions the feature variances were too small. For RV-LDA, automatic stopping and the maximum of 10 iterations was used. SVM with the polynomial kernel was not applied on Dataset 1 due to its computational complexity.

4.2. Experimental Results

4.2.1. CNN Classification

Table 3 shows specimen-based classification accuracies obtained for all three datasets of the proposed database using the described CNN model. As expected, the accuracies are highest for Dataset 3, which is the most balanced of the datasets. Dataset 1 has the highest imbalance between the categories, e.g., *Callicorixa wollastoni* and *Neureclipsis bimaculata* have less than five training images.

Table 3: Classification accuracy of AlexNet on the proposed datasets

Data Partition	Dataset 1	Dataset 2	Dataset 3
1	75.31	81.36	91.41
2	75.52	81.47	92.38
3	75.69	81.47	89.84
4	76.12	79.71	87.50
5	75.01	81.52	86.33
6	76.38	79.33	91.60
7	76.63	81.58	90.43
8	76.29	80.81	90.43
9	74.37	80.81	91.80
10	76.04	82.35	89.65
mean	75.74	81.04	90.14
std	0.70	0.91	1.93

The accuracies are shown separately for each data partition along with their mean and standard deviation. To give further insight into occurring errors, we also give the average confusion matrix for Dataset 2 in Table 4 and for Dataset 3 in Table 5. The values shown are the average numbers of test specimens falling to each category and all the values of 3.0 or higher have been bolded. Unless on the diagonal, the bolded values show where the main confusion in the classification occurs. The j^{th} column of the confusion matrix shows into which categories the test specimens from the j^{th} category are classified. In the optimal case, all the specimens are classified to the j^{th} category and the only non-zero value appears on the diagonal. Similarly the i^{th} row shows from which other categories specimens have been misclassified to the i^{th} category.

The confusion matrix for Dataset 2 reveals that all the 87 test specimens from category 1, *Ephemerella aroni*, were always correctly classified. When considering the sample images in Figure 5, it is evident that this category has some distinct features that even a non-expert can detect. For specimens from categories 17, *Psychodiidae*, and 27, *Ceratopogonidae*, the largest confusion values in the matrix are 0.5. Also these categories look somehow distinguishable to a non-expert viewer. Specimens from categories 4, 9, 11, 13, 16, 19, 21, 24, 26, and 28 have been classified quite successfully as well with largest confusion values being less than 3, but these categories are already more difficult to distinguish. For example, some specimens from categories 13 and 19 (*Simuliidae* vs. *Limnius volckmari* adult) or from categories 28 and 29 (*Hemerodromia* vs. *Atherix ibis*) look very similar. The largest confusion has occurred between categories 6 and 20 (*Isoperla sp.* vs. *Protonemura sp.*), categories 10 and 12 (*Oxyethira sp.* vs. *Ithytrichia lamellaris*), categories 2 and 7 (*Leptophlebia*

Table 4: Average confusion matrix of AlexNet for Dataset 2. The value at (i, j) is the average number of test specimens from category j classified to category i.

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27	28	29
1	**87.0**	0.0	0.0	0.0	0.0	0.0	0.0	0.2	0.2	0.0	0.0	0.0	0.0	0.0	0.0	0.7	0.0	0.0	0.0	0.2	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.1
2	0.0	**49.9**	2.4	0.0	0.0	1.8	**7.5**	2.6	0.1	0.1	0.0	0.4	0.8	1.1	0.0	1.4	0.0	0.0	0.1	0.1	0.0	0.9	0.7	0.3	0.3	0.1	0.0	0.0	0.1
3	0.0	1.5	**55.0**	0.0	0.2	1.6	0.3	**3.0**	1.1	0.2	0.0	0.1	0.0	0.0	0.2	0.5	0.0	0.2	0.0	0.3	0.0	0.0	0.1	0.0	**5.0**	0.0	0.0	0.0	0.0
4	0.0	0.0	0.1	**66.2**	**5.2**	0.1	0.0	0.0	0.2	0.0	0.2	0.2	0.0	0.2	0.0	0.1	0.1	0.0	0.0	0.2	0.1	0.0	0.0	0.0	0.3	0.0	0.1	0.0	0.5
5	0.0	0.0	0.1	2.2	**61.7**	0.4	0.3	1.8	0.0	0.0	0.0	0.0	0.0	0.2	0.0	0.0	0.3	0.0	0.2	0.1	0.0	0.6	0.2	0.0	1.5	0.0	0.0	0.6	0.0
6	0.0	0.9	0.2	0.6	0.2	**40.3**	0.5	0.1	1.6	0.0	0.0	0.2	0.0	0.1	**3.4**	2.6	0.0	0.6	0.0	**8.0**	0.0	1.1	0.0	0.0	0.3	0.6	0.0	0.0	0.6
7	0.0	**8.6**	1.9	0.0	0.3	0.9	**44.5**	**3.6**	1.0	0.4	0.0	0.1	1.9	1.6	0.1	0.3	0.3	0.0	1.4	0.3	0.0	**3.7**	0.1	0.6	1.2	0.8	0.0	0.2	0.4
8	0.0	**4.7**	**4.1**	0.0	0.5	0.4	**4.9**	**45.3**	0.0	1.2	0.0	0.4	0.6	0.2	0.0	0.2	0.4	0.0	0.4	0.0	0.0	0.0	0.2	0.0	**8.1**	0.0	0.1	0.1	0.2
9	0.0	0.6	**3.3**	0.0	0.0	1.1	0.3	0.0	**58.4**	0.0	0.0	0.0	0.2	0.0	**4.6**	0.2	0.0	0.4	0.0	1.2	0.0	0.2	0.0	0.0	0.4	0.1	0.0	0.0	0.0
10	0.0	0.5	0.1	0.7	0.1	0.0	0.5	1.4	0.0	**54.2**	0.4	**9.7**	0.6	2.4	0.0	0.0	0.3	0.0	0.0	0.0	0.0	0.0	0.3	0.5	0.4	0.0	0.1	1.0	0.2
11	0.0	0.0	0.0	0.0	0.1	0.0	0.0	0.1	0.0	0.0	**64.1**	0.0	0.0	0.0	0.0	0.1	0.2	0.0	0.0	0.0	1.6	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.2
12	0.0	0.1	0.0	0.1	0.1	0.0	0.1	0.9	0.0	**8.7**	0.1	**53.2**	0.0	0.6	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.4	0.0	0.0	0.0	0.0	0.2	0.1
13	0.0	0.3	0.2	0.0	0.0	0.0	0.5	0.8	0.0	0.3	0.0	0.0	**58.0**	0.2	0.0	0.3	0.2	0.0	0.5	0.0	0.0	0.0	0.1	0.4	0.0	0.1	0.0	0.0	0.2
14	0.0	0.5	0.0	0.0	0.0	0.1	0.1	0.2	0.0	0.2	0.0	0.0	0.2	**49.3**	0.0	1.0	0.0	0.0	0.1	0.0	0.0	0.0	**5.0**	1.0	0.1	0.3	0.0	0.2	0.6
15	0.0	0.2	0.3	0.1	0.3	**4.6**	1.3	0.0	1.5	0.0	0.0	0.0	0.0	0.0	**44.9**	0.2	0.0	**4.4**	0.0	**6.2**	0.0	0.5	0.0	0.0	0.1	0.0	0.0	0.0	0.0
16	0.0	0.5	0.3	0.8	0.1	1.2	0.1	0.0	0.4	0.0	0.0	0.0	0.1	0.5	0.0	**48.2**	0.0	0.0	0.0	1.5	0.0	0.2	0.0	0.2	0.1	0.6	0.0	0.2	2.9
17	0.0	0.0	0.0	0.3	0.3	0.0	0.2	0.1	0.0	0.0	0.8	0.2	0.5	0.6	0.0	0.0	**59.3**	0.0	0.6	0.0	0.0	0.0	0.2	0.2	0.1	0.0	0.1	0.8	0.0
18	0.0	0.0	0.4	0.0	0.0	1.5	0.0	0.0	1.6	0.0	0.0	0.0	0.0	0.0	**5.7**	0.0	0.0	**54.4**	0.0	**4.1**	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
19	0.0	0.0	1.2	0.1	0.3	0.0	0.9	0.7	0.0	0.0	0.1	0.0	1.1	0.2	0.1	0.1	0.1	0.0	**56.9**	0.0	0.0	0.0	0.1	0.0	0.3	0.0	0.0	0.1	0.2
20	0.0	0.4	0.1	0.0	0.0	**10.5**	0.1	0.0	1.0	0.1	0.1	0.0	0.0	0.0	2.5	2.0	0.0	2.0	0.0	**34.2**	0.3	0.2	0.1	0.0	0.0	0.1	0.0	0.0	0.5
21	0.0	0.0	0.0	0.3	0.0	0.0	0.0	0.0	0.0	0.0	1.2	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	**60.0**	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
22	0.0	0.4	0.6	0.0	0.2	0.9	**4.5**	1.0	0.5	0.3	0.0	0.1	0.1	0.1	1.1	0.0	0.0	0.0	0.4	0.9	0.0	**46.0**	0.3	1.1	1.6	0.7	0.1	1.0	0.2
23	0.0	0.3	0.1	0.1	0.0	0.0	0.0	0.2	0.0	0.9	0.0	0.7	0.2	**4.0**	0.0	0.3	0.3	0.0	0.3	0.0	0.0	0.0	**50.7**	0.1	0.0	0.0	0.0	0.0	0.3
24	0.0	0.9	0.0	0.0	0.0	0.5	0.2	0.0	0.6	1.0	0.0	0.0	0.4	0.5	0.0	0.7	0.0	0.0	0.0	0.5	0.0	1.0	0.2	**47.7**	0.2	2.9	0.0	0.6	**3.8**
25	0.0	0.4	1.2	0.1	1.2	0.9	1.4	**6.0**	0.1	0.2	0.0	0.1	1.0	0.7	0.0	0.6	0.0	0.0	0.0	0.7	0.0	0.3	0.0	0.0	**30.3**	0.4	0.5	0.2	0.2
26	0.0	0.9	0.1	0.0	0.0	1.2	0.2	0.1	1.2	0.0	0.0	0.0	0.0	1.2	0.1	1.2	0.0	0.0	0.0	0.2	0.0	0.1	0.2	1.9	0.3	**44.8**	0.0	0.0	0.5
27	0.0	0.0	0.3	0.0	0.0	0.4	0.1	0.1	0.2	0.0	0.0	0.2	0.1	0.0	0.1	0.2	0.4	0.0	0.1	0.0	0.0	0.8	0.0	0.0	0.5	0.0	**51.9**	0.2	0.0
28	0.0	0.0	0.0	0.4	0.2	0.0	0.5	0.7	0.2	0.8	0.0	0.4	0.1	0.0	0.1	0.0	0.1	0.0	0.0	0.1	0.0	1.3	0.0	0.9	0.9	0.3	0.0	**37.5**	0.9
29	0.0	0.4	0.0	0.0	0.0	0.6	0.0	0.1	0.1	0.4	0.0	0.0	0.1	0.3	0.1	1.1	0.0	0.0	0.0	0.2	0.0	0.1	0.1	2.1	0.0	0.2	0.1	0.1	**24.3**

Table 5: Average confusion matrix of AlexNet for Dataset 3. The value at (i, j) is the average number of test specimens from category j classified to category i.

	19	20	21	22	23	24	25	26	27
19	**57.5**	0.2	0.4	0.1	1.6	0.6	0.6	0.0	0.6
20	0.1	**52.3**	0.1	0.7	0.4	0.2	2.3	0.2	0.0
21	0.0	0.2	**61.3**	0.0	0.2	0.0	0.1	0.0	0.0
22	1.7	**3.2**	0.0	**52.0**	0.4	**4.5**	**5.5**	1.7	0.0
23	0.4	0.4	0.2	0.0	**54.0**	0.9	0.4	0.0	0.0
24	0.1	0.3	0.0	1.8	0.7	**48.1**	0.8	**4.5**	0.0
25	0.3	2.0	0.0	1.5	1.5	0.5	**39.7**	1.0	0.4
26	0.0	0.4	0.0	0.4	0.2	2.2	1.7	**44.6**	0.0
27	0.9	0.0	0.0	0.5	0.0	0.0	0.9	0.0	**52.0**

sp. vs. *Habrophlebia sp.*), and categories 8 and 25 (*Baetis niger* vs. *Ameletus inopinatus*). Indeed, all these category pairs look easily confusable in Figures 5 and 6. Thus, it can be concluded that the error sources for the CNN classification correlates with the human (non-expert) impression of category similarity.

Similar conclusions can be made looking at the confusion matrix for Dataset 3. The specimens from categories 21, *Elmis aenea* adult, and 27, *Ceratopogonidae*, have been classified with the highest accuracy. The largest confusion has occurred between categories 22 and 25 (*Leuctra sp.* vs. *Ameletus inopinatus*), categories 22 and 24 (*Leuctra sp.* vs. *Dicranota*), and categories 24 and 26 (*Dicranota* vs. *Philopotamus montanus*). Also these results agree with the human (non-expert) impression of category similarity.

The accuracy achieved by CNNs is already beyond the ability of a non-expert to distinguish specimens from different categories. Nevertheless, the database size is still too small to efficiently train the AlexNet model having over 6 million parameters from scratch. We have already considered this problem in [41], where some improvement was gained using pretrained networks and dataset enrichment. Furthermore, the AlexNet architecture has not been optimized for this particular classification task. Thus, improved results are expected in the future.

We trained the networks using a Windows laptop having Intel Core i7-7820HQ 4-core processor (at 2.90 GHz) and 64GB of RAM. When the laptop was free from other load, the average training times per epoch were approximately 9 minutes for Dataset 1, 7 minutes 30 seconds for Dataset 2, and 2 minutes 10 seconds for Dataset 3.

4.2.2. Performance Comparison of Different Classifiers

The specimen-based accuracies obtained by applying classifiers listed in Section 4.1.2 over the deep features extracted from the images are provided per data partition for Dataset 2 in Table 6. For Dataset 1 and Dataset 3, we give only means and standard deviations over the partitions in Table 7. 'Regr.' refers to Ridge Regression classifier. The original CNN results are repeated in the first column for easier comparison. The column is titled as MLP, because in that case the same deep features are handled by one more MLP layer.

The results show that for Dataset 1 and Dataset 2 the other applied classifiers could not outperform the original CNN results. The best performing classifier was SVM with the RBF kernel, while the linear SVM achieved almost similar results. SVM with the polynomial kernel, Ridge Regression, RELM, and GEELM obtained relatively high accuracies. As expected, the simplest classifiers, k-NN and

Table 6: Classification accuracies of different classifiers on deep features for Dataset 2

Data Partition	MLP	k-NN	NCC	LDA+ NCC	RV-LDA+ NCC	SVM (lin)	SVM (poly)	SVM (RBF)	RF	RBA	Regr.	RELM	GEELM
1	81.36	72.97	74.23	74.51	74.40	80.59	78.84	81.36	68.97	74.45	80.04	80.15	0.7664
2	81.47	71.55	74.67	72.04	72.04	80.70	77.80	80.65	68.86	72.57	80.21	79.99	0.7664
3	81.47	74.51	74.34	75.44	75.27	81.47	80.76	82.02	71.66	73.30	80.26	81.25	0.8037
4	79.71	69.24	71.11	71.44	71.38	80.65	77.19	79.82	68.70	68.42	78.29	79.11	0.7478
5	81.52	73.63	75.33	71.82	71.55	80.81	79.06	81.30	69.90	73.46	80.37	80.98	0.7314
6	79.33	71.16	72.26	69.46	69.30	79.22	76.26	78.56	67.21	71.33	78.62	78.56	0.7769
7	81.58	74.34	74.84	71.88	71.88	80.43	79.50	81.25	70.89	74.56	80.87	81.30	0.8026
8	80.81	72.92	73.96	73.85	74.01	81.25	78.34	80.98	71.05	71.93	80.43	80.37	0.7922
9	80.81	70.01	71.60	70.34	70.45	79.17	76.81	78.56	68.20	67.49	77.91	79.11	0.7785
10	82.35	71.38	74.45	72.97	73.25	81.74	78.78	82.51	71.18	72.86	80.87	81.30	0.7741
mean	81.04	72.17	73.68	72.37	72.35	80.60	78.33	80.70	69.56	72.14	79.79	80.21	77.40
std	0.91	1.79	1.47	1.84	1.86	0.85	1.35	1.34	1.41	1.01	1.09	1.01	2.28

Table 7: Classification accuracies of different classifiers on deep features for Dataset 1 and Dataset 3

Data Partition	MLP	k-NN	NCC	LDA+ NCC	RV-LDA+ NCC	SVM (lin)	SVM (poly)	SVM (RBF)	RF	RBA	Regr.	RELM	GEELM
Dataset 1													
mean	75.74	64.26	67.37	67.61	67.64	74.65		75.19	61.70		72.50	73.05	71.12
std	0.70	0.65	1.17	1.01	1.03	0.52		0.42	0.64		0.47	0.58	1.74
Dataset 3													
mean	90.14	81.23	85.00	19.98	21.04	91.04	88.93	89.57	83.50	86.95	91.00	91.15	89.39
std	1.93	1.56	1.57	4.92	4.82	1.43	1.68	1.41	1.59	1.74	1.46	1.34	3.23

NCC, have clearly more moderate results. The worst performing classifier is RF, which likely suffers from the high dimensionality (4096) of the deep features. The feature reduction techniques, LDA and RV-LDA, kept the NCC performance close to the original, while the feature vector dimensionality was significantly lower.

For the smallest dataset, Dataset 3, the results are somewhat different. The highest accuracy was obtained by RELM. Also the linear SVM and Ridge Regression classifier could further improve the CNN results. The feature reduction techniques were not able to maintain the discrimination power of the original deep features. Likely the final feature vector dimensionality (8) is already too low.

5. Conclusions

In this paper, we presented a new benchmark database of benthic macroinvertebrates (often used in freshwater biomonitoring) for the purpose of automatic fine-grained classification. The database contains 64 categories of macroinvertebrates and is the largest publicly available macroinvertebrate database, in terms of numbers of both categories and images. We also described a semi-automatic imaging technique, which significantly reduces the expert time and effort along with the overall cost for water quality monitoring. Furthermore, we performed preliminary classification experiments using deep CNNs and also with several other well-known classifiers using the deep features extracted from the trained CNNs. The results show that the automatic classification methods can already outperform the ability of a non-expert human to distinguish between different macroinvertebrate taxa and classification accuracy is approaching that of a trained expert.

In the future, we strive for automated macroinvertebrate classification in freshwater biomonitoring. Once CNNs are trained, classification is quick and, therefore, the classification step can be incorporated directly into the imaging pipeline. We will further optimize the parameters and network architecture used for CNN classification. We will do experiments with non-square cropped regions. We will use pretrained networks and exploit transfer learning techniques. Transfer learning techniques will be necessary also if the imaging technique is further improved and the image properties are slightly different, e.g. in terms of lighting or color balance, in the new images. Transfer learning will be necessary to exploit the previously acquired learning with the new samples.

Acknowledgment

The authors would like to thank the Academy of Finland for the grants no. 288584, 289076, and 289104 funding the DETECT consortium's project (Advanced Computational and Statistical Techniques for Biomonitoring and Aquatic Ecosystem Service Management).

References

[1] Directive 2000/60/EC of the European Parliament and of the Council of 23 October 2000 on establishing a framework for community action in the field of water policy. *Journal of the European Communities*, L327/1:1–72, 2000.

[2] Directive 2008/56/EC of the European Parliament and of the Council of 17 June 2008 establishing a framework for community action in the field of marine environmental policy. *Official Journal of the European Union*, L164/19, 2008.

[3] 95th Congress. US Public Law 95-217, December 17 1977, Clean Water Act. 1977.

[4] Shivani Agarwal and Dan Roth. Learning a sparse representation for object detection. In *Proc. of European Conference on Computer Vision (ECCV)*, pages 113–127, 2002.

[5] J Ärje, S Kärkkäinen, T Turpeinen, and K Meissner. 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, 2013.

[6] Thomas Berg, Jiongxin Liu, Seung Woo Lee, Michelle L Alexander, David W Jacobs, and Peter N Belhumeur. Birdsnap: Large-scale fine-grained visual categorization of birds. In *Proc. of IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2019–2026, 2014.

[7] Sebastian Birk, Wendy Bonne, Angel Borja, Sandra Brucet, Anne Courrat, Sandra Poikane, Angelo Solimini, Wouter Van de Bund, Nikolaos Zampoukas, and Daniel Hering. Three hundred ways to assess Europe's surface waters: an almost complete overview of biological methods to implement the Water Framework Directive. *Ecological Indicators*, 18:31–41, 2012.

[8] Bernhard E Boser, Isabelle M Guyon, and Vladimir N Vapnik. A training algorithm for optimal margin classifiers. In *Proc. of fifth annual workshop on Computational learning theory*, pages 144–152. ACM, 1992.

[9] Leo Breiman. Random forests. *Machine Learning*, 45(1):5–32, 2001. ISSN 1573-0565. doi: 10.1023/A:1010933404324.

[10] Corinna Cortes and Vladimir Vapnik. Support-vector networks. *Machine learning*, 20(3):273–297, 1995.

[11] Phil F Culverhouse. Human and machine factors in algae monitoring performance. *Ecological Informatics*, 2(4):361–366, 2007.

[12] Philip F Culverhouse, Norman Macleod, Robert Williams, Mark C Benfield, Rubens M Lopes, and Marc Picheral. An empirical assessment of the consistency of taxonomic identifications. *Marine Biology Research*, 10(1):73–84, 2014.

[13] David Dudgeon, Angela H Arthington, Mark O Gessner, Zen-Ichiro Kawabata, Duncan J Knowler, Christian Lévêque, Robert J Naiman, Anne-Hélène Prieur-Richard, Doris Soto, Melanie LJ Stiassny, et al. Freshwater biodiversity: importance, threats, status and conservation challenges. *Biological reviews*, 81(02):163–182, 2006.

[14] Vasco Elbrecht, Ecaterina Vamos, Kristian K. Meissner, Jukka Aroviita, and Florian Leese. Assessing strengths and weaknesses of dna metabarcoding based macroinvertebrate identification for routine stream monitoring. *Methods in Ecology and Evolution*, 2017. doi: 10.1111/2041-210X.12789.

[15] Mark Everingham, SM Ali Eslami, Luc Van Gool, Christopher KI Williams, John Winn, and Andrew Zisserman. The Pascal visual object classes challenge: A retrospective. *International Journal of Computer Vision*, 111(1):98–136, 2015.

[16] Peter Haase, Steffen U Pauls, Karin Schindehütte, and Andrea Sundermann. 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, 2010.

[17] Haibo He and Edwardo A. Garcia. Learning from imbalanced data. *IEEE Transactions on Knowledge and Data Engineering*, 21(9):1263–1284, September 2009. ISSN 1041-4347. doi: 10.1109/TKDE.2008.239.

[18] David U Hooper, E Carol Adair, Bradley J Cardinale, Jarrett EK Byrnes, Bruce A Hungate, Kristin L Matulich, Andrew Gonzalez, J Emmett Duffy, Lars Gamfeldt, and Mary I O'Connor. A global synthesis reveals biodiversity loss as a major driver of ecosystem change. *Nature*, 486(7401):105–108, 2012.

[19] G. B. Huang, H. Zhou, X. Ding, and R. Zhang. Extreme learning machine for regression and multiclass classification. *IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics)*, 42(2):513–529, April 2012. ISSN 1083-4419.

[20] A. Iosifidis, A. Tefas, and I. Pitas. Graph embedded extreme learning machine. *IEEE Transactions on Cybernetics*, 46(1): 311–324, Jan 2016. ISSN 2168-2267. doi: 10.1109/TCYB.2015.2401973.

[21] Henry Joutsijoki and Martti Juhola. DAGSVM vs. DAGKNN: An experimental case study with benthic macroinvertebrate dataset. In Petra Perner, editor, *Proc. of International Conference in Machine Learning and Data Mining in Pattern Recognition (MLDM)*, pages 439–453. Springer Berlin Heidelberg, 2012. ISBN 978-3-642-31537-4. doi: 10.1007/978-3-642-31537-4_35.

[22] Henry Joutsijoki, Kristian Meissner, Moncef Gabbouj, Serkan Kiranyaz, Jenni Raitoharju, Johanna Ärje, Salme Kärkkäinen, Ville Tirronen, Tuomas Turpeinen, and Martti Juhola. Evaluating the performance of artificial neural networks for the classification of freshwater benthic macroinvertebrates. *Ecological Informatics*, 20:1–12, 2014.

[23] Aditya Khosla, Nityananda Jayadevaprakash, Bangpeng Yao, and Fei fei Li. Novel dataset for fine-grained image categorization. In *Proc. of First Workshop on Fine-Grained Visual Categorization (CVPR)*, 2011.

[24] Serkan Kiranyaz, Moncef Gabbouj, Jenni Pulkkinen, Turker Ince, and Kristian Meissner. Network of evolutionary binary classifiers for classification and retrieval in macroinvertebrate databases. In *Proc. of IEEE International Conference on Image Processing (ICIP)*, pages 2257–2260, 2010.

[25] Serkan Kiranyaz, Turker Ince, Jenni Pulkkinen, Moncef Gabbouj, Johanna Ärje, Salme Kärkkäinen, Ville Tirronen, Martti Juhola, Tuomas Turpeinen, and Kristian Meissner. Classification and retrieval on macroinvertebrate image databases. *Computers in biology and medicine*, 41(7):463–472, 2011.

[26] Jonathan Krause, Michael Stark, Jia Deng, and Li Fei-Fei. 3D object representations for fine-grained categorization. In *Proc. of the IEEE International Conference on Computer Vision (ICCV) Workshops*, pages 554–561, 2013.

[27] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. ImageNet classification with deep convolutional neural networks. In *Proc. of Advances in Neural Information Processing Systems (NIPS)*, pages 1097–1105, 2012.

[28] Neeraj Kumar, Peter N Belhumeur, Arijit Biswas, David W Jacobs, W John Kress, Ida C Lopez, and João VB Soares. Leafsnap: A computer vision system for automatic plant species identification. In *Prof. of European Conference on Computer Vision (ECCV)*, pages 502–516. Springer, 2012.

[29] N Larios, Junyuan Lin, Mengzi Zhang, D Lytle, Andrew Moldenke, L Shapiro, and T Dietterich. Stacked spatial-pyramid kernel: An object-class recognition method to combine scores from random trees. In *Proc. of IEEE Workshop on Applications of Computer Vision (WACV)*, pages 329–335, 2011.

[30] Jeffrey Lin, Natalia Larios, David Lytle, Andrew Moldenke, Robert Paasch, Linda Shapiro, Sinisa Todorovic, and Tom Dietterich. Fine-grained recognition for arthropod field surveys: Three image collections. In *Proc. of First Workshop on Fine-Grained Visual Categorization (CVPR)*, volume 1, 2011.

[31] Yen-Liang Lin, Vlad I. Morariu, Winston Hsu, and Larry S. Davis. Jointly optimizing 3D model fitting and fine-grained classification. In *Prof. of European Conference on Computer Vision (ECCV)*, pages 466–480, Cham, 2014. Springer International Publishing. ISBN 978-3-319-10593-2. doi: 10.1007/978-3-319-10593-2_31.

[32] Jiongxin Liu, Angjoo Kanazawa, David Jacobs, and Peter Belhumeur. Dog breed classification using part localization. In *Prof. of European Conference on Computer Vision (ECCV)*, pages 172–185. Springer, 2012.

[33] David A Lytle, Gonzalo Martínez-Muñoz, Wei Zhang, Natalia Larios, Linda Shapiro, Robert Paasch, Andrew Moldenke, Eric N Mortensen, Sinisa Todorovic, and Thomas G Dietterich. Automated processing and identification of benthic invertebrate samples. *Journal of the North American Benthological Society*, 29(3):867–874, 2010.

[34] David A Lytle, Gonzalo Martínez-Muñoz, Wei Zhang, Natalia Larios, Linda Shapiro, Robert Paasch, Andrew Moldenke, Eric N Mortensen, Sinisa Todorovic, and Thomas G Dietterich. Automated processing and identification of benthic invertebrate samples. *Journal of the North American Benthological Society*, 29(3):867–874, 2010.

[35] S. Maji, J. Kannala, E. Rahtu, M. Blaschko, and A. Vedaldi. Fine-grained visual classification of aircraft. *arXiv, 1306.5151*, 2013.

[36] Kristian Meissner, Henrik Nygrd, Katarina Bjrklf, Marko Jaale, Miikka Hasari, Lauri Laitila, Jouko Rissanen, and Mirja Leivuori. Proficiency test 04/2016 - taxonomic identification of boreal freshwater lotic, lentic, profundal and North-Eastern Baltic benthic macroinvertebrates. *Reports of the Finnish Environment Institute*, 2017.

[37] Millennium Ecosystem Assessment Panel. *Ecosystems and human well-being*, volume 5. Island press Washington, DC, 2005.

[38] Maria-Elena Nilsback and Andrew Zisserman. A visual vocabulary for flower classification. In *Proc. of IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, volume 2, pages 1447–1454, 2006.

[39] Christer Nilsson, Catherine A Reidy, Mats Dynesius, and Carmen Revenga. Fragmentation and flow regulation of the world's large river systems. *Science*, 308(5720):405–408, 2005.

[40] Nobuyuki Otsu. A threshold selection method from gray-level histograms. *Automatica*, 11(285-296):23–27, 1975.

[41] J. Raitoharju, E. Riabchenko, K. Meissner, I. Ahmad, A. Iosifidis, M. Gabbouj, and S. Kiranyaz. Data enrichment in fine-grained classification of aquatic macroinvertebrates. In *Proc. of ICPR 2nd Workshop on Computer Vision for Analysis of Underwater Imagery (CVAUI)*, pages 43–48, 2016. doi: 10.1109/CVAUI.2016.020.

[42] Ekaterina Riabchenko, Kristian Meissner, Iftikhar Ahmad, Alexandros Iosifidis, Ville Tirronen, Moncef Gabbouj, and Serkan Kiranyaz. Learned vs. engineered features for fine-grained classification of aquatic macroinvertebrates. In *Proc. of International Conference on Pattern Recognition (ICPR)*, 2016.

[43] Sonia Rochatte and Alain Mestre. Syndex Report for the European Federation of Public Service Unions 2012-03 EPSU Brussels. 2012.

[44] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, et al. ImageNet large scale visual recognition challenge. *International Journal of Computer Vision*, 115(3):211–252, 2015.

[45] Grant Van Horn, Steve Branson, Ryan Farrell, Scott Haber, Jessie Barry, Panos Ipeirotis, Pietro Perona, and Serge Belongie. Building a bird recognition app and large scale dataset with citizen scientists: The fine print in fine-grained dataset collection. In *Proc. of IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 595–604, 2015.

[46] Andrea Vedaldi and Karel Lenc. MatConvNet: Convolutional neural networks for Matlab. In *Proc. of International Conference on Multimedia*, pages 689–692, 2015.

[47] Andrea Vedaldi, Siddharth Mahendran, Stavros Tsogkas, Subhransu Maji, Ross Girshick, Juho Kannala, Esa Rahtu, Iasonas Kokkinos, Matthew Blaschko, David Weiss, et al. Understanding objects in detail with fine-grained attributes. In *Proc. of IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3622–3629, 2014.

[48] Marguerite A Xenopoulos, David M Lodge, Joseph Alcamo, Michael Märker, Kerstin Schulze, and Detlef P Van Vuuren. Scenarios of freshwater fish extinctions from climate change and water withdrawal. *Global Change Biology*, 11(10):1557–1564, 2005.

Highlights

- We publish a database with 64 types of freshwater macroinvertebrates and more than 15,000 images.
- CNNs outperforms a non-expert human in classification accuracy.
- Several other well-known classifiers are applied using the features extracted from the CNNs.
- In most cases, the classifiers cannot further improve the CNN results.