water

MDPI

Article

The Effects of Data Quality on Deep Learning Performance for Aquatic Insect Identification: Advances for Biomonitoring Studies

Predrag Simović [1], Aleksandar Milosavljević [2], Katarina Stojanović [3], Dimitrija Savić-Zdravković [4], Ana Petrović [1], Bratislav Predić [2] and Djuradj Milošević [4,*]

[1] Department of Biology and Ecology, Faculty of Science, University of Kragujevac, Radoja Domanovića 12, 34000 Kragujevac, Serbia; predrag.simovic@pmf.kg.ac.rs (P.S.); ana.petrovic@pmf.kg.ac.rs (A.P.)
[2] Faculty of Electronic Engineering, University of Niš, Aleksandra Medvedeva 14, 18000 Niš, Serbia; aleksandar.milosavljevic@elfak.ni.ac.rs (A.M.); bratislav.predic@elfak.ni.ac.rs (B.P.)
[3] Department of Zoology, Faculty of Biology, University of Belgrade, Studentski trg 16, 11000 Belgrade, Serbia; k.bjelanovic@bio.bg.ac.rs
[4] Department of Biology and Ecology, Faculty of Sciences and Mathematics, University of Niš, Višegradska 33, 18000 Niš, Serbia; dimitrija.savic@pmf.edu.rs
* Correspondence: djuradj.milosevic@pmf.edu.rs

Abstract: Deep learning models, known as convolutional neural networks (CNNs), have paved the way for reliable automated image recognition. These models are increasingly being applied in research on freshwater biodiversity, aiming to enhance efficiency and taxonomic resolution in biomonitoring. However, insufficient or imbalanced datasets remain a significant bottleneck for creating high-precision classifiers. The highly imbalanced data, where some species are rare and others are common, are typical of the composition of most benthic communities. In this study, a series of CNN models was built using 33 species of aquatic insects, with datasets ranging from 10 to 80 individuals, to determine the optimal number of individuals each class should have to build a high-precision classifier. We also consider the effect of class imbalance in the training dataset and the use of oversampling technique. The results showed that a robust model with acceptable accuracy (99.45%) was achieved with at least 30 individuals per class. A strongly imbalanced dataset caused an approximately 2% decrease in classification accuracy, while a moderately imbalanced dataset had no significant effect. The application of the oversampling technique enhanced in 1.88% the accuracy of strongly imbalanced models. These findings can help effectively tailor future aquatic macroinvertebrate training datasets.

Keywords: convolutional neural network; imbalance dataset; data size; oversampling technique; Ephemeroptera; Plecoptera; Trichoptera; EPT taxa

check for updates

Academic Editors: Marcos Gomes Nogueira and Douglas Kane

Received: 13 November 2024
Revised: 17 December 2024
Accepted: 19 December 2024
Published: 25 December 2024

Citation: Simović, P.; Milosavljević, A.; Stojanović, K.; Savić-Zdravković, D.; Petrović, A.; Predić, B.; Milošević, D. The Effects of Data Quality on Deep Learning Performance for Aquatic Insect Identification: Advances for Biomonitoring Studies. *Water* **2025**, *17*, 21. https://doi.org/10.3390/w17010021

Copyright: © 2024 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (https://creativecommons.org/licenses/by/4.0/).

1. Introduction

Biomonitoring with high taxonomic resolution is fundamental for understanding ecosystem dynamics, especially in an era where freshwater ecosystems are facing increasing pressures from climate change and anthropogenic stressors [1]. While conventional methods based on physical and chemical parameters provide information on current environmental conditions and pollutant levels, biological monitoring offers a profound understanding of ecosystem health and functioning over time [2]. The biomonitoring approach uses biological indicators to assess the quality of habitats and ecosystems, offering valuable insights into the impacts of human activities and natural changes [3]. In freshwater ecosystems research, the larvae of insects from the orders Ephemeroptera, Plecoptera, and Trichoptera (collectively referred to as EPT) are considered the best bioindicators of water

quality because they typically dominate benthic fauna in terms of both abundance and species richness [4,5].

Ecologists have most often relied on traditional identification approaches based on morphological features, which often result in data with low or unverifiable taxonomic precision [6]. Furthermore, this approach is quite time and labor intensive and requires specialized knowledge and experience [6]. Additionally, challenges in identification are also attributed to phenotypic plasticity, the existence of cryptic species, and variations in size and sexual dimorphism [7,8]. Therefore, the challenges associated with the traditional identification of these larvae highlight the need for more reliable and efficient ways of processing benthic samples [1].

In recent years, there has been an exponential increase in the number of scientific publications combining the terms "biodiversity" and "deep learning" [9,10]. Special attention has been paid to the type of deep learning models known as convolutional neural networks (CNNs), which have proven effective in domains such as image classification and recognition [11,12]. A Convolutional Neural Network (CNN) analyzes images in segments, learning by identifying approximate feature correspondences at corresponding locations between paired images [13]. The main benefit of deep learning CNN architecture is its ability to self-learn and self-organize without the need for manual supervision [14].

Many authors have demonstrated the application of deep learning for the automatic identification of aquatic macroinvertebrates, both in laboratory settings [15–17] and under field conditions [18], which is vital for monitoring species populations. If widely adopted, automatic classification techniques could drastically reduce the costs of water quality analysis in aquatic ecosystems, while also decreasing the time and effort required by experts [2]. However, several technical challenges currently hinder the widespread application of computer vision in biomonitoring programs, biodiversity, and conservation studies. To achieve reliable predictions, CNN models often require many samples and the creation of balanced datasets for training [19,20]. The task of acquiring an extensive array of datasets covering relevant biomonitoring taxa is often hindered by the localized and uneven distribution of some species. This can lead to the 'imbalanced dataset problem', where some classes, the minority groups, contain significantly fewer examples than the other classes, the majority groups [21,22]. Therefore, typically rare and endangered species tend to perform poorly in automatic species identification due to limited training data [17]. Addressing this issue by excluding or misclassifying rare species can directly impact the calculation of multimetric indexes used to assess ecosystem status, potentially leading to inaccurate evaluations [23]. At the same time, CNNs tend to exhibit bias towards the majority class [19], i.e., common species that have a disproportionately large number of training data, typically those that are cosmopolitan and conspicuous [24]. More closely related species are harder to separate in classification models [17,25], and a severe problem may arise if one of these species is represented in lower numbers during the training of CNN.

Despite the increasing application of deep learning and computer vision in biomonitoring [16,17,26], very little empirical work in this field examines the effects of sample size and dataset imbalance on the performance of species classification models [25,27]. The objectives of this study are as follows: (1) to examine how classification performance changes with increasing training sample sizes, using a balanced dataset; (2) to test the influence of imbalanced datasets on the accuracy of CNN models, using a varying number of image samples for certain target species; and (3) to evaluate the effectiveness of oversampling techniques applied to minority class samples during training in improving class balance. This study is significant as it proposes a new design of standards for reference image datasets, enabling cost-effective training of more accurate algorithms for the automatic identification of indicator taxa.

2. Materials and Methods

2.1. Dataset Acquisition

Figure 1 provides an overview of the steps taken to construct the training datasets for the CNN models. The larvae of EPT species were collected alongside other macroinvertebrates as part of hydrobiological studies in aquatic ecosystems within Serbia, during the development of a freshwater biomonitoring program for Serbian surface waters. The sampling sites included various lotic and lentic habitats from the Middle Danube Basins [28]. The larvae were collected from different substrate types using a Surber sampler and preserved in 96% ethanol. They were transported and stored at the Department of Biology and Ecology, Faculty of Science, University of Kragujevac, Republic of Serbia. After separating the collected samples from extraneous materials, taxonomic identification was performed. The material was identified at the species level using relevant identification literature [29–32] by experts for this group of insects. Morphological identification of bioindicators is especially challenging for immature and early life stages [17]. To address this, our study focused on late instar larvae, which have well-developed and clearly visible morphological traits that enable accurate identification. In total, we gathered 100 individuals for each of the 33 target species (Table 1; Figure 2). Each individual was positioned in a dorsal view and photographed using either a Nikon SMZ800 stereomicroscope (Nikon Corporation, Tokyo, Japan) equipped with a Leica Flexacam C3 microscope camera (Leica Microsystems, Wetzlar, Germany) or a Nikon D900 camera (Nikon Corporation, Tokyo, Japan). According to the previous study [17] the dorsal view of these larvae provides morphological characteristics that are sufficiently informative for reliable identification. We completed our training dataset with a total of 3330 photos (one image per individual).

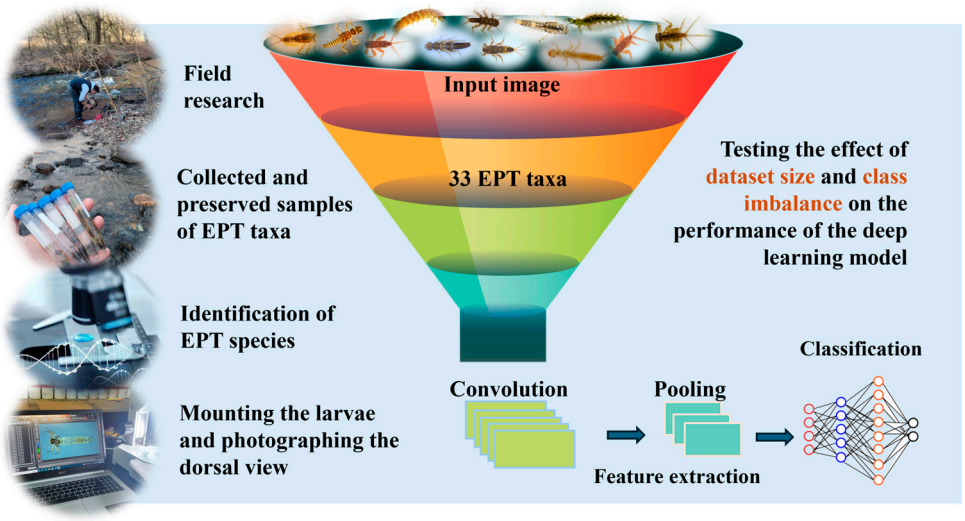


Figure 1. Schematic overview of research methodology.

Table 1. The EPT species included in the study.

Code	Order	Family	Species
1	Ephemeroptera	Baetidae	*Baetis alpinus* (Pictet, 1843)
2	Ephemeroptera	Baetidae	*Baetis rhodani* (Pictet, 1843)
3	Ephemeroptera	Caenidae	*Caenis macrura* (Stephens, 1835)
4	Ephemeroptera	Ephemeridae	*Ephemera danica* (Müller, 1764)
5	Ephemeroptera	Ephemerellidae	*Ephemerella mucronata* (Bengtsson, 1909)
6	Ephemeroptera	Ephemerellidae	*Ephemerella (Serratella) ignita* (Poda, 1761)
7	Ephemeroptera	Ephemerellidae	*Torleya major* (Klapálek, 1905)

Table 1. *Cont.*

Code	Order	Family	Species
8	Ephemeroptera	Heptagenidae	*Rhithrogena semicolorata* (Curtis, 1834)
9	Ephemeroptera	Heptagenidae	*Epeorus assimilis* (Eaton, 1865)
10	Ephemeroptera	Leptophlebidae	*Habrophlebia fusca* (Curtis, 1834)
11	Ephemeroptera	Oligoneuriidae	*Oligoneuriella rhenana* (Imhoff, 1852)
12	Ephemeroptera	Potamantidae	*Potamanthus luteus* (Linneaus, 1767)
13	Ephemeroptera	Siphlonuridae	*Siphlonurus aestivalis* (Eaton, 1903)
14	Plecoptera	Leuctridae	*Leuctra hippopus* Kempny, 1899
15	Plecoptera	Nemouridae	*Protonemura hrabei* Raušer, 1956
16	Plecoptera	Perlidae	*Perla palida* Guérin-Méneville, 1838
17	Plecoptera	Perlodidae	*Isoperla tripartita* Illies, 1954
18	Trichoptera	Brachycentridae	*Oligoplectrum maculatum* (Fourcroy, 1785)
19	Trichoptera	Brachycentridae	*Lepidostoma basale* (Kolenati, 1848)
20	Trichoptera	Goeridae	*Lithax obscurus* (Hagen, 1859)
21	Trichoptera	Goeridae	*Silo pallipes* Fabricius, 1781
22	Trichoptera	Hydropsychidae	*Hydropsyche incognita* Pitshc, 1993
23	Trichoptera	Hydropsychidae	*Cheumatopsyche lepida* Pictet, 1834
24	Trichoptera	Limnephilidae	*Limnephilus rhombicus* Linnaeus, 1758
25	Trichoptera	Limnephilidae	*Potamophylax luctuosus* (Piller & Mitterpacher, 1783)
26	Trichoptera	Limnephilidae	*Halesus digitatus* (von Paula Schrank, 1781)
27	Trichoptera	Psychomiidae	*Tinodes unicolor* (Pictet, 1834)
28	Trichoptera	Psychomiidae	*Psychomyia pusilla* (Fabricius, 1781)
29	Trichoptera	Philopotamidae	*Philopotamus montanus* (Donovan, 1813)
30	Trichoptera	Rhyacophilidae	*Rhyacophila fasciata* Hagen, 1859
31	Trichoptera	Rhyacophilidae	*Rhyacophila tristis* Pictet, 1834
32	Trichoptera	Sericostomatidae	*Sericostoma flavicorne* Schneider, 1845
33	Trichoptera	Ueonidae	*Thremma anomalum* McLachlan, 1876

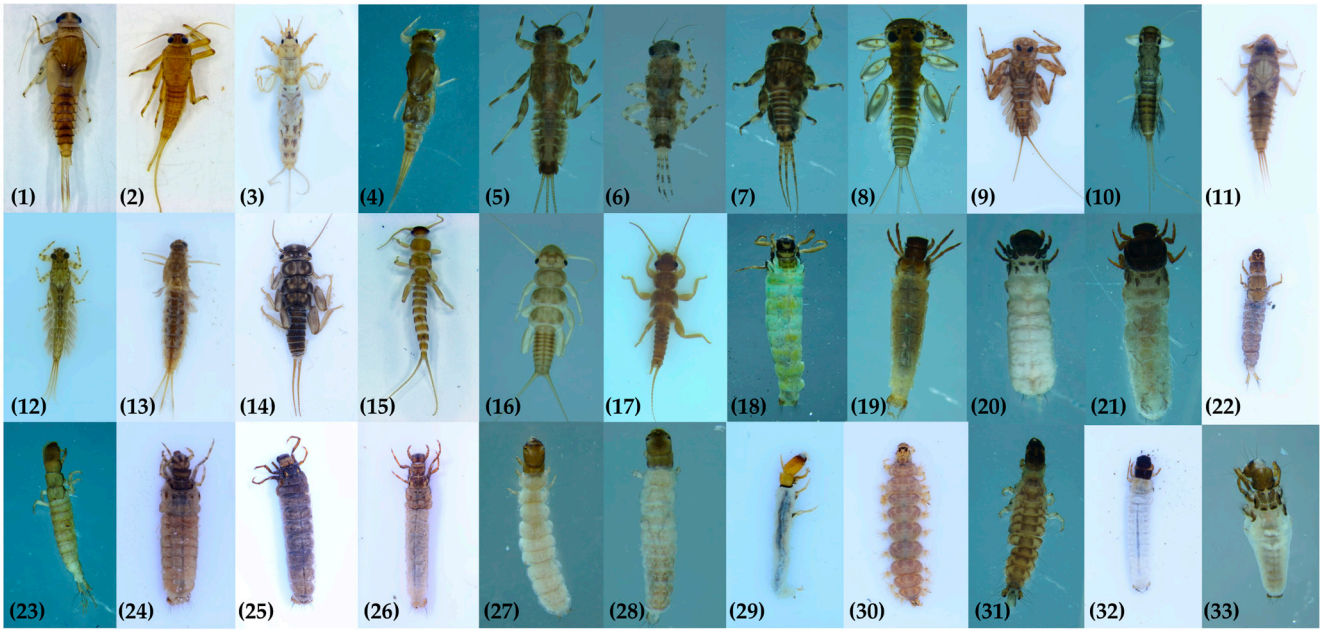


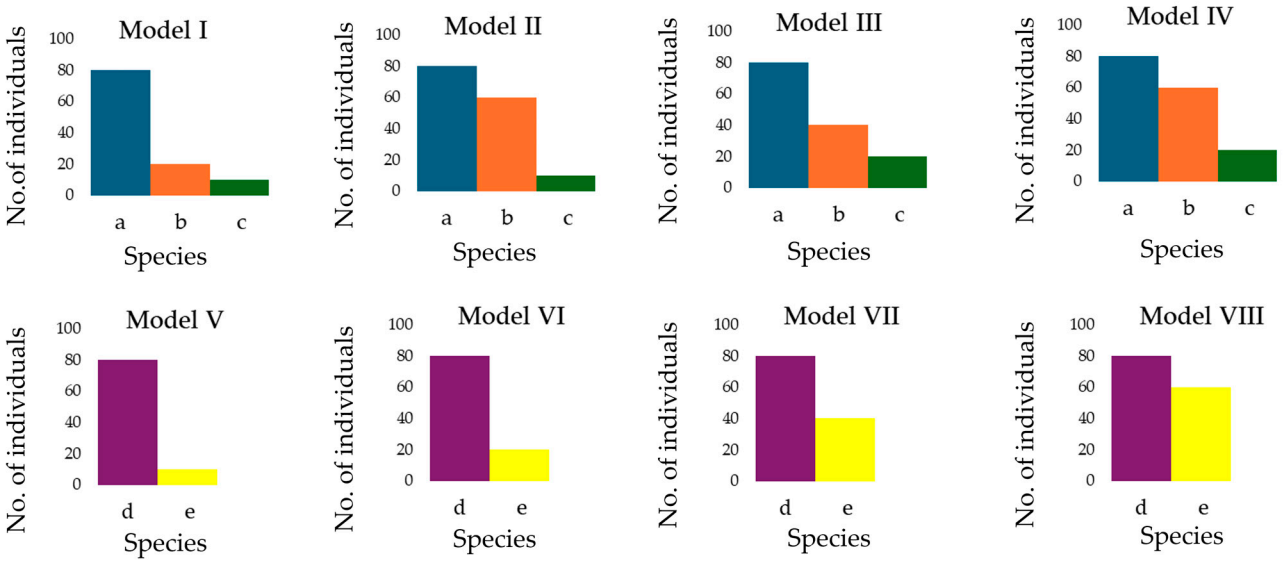
Figure 2. Dorsal view of larvae from the 33 EPT species included in the training dataset (species codes are detailed in Table 1).

2.2. Model Development

We trained a series of CNN models based on EfficientNetB2 architecture using different training image datasets [33]. A comprehensive description of the model training is provided in the study by Simović et al. [17]. For model validation and testing, 20% (10% for

validation and 10% for testing) from each species were consistently used, while the training percentages varied across models.

To assess how the number of images (individuals) affects classification accuracy, we trained a series of eight CNN models with three different configurations. In the first configuration we varied the number of individuals in the training set from 10 to 80. In the second configuration, eight subsets were randomly created to form a series of smaller training sets, allowing us to investigate the effects of dataset imbalance on the performance of species classification models. For all eight models, one species group contained 80 individuals, while the other species had progressively smaller sample sizes of 60, 40, 20, or 10 individuals, following the pattern outlined in Figure 3. In the third configuration, we used the same unbalanced dataset and applied the oversampling technique, which duplicates random examples from the minority class to balance the dataset.



Species codes according to Table 1.
a - 1, 2, 4, 5, 11, 12, 15, 17, 22, 24, 26, 28, 33
b - 3, 9, 10, 13, 14, 18, 19, 21, 23, 31
c - 6, 7, 8, 16, 20, 25, 27, 29, 30, 32
d - 2, 3, 4, 5, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 21, 22, 24, 26, 31, 33
e - 1, 6, 7, 20, 23, 25, 27, 28, 29, 30, 32

Figure 3. Sample sizes and composition of training and validation data in eight imbalanced datasets.

The minimum number of training individuals used for certain species was set at 10, as preliminary analysis indicated that sample sizes below this threshold led to challenges in cross-validation parameter tuning due to the limited representation within the class. Since model training starts with random weights, to statistically test our hypothesis, we trained five models for each experimental setup to obtain mean accuracy and standard deviation. The differences in classification accuracy of the CNN models were evaluated using a one-way Analysis of Variance (ANOVA) to assess the effects of sample size and degree of imbalance. Post hoc pairwise comparisons were conducted using Tukey's Honest Significant Difference (HSD) test to identify significant differences between groups. To evaluate the effect of the oversampling technique on classification accuracy, an independent samples *t*-test was conducted for each of the eight models. Prior to the ANOVA and *t*-test, the data were tested for normality and homogeneity of variances to ensure the validity of the analysis. All statistical analyses were performed using R software (accessed on 10 December 2024) [34].

A confusion matrix (CM) was generated for each dataset size and class imbalance in the deep learning model to reveal the true and incorrect categories of predictions. A CM is a

graphical representation that visualizes the class-wise distribution of a classification model's predictive performance, showing how the errors are distributed [35]. Python scripts used to train the model and produce visualizations are open-sourced on the project's GitHub page (https://github.com/a-milosavljevic/aiaquami-ept33, accessed on 12 December 2024). The training results used in this research are available for download in the Releases section of the GitHub project (https://github.com/a-milosavljevic/aiaquami-ept33/releases, accessed on 12 December 2024).

3. Results

3.1. The Effect of Varied Sample Size

The classification accuracy of the CNN model across five training repetitions with varying sample sizes per class (ranging from 10 to 80) significantly varied (ANOVA, F = 49,149, $p < 0.001$) (Figure 4). The model achieved a remarkable mean classification accuracy of 97.09% (range: 96.36–97.58%) when trained on only 10 individuals per each of the 33 EPT species. This accuracy was significantly improved to 99.45% with larger training dataset sizes of 30 individuals per species (Tukey's HSD, $p < 0.01$). The highest accuracy was 99.70% (range: 99.39–100%) when 50 individuals per species were used for training. The same highest average accuracy of 99.70% was also confirmed with 70 individuals per species used for training, but it decreased to 99.58% when the number of individuals increased to 80. However, the accuracy percentage remained high, so these oscillations can be considered negligible (Figure 4). In addition, low standard deviation values (Std.) indicate that the results are consistent across different tests, which means that the model is stable and reliable.

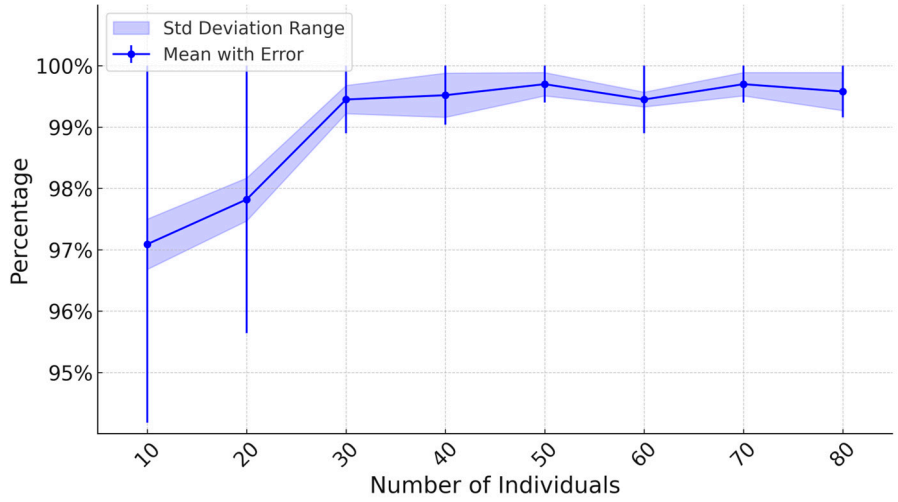


Figure 4. The variability in CNN model accuracy with increasing training set sizes (ranging from 10 to 80 individuals per species) for classifying 33 EPT species.

The deep CNN model produced confusion matrices displaying the accuracy percentage for each species, with classification errors for models with balanced datasets (Appendix A). When the CNN model was trained on 10 individuals per species, classification errors ranged from 10% to 30% across 11 species (Appendix A). However, as the training dataset size increased, error rates for individual species decreased substantially. With a test dataset of 40 individuals per species, classification errors for individual species did not exceed 10%. Only the species *Sericostoma flavicorne* and *Torleya major* showed classification errors of 10% when tested with 50 individuals per species. Misclassified individuals were identified as *Limnephilus rhombicus* and *Caenis macrura*, respectively. Similar misclassification patterns were observed in models tested with 60, 70, and 80 individuals of each

species. Additionally, in some cases, *Ephemerella mucronata* showed a classification error, with 10% of individuals being misclassified as *Ephemerella ignita* (Appendix A).

3.2. The Effect of Imbalanced Databases

The performance of the CNN model on the test set was influenced by the class imbalance levels in the training sets, resulting in significant changes in classification accuracy (ANOVA, F = 86.816, $p < 0.001$). The model achieved the best performance when the imbalance ratio in the tested EPT species was the smallest (99.82 ± 0.41% Model VIII and 99.88 ± 0.15 Model VI) (Figure 5). In those cases, we noted low classification errors for only two species, *Thremma anomalum* and *Torleya major* (Appendix B). When the degree of imbalance among individuals used for the tested EPT species was highest (Models I and II), the model's performance was significantly the lowest (Tukey's HSD, $p < 0.01$), with an accuracy of 97.09 ± 0.41% and 97.94 ± 0.59, respectively (Figure 5). When considering individual species, the CNN models generally struggled to accurately identify species with the fewest available individuals (only 10), such as *Lithax obscurus*, *T. major*, *S. flavicorne*, and *Philopotamus montanus* (Appendix B). Species *L. obscurus* and *T. major* exhibited the highest error rates, ranging between 10% and 40%. The misclassified individuals of *L. obscurus* were assigned to the image class of the species *Silo pallipes*, while *T. major* individuals were misclassified as belonging to the species *E. mucronata* (Appendix B).

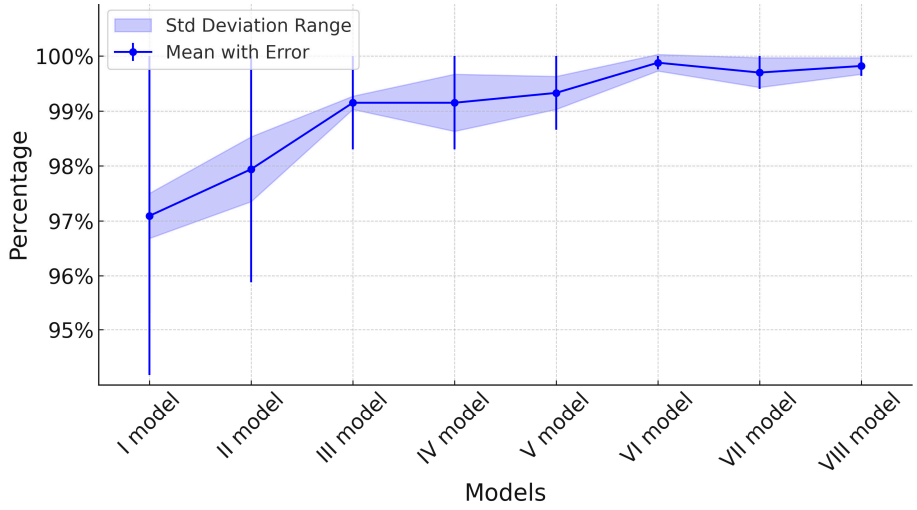


Figure 5. The variability in CNN model accuracy with increasing imbalance levels for the classification of 33 EPT species.

3.3. The Effect of Oversampling on Imbalanced Databases

The application of the oversampling technique significantly improved classification accuracy in Models I, II, and III (*t*-test, $p < 0.05$; Figure 6). Accuracy increased from 97.09% ± 0.41% to 98.97% ± 0.31% in Model I, from 97.94% ± 0.59% to 98.85% ± 0.23% in Model II, and from 99.15% ± 0.12% to 99.76% ± 0.12% in Model III (Figure 6). In contrast, Model IV showed a slight but significant decrease in accuracy, dropping from 99.88% ± 0.15% to 99.39% ± 0.27% (*t*-test, $p < 0.05$).

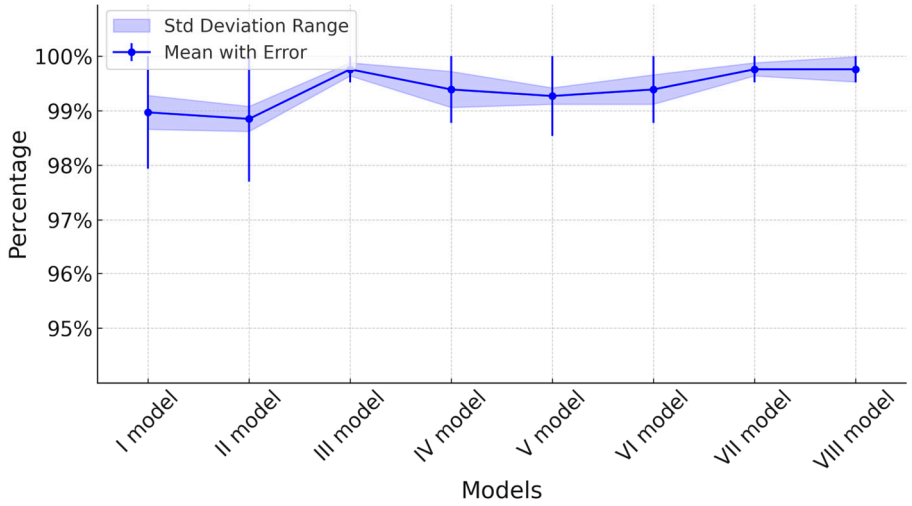


Figure 6. The variability in CNN model accuracy with increasing imbalance levels for the classification of 33 EPT with oversampling techniques applied to the minority class.

4. Discussion

Deep learning has the potential to enhance investigation processes in various ecological fields, including biomonitoring; however, a significant bottleneck in this process is the need for accurate and balanced training data [36]. Obtaining a dataset of a size sufficient for automatic classification of high accuracy is challenging, particularly when sample collection is conducted during field campaigns [37]. Establishing a sufficiently large and balanced dataset for aquatic insects requires extensive sampling along various spatial and temporal gradients. This process can be time-consuming and particularly challenging when it comes to locating and collecting species with limited geographical distribution ranges. So far, only a few authors have explored the sufficient number of images per individual required to achieve a highly accurate classifier for the automatic identification of aquatic insects [25,38]. As expected, the classification accuracy of our model gradually increased with an increasing number of individuals used for training. The accuracy of our models increased from 97.09 $\pm$ 0.41% with the smallest training dataset to 99.70 $\pm$ 0.19% with the largest training dataset, which is similar to the results reported in other studies. Høye et al. [25], demonstrated a 97% classification accuracy when the CNN model was trained on a balanced set of 10 individuals from 16 aquatic macroinvertebrate taxa and a striking 99.2% classification accuracy when trained on 50 individuals from the same 16 taxa. In this study, the model was trained on species from seven morphologically and phylogenetically distinct taxonomic orders, including mollusks and crustaceans, which makes the classification task less challenging than one in our study. Although it is expected that the similarity among taxa could make classification more challenging for CNN algorithms, our analysis of performance on a balanced dataset found no evidence that morphologically similar and closely related species within the same genus (e.g., *Baetis rhodani* and *B. alpinus*) or family (e.g., *Silo pallipes* and *Lithax obscurus* from the family Glossosomatidae) were more difficult to classify than species with greater morphological differences. CNNs typically show improved performance with increased data volume, but beyond a certain saturation point, the rate of improvement can slow and eventually diminish [39]. However, the number of data considered 'sufficient' for optimal CNN performance depends on factors such as model complexity, task requirements, and data quality [25,38]. Accordingly, our findings suggest that a sample size of 30 individuals per species was optimal for achieving high classification precision. Training with a larger sample size introduced fluctuations

and even a slight decrease in model accuracy, indicating diminishing returns beyond this threshold in our case.

The frequency of species within freshwater macroinvertebrate communities is naturally unbalanced, with fewer dominant taxa and a higher number of rare ones [40]. At the same time, the minority class is usually more important from a data mining perspective, as despite its rarity, it may contain significant and useful information [41]. This imbalance is consequently reflected in the construction of training data for deep learning models. The overall performance of the CNN models after the decrease of the imbalance ratio increased from 97.09% to 99.88%. Various studies have observed that CNN models struggle to accurately identify classes with small and imbalanced training samples [22,42]. The findings of some authors indicate that sensitivity to class imbalance increases as problem complexity increases (i.e., the number of training variables and the degree of class imbalance) [22,43,44]. When we analyzed the performance on an imbalanced dataset, we found that the species with the poorest performance were those with the lowest number of images in the dataset (Models I and II, in which the smallest number of individuals for some species was 10), which is consistent with other observations [45]. We acknowledge the potential impact of class imbalance on classification accuracy within CNN models; while our findings showed a significant decrease in accuracy due to imbalanced samples, the loss was deemed acceptable and unlikely to compromise our primary objective of accurate taxa identification. In our study, we observed a reduction in accuracy, approximately 2%, due to imbalanced samples, a trend also noted by Thabtah et al. [44] across various domains, including medical diagnosis and credit card fraud detection, where minor classes contributed to a 2% accuracy drop. Similarly, in Lee et al.'s [46] study on plankton image classification, balancing the class distribution led to a 5% improvement in accuracy for minority classes. Despite this increase, the study's overall classification accuracy remained lower than ours at around 88%, reinforcing that a well-tuned model can achieve strong performance even with some degree of class imbalance. Given these findings, we believe that the current level of accuracy in our model is sufficiently high, and the minor error introduced by sample imbalance—estimated to be under 2%—is unlikely to detract from the reliability of our taxa classification outcomes. Krawczyk [41] suggests that the imbalance ratio itself may not be the primary source of learning challenges; rather, satisfactory results can be achieved if each class has sufficient representation. In experiments with artificially generated data of varying complexity and degrees of imbalance, some authors [47] demonstrate that when each class has at least 50 examples, even at high levels of concept complexity, the classification error remains below 1%, indicating that robust performance can be maintained with adequate representation across classes. Therefore, while we recognize the potential for imbalance-induced errors in CNN models, the 2% accuracy loss observed in our study indicates that taxa identification remains robust and effective, ensuring reliable classification outcomes even in the presence of a slight class imbalance.

When observing individual species, the common classification errors included misclassifying *L. obscurus* as *S. pallipes* and *T. major* as *E. mucronata*. Both cases resulted in confusion among species within the families Glossosomatidae and Ephemerellidae, respectively. This is not surprising because species that are morphologically similar and closely related (within the same genus or family) are considered more difficult to separate, as has been previously shown, not only for humans but also for deep learning models [17,25,45,48]. Although the misclassified species are closely related, we cannot conclude that morphological or phylogenetic similarity is the cause of misclassification. In fact, for some morphologically similar species within the same genera (*Baetis rhodani* and *B. alpinus*, *Rhyacophila tristis* and *R. fasciata*, *Ephemerella mucronata* and *E. ignita*), our models achieved a perfect classification rate of 100%.

Misclassification of individuals during routine species monitoring can lead to inaccurate calculations of multimetric index scores, which, according to the Water Framework Directive (WFD) regulations, are used across Europe to assess the health of aquatic ecosystems [49]. Closely related species within the same genus may exhibit different preferences for various water qualities [49]. For example, *B. rhodani* is a cosmopolitan species that inhabits waters of varying quality, while the occurrence of *B. alpinus* in some ecosystems is more indicative of high-quality waters. The presence of *S. pallipes* indicates oligotrophic waters, while *L. obscurus* is a bioindicator of oligotrophic to beta-mesosaprobic waters [50]. This emphasizes that reliable species identification is a crucial step in assessing the health of water ecosystems and recognizing areas of high conservation priority.

Synthetic data generation like data augmentation, including transformations such as rotation, flipping, scaling, and cropping, can significantly increase the number of images in a dataset for CNN training, enhancing model generalization without additional data collection [51,52]. Our study demonstrates that, with the exception of Models I and II, species in all models were represented by a sufficient number of data. Class imbalance had a significant impact on the performance of Models I and II, leading to an approximate 2% reduction in classification accuracy. By applying the oversampling technique, we successfully mitigated the errors caused by imbalanced samples, almost completely restoring the species classification accuracy to 98.97 $\pm$ 0.31% and 98.85 $\pm$ 0.23%, respectively. This technique, which duplicates random examples from the minority class to balance the dataset, is widely supported in the literature as an effective approach to improving model performance [53,54]. Therefore, the oversampling technique offers a viable solution to address the challenges associated with detecting rare taxa in automated species monitoring.

In conclusion, building a robust model with acceptable accuracy for automated species identification largely depends on having a sufficient sample size in the training dataset, with at least 30 images per class. The impact of sample imbalance on the classification accuracy of CNN models in our study was observed as a 2% reduction; however, this level remains within an acceptable range for achieving the primary goal of sufficient taxonomic resolution in routine biomonitoring of surface waters, with the adverse effects further mitigated through the application of oversampling techniques. Our findings demonstrate the feasibility of automated classification for less common species, but further research is needed to understand the effects of training convolutional neural network (CNN) models on a broader range of species, particularly with high-class imbalance and increased diversity among closely related species.

The findings of this research provide valuable insights into determining the optimal amount of information needed to develop an effective deep-learning model capable of automatically identifying species essential for biomonitoring, biodiversity, and conservation studies at high taxonomic resolution. This research supports the organization of reference datasets to improve existing models, enhance their quality, and streamline the application of deep learning techniques for accurate species identification. These approaches have the potential to tackle critical biomonitoring challenges by significantly reducing both time and effort.

Author Contributions: Conceptualization, P.S., D.M. and A.M.; Methodology, P.S., D.M. and A.M.; Software, A.M.; Validation, P.S., A.M. and D.M.; Formal Analysis, P.S.; Investigation, P.S.; Resources, P.S., K.S., A.M., D.S.-Z., B.P., A.P. and D.M.; Data Curation, P.S. and K.S.; Writing—Original Draft Preparation, P.S.; Writing—Review & Editing, P.S. and D.M.; Visualization, P.S.; Supervision, D.M.; Project Administration, D.S.-Z. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by the Science Fund of the Republic of Serbia (Grant #7751676) for the project 'Application of Deep Learning in Bioassessment of Aquatic Ecosystems: Toward the

Construction of an Automatic Identifier for Aquatic Macroinvertebrates—AIAQUAMI'. Additional support was provided by the Ministry of Science, Technological Development and Innovation of the Republic of Serbia under Grant No. 451-03-66/2024-03/200122.

Data Availability Statement: The training results, including the images used in this research, are available for download in the Releases section of the GitHub project: https://github.com/a-milosavljevic/aiaquami-ept33/releases (accessed on 12 December 2024).

Conflicts of Interest: The authors declare no conflicts of interest. The funders were not involved in the study design, data collection, analysis, interpretation, manuscript writing, or the decision to publish the results.

Appendix A

Classification errors and their distribution among species for each model using balanced datasets of 10 to 80 individuals per EPT taxon, across five replicates.

Appendix B

Classification errors and their distribution among species for each model using imbalanced datasets, both with and without applying the oversampling technique for the minority class in the dataset.

References

1. Besson, M.; Alison, J.; Bjerge, K.; Gorochowski, T.E.; Høye, T.T.; Jucker, T.; Mann, H.M.R.; Clements, C.F. Towards the fully automated monitoring of ecological communities. *Ecol. Lett.* **2022**, *25*, 2753–2775. [CrossRef] [PubMed]
2. Riabchenko, E.; Meissner, K.; Ahmad, I.; Iosifidis, A.; Tirronen, V.; Gabbouj, M.; Kiranyaz, S. Learned vs. engineered features for fine-grained classification of aquatic macroinvertebrates. In Proceedings of the 2016 23rd International Conference on Pattern Recognition (ICPR), Cancun, Mexico, 4–8 December 2016; pp. 2276–2281. [CrossRef]
3. Batzias, F.A.; Siontorou, C.G.A. Knowledge-Based Approach to Environmental Biomonitoring. *Environ. Monit. Assess.* **2006**, *123*, 167–197. [CrossRef] [PubMed]
4. Amaral, P.H.M.d.; Silveira, L.S.d.; Rosa, B.F.J.V.; Oliveira, V.C.d.; Alves, R.d.G. Influence of Habitat and Land Use on the Assemblages of Ephemeroptera, Plecoptera, and Trichoptera in Neotropical Streams. *J. Insect Sci.* **2015**, *15*, 60. [CrossRef] [PubMed]
5. Simović, P.; Milošević, D.; Simić, V.; Stojanović, K.; Atanacković, A.; Jakovljević, A.; Petrović, A. Benthic macroinvertebrates in a tufa-depositing environment: A case study of highly vulnerable karst lotic habitats in Southeast Europe. *Hydrobiologia* **2024**, *851*, 4761–4779. [CrossRef]
6. Haase, P.S.; Pauls, K.; Schindehu, T.T.E.; Sundermann, A. First audit of macroinvertebrate samples from an EU Water Framework Directive monitoring program: Human error greatly lowers precision of assessment results. *J. N. Am. Benthol. Soc.* **2010**, *29*, 1279–1291. [CrossRef]
7. Zhou, X.; Jacobus, L.M.; DeWalt, R.E.; Adamowicz, S.J.; Hebert, P.D.N. Ephemeroptera, Plecoptera, and Trichoptera fauna of Churchill (Manitoba, Canada): Insights into biodiversity patterns from DNA barcoding. *J. N. Am. Benthol.* **2010**, *29*, 814–837. [CrossRef]
8. Suh, K.I.; Hwang, J.M.; Bae, Y.J.; Kang, J.H. Comprehensive DNA barcodes for species identification and discovery of cryptic diversity in mayfly larvae from South Korea: Implications for freshwater ecosystem biomonitoring. *Entomol. Res.* **2019**, *49*, 46–54. [CrossRef]
9. Villon, S.; Iovan, C.; Mangeas, M.; Vigliola, L. Confronting Deep-Learning and Biodiversity Challenges for Automatic Video-Monitoring. *Sensors* **2022**, *10*, 497. [CrossRef]
10. Chang, G.J. Biodiversity estimation by environment drivers using machine/deep learning for ecological management. *Ecol. Inform.* **2023**, *78*, 102319. [CrossRef]
11. LeCun, Y.; Bengio, Y.; Hinton, G. Deep learning. *Nature* **2015**, *521*, 436–444. [CrossRef]
12. Chandrakar, R.; Raja, R.; Miri, R.; Tandan, S.R.; Ramya Laxmi, K. Detection and Identification of Animals in Wildlife Sanctuaries using Convolutional Neural Network. *Int. J. Recent Technol. Eng.* **2020**, *8*, 2277–3878. [CrossRef]
13. Badre, P.; Chandanshive, P.; Bandiwadekar, S.; Chaudhari, A.; Jadhav, S. Automatically Identifying Animals Using Deep Learning. *Int. J. Recent Innov. Trends Comput. Commun.* **2018**, *6*, 194–197.
14. Rauf, H.T.; Lali, M.I.U.; Zahoor, S.; Shah, S.Z.H.; Rehman, A.U.; Bukhari, S.A.C. Visual features based automated identification of fish species using deep convolutional neural networks. *Comput. Electron. Agric.* **2019**, *167*, 105075. [CrossRef]

15. Larios, N.; Lin, J.; Zhang, M.; Moldenke, A.; Shapiro, L.; Dietterich, T. Stacked spatial-pyramid kernel: An object-class recognition method to combine scores from random trees. In Proceedings of the 2011 IEEE Workshop on Applications of Computer Vision (WACV), Kona, HI, USA, 5–7 January 2011; pp. 329–335. [CrossRef]
16. Raitoharju, J.; Riabchenko, E.; Ahmad, I.; Iosifidis, A.; Gabbouj, M.; Kiranyaz, S.; Tirronen, V.; Ärje, J.; Kärkkäinen, S.; Meissner, K. Benchmark database for fine-grained image classification of benthic macroinvertebrates. *Image Vis. Comput.* **2018**, *78*, 73–83. [CrossRef]
17. Simović, P.; Milosavljević, A.; Stojanović, K.; Radenković, M.; Predić, B.; Božanić, M.; Petrović, A.; Milošević, D. Automated identification of aquatic insects: A case study using deep learning and computer vision techniques. *Sci. Total Environ.* **2024**, *935*, 172877. [CrossRef]
18. Jaballah, S.; Fernandez Garcia, S.; Martignac, F.; Parisey, N.; Jumel, S.; Roussel, J.-M.; Dézerald, O. A deep learning approach to detect and identify live freshwater macroinvertebrates. *Aquat. Ecol.* **2023**, *57*, 933–949. [CrossRef]
19. Johnson, J.M.; Khoshgoftaar, T.M. Survey on deep learning with class imbalance. *J. Big Data* **2019**, *6*, 29. [CrossRef]
20. Ghosh, K.; Bellinger, C.; Corizzo, R.; Branco, P.; Krawczyk, B.; Japkowicz, N. The class imbalance problem in deep learning. *Mach. Learn.* **2024**, *113*, 4845–4901. [CrossRef]
21. Li, J.; Fong, S.; Mohammed, S.; Fiaidhi, J. Improving the classification performance of biological imbalanced datasets by swarm optimization algorithms. *J. Supercomput.* **2016**, *72*, 3708–3728. [CrossRef]
22. Sotiropoulos, D.N.; Tsihrintzis, G.A. The Class Imbalance Problem. In *Machine Learning Paradigms*; Springer: Cham, Switzerland, 2016; Volume 118, pp. 51–78. [CrossRef]
23. Wach, M.; Guéguen, J.; Christian, C.; Delmas, F.; Dagens, N.; Feret, T.; Loriot, S.; Tison-Rosebery, J. Probability of misclassifying river ecological status: A large-scale approach to assign uncertainty in macrophyte and diatom-based biomonitoring. *Ecol. Indic.* **2019**, *101*, 285–295. [CrossRef]
24. Høye, T.T.; Ärje, J.; Bjerge, K.; Hansen, O.L.P.; Iosifidis, A.; Leese, F.; Mann, H.M.R.; Meissner, K.; Melvad, C.; Raitoharju, J. Deep learning and computer vision will transform entomology. *Proc. Natl. Acad. Sci. USA* **2021**, *118*, e2002545117. [CrossRef] [PubMed]
25. Høye, T.T.; Dyrmann, M.; Kjær, C.; Nielsen, J.; Bruus, M.; Mielec, C.L.; Vesterdal, M.S.; Bjerge, K.; Madsen, S.A.; Jeppesen, M.R.; et al. Accurate image-based identification of macroinvertebrate specimens using deep learning—How much training data is needed? *PeerJ* **2022**, *10*, e13837. [CrossRef] [PubMed]
26. Kiranyaz, S.; Ince, T.; Pulkkinen, J.; Gabbouj, M.; Ärje, J.; Kärkkäinen, S.; Tirronen, V.; Juhola, M.; Turpeinen, T.; Meissner, K. Classification and retrieval on macroinvertebrate image databases. *Comput. Biol. Med.* **2011**, *41*, 463–472. [CrossRef] [PubMed]
27. Durden, J.M.; Hosking, B.; Bett, B.J.; Cline, D.; Ruhl, H.A. Automated classification of fauna in seabed photographs: The impact of training and validation dataset size, with considerations for the class imbalance. *Progr. Oceanogr.* **2021**, *196*, 102612. [CrossRef]
28. Simić, V.; Bănăduc, D.; Curtean-Bănăduc, A.; Petrović, A.; Veličković, T.; Stojković-Piperac, M.; Simić, S. Assessment of the ecological sustainability of river basins based on the modified theESHIPPOfish model on the example of the Velika Morava basin (Serbia, Central Balkans). *Front. Environ. Sci.* **2022**, *10*, 952692. [CrossRef]
29. Aubert, J. Plecoptera. In *Insecta Helvetica, Fauna*; Imprimerie La Concorde: Lausanne, Switzerland, 1959; Volume 1, pp. 91–136.
30. Eiseler, B. Identification key to the mayfly larvae of the German Highlands und Lowlands. *Lauterbornia* **2005**, *53*, 1–112.
31. Rozkošný, R. *Klíč Vodnich Lareu Hmyzu*; Academia Nakladatelstvi Československé Akademie Véd: Praha, Czech Republic, 1959.
32. Waringer, J.; Graf, W. *Atlas der Mitteleuropäischen Köcherfiegenlarven: Atlas of Central-European Trichoptera Larvae*; Erik Mauch Verlag: Dinkelscher, Germany, 2011.
33. Tan, M.; Le, Q. EfficientNet: Rethinking Model Scaling for Convolutional Neural Networks. In Proceedings of the 36th International Conference on Machine Learning, PMLR, Long Beach, CA, USA, 9–15 June 2019; Volume 97, pp. 6105–6114.
34. R Core Team. *R: A Language and Environment for Statistical Computing*; R Foundation for Statistical Computing: Vienna, Austria, 2023. Available online: https://www.R-project.org/ (accessed on 10 December 2024).
35. Fernandes, J.A.; Irigoien, X.; Boyra, G.; Lozano, J.A.; Inza, I. Optimizing the number of classes in automated zooplankton classification. *J. Plankton Res.* **2009**, *31*, 19–29. [CrossRef]
36. Benkendorf, D.J.; Hawkins, C.P. Effects of sample size and network depth on a deep learning approach to species distribution modeling. *Ecol. Inform.* **2020**, *60*, 101137. [CrossRef]
37. Ramezan, C.A.; Warner, T.A.; Maxwell, A.E.; Price, B.S. Effects of Training Set Size on Supervised Machine-Learning Land-Cover Classification of Large-Area High-Resolution Remotely Sensed Data. *Remote Sens.* **2021**, *13*, 368. [CrossRef]
38. Ärje, J.; Melvad, C.; Jeppesen, M.R.; Madsen, S.A.; Raitoharju, J.; Meissner, K.; Rasmussen, M.S.; Iosifidis, A.; Tirronen, V.; Gabbouj, M.; et al. Automatic image-based identification and biomass estimation of invertebrates. *Methods Ecol. Evol.* **2020**, *11*, 922–931. [CrossRef]
39. Davidian, M.; Lahav, A.; Joshua, B.-Z.; Wand, O.; Lurie, Y.; Mark, S. Exploring the Interplay of Dataset Size and Imbalance on CNN Performance in Healthcare: Using X-rays to Identify COVID-19 Patients. *Diagnostics* **2024**, *14*, 1727. [CrossRef] [PubMed]
40. Magurran, A.E. Measuring biological diversity. *Curr. Biol.* **2021**, *31*, R1174–R1177. [CrossRef] [PubMed]

41. Krawczyk, B. Learning from imbalanced data: Open challenges and future directions. *Prog. Artif. Intell.* **2016**, *5*, 221–232. [CrossRef]
42. Buda, M.; Maki, A.; Mazurowski, M.A. A systematic study of the class imbalance problem in convolutional neural networks. *Neural Netw.* **2018**, *106*, 249–259. [CrossRef]
43. Japkowicz, N. The class imbalance problem: Significance and strategies. In Proceedings of the 2000 International Conference on Artificial Intelligence (ICAI), Vancouver, BC, Canada, 13–15 November 2000; pp. 111–117.
44. Thabtah, F.; Hammoud, S.; Kamalov, F.; Gonsalves, A. Data imbalance in classification: Experimental evaluation. *Inf. Sci.* **2020**, *513*, 429–441. [CrossRef]
45. Valan, M.; Makonyi, K.; Maki, A.; Vondráček, D.; Ronquist, F. Automated Taxonomic Identification of Insects with Expert-Level Accuracy Using Effective Feature Transfer from Convolutional Networks. *Syst. Biol.* **2019**, *68*, 876–895. [CrossRef]
46. Lee, H.; Park, M.; Kim, J. Plankton classification on imbalanced large-scale database via convolutional neural networks with transfer learning. In Proceedings of the 2016 IEEE International Conference on Image Processing (ICIP), Phoenix, AZ, USA, 25–28 September 2016; pp. 3713–3717. [CrossRef]
47. An, A.; Cercone, N.; Huang, X. Case Study for Learning from Imbalanced Data. In *Advances in Artificial Intelligence: 14th Biennial Conference of the Canadian Society for Computational Studies of Intelligence, AI 2001 Ottawa, Canada, June 7–9, 2001 Proceedings*; Stroulia, E., Matwin, S., Eds.; Springer: Berlin/Heidelberg, Germany, 2001. [CrossRef]
48. Larios, N.; Deng, H.; Zhang, W.; Sarpola, M.; Yuen, J.; Paasch, R.; Moldenke, A.; Lytle, D.A.; Ruiz Correa, S.; Mortensen, E.N.; et al. Automated insect identification through concatenated histograms of local appearance features: Feature vector generation and region detection for deformable objects. *Mach. Vis. Appl.* **2008**, *19*, 105–123. [CrossRef]
49. *Directive 2000/60/EC of the European Parliament and of the Council of 23 October 2000 Establishing a Framework for Community Action in the Field of Water Policy*; European Union: Brussels, Belgium, 2000; pp. 1–73.
50. Sládeček, V. System of water quality from the biological point of view. *Arch. Hydrobiol.* **1973**, *7*, 1–218.
51. Han, D.; Liu, Q.; Fan, W. A new image classification method using CNN transfer learning and web data augmentation. *Expert Syst. Appl.* **2018**, *95*, 43–56. [CrossRef]
52. Talukdar, J.; Biswas, A.; Gupta, S. Data Augmentation on Synthetic Images for Transfer Learning using Deep CNNs. In Proceedings of the 5th International Conference on Signal Processing and Integrated Networks (SPIN), Noida, India, 22–23 February 2018; pp. 215–219. [CrossRef]
53. Yang, Z.; Zhu, Y.; Liu, T.; Zhao, S.; Wang, Y.; Tao, D. Output Layer Multiplication for Class Imbalance Problem in Convolutional Neural Networks. *Neural Process. Lett.* **2020**, *52*, 2637–2653. [CrossRef]
54. Dablain, D.; Jacobson, K.N.; Bellinger, C.; Roberts, M.; Chawla, N. Understanding CNN fragility when learning with imbalanced data. *Mach. Learn.* **2024**, *113*, 4785–4810. [CrossRef]

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.