Applying convolutional neural networks to speed up environmental DNA annotation in a highly diverse ecosystem

Benjamin Flück, Laëtitia Mathon, Stéphanie Manel, Alice Valentini, Tony Dejean, Camille Albouy, David Mouillot, Wilfried Thuiller, Jérôme Murienne, Sébastien Brosse, et al.

scientific reports

Check for updates

OPEN

Applying convolutional neural networks to speed up environmental DNA annotation in a highly diverse ecosystem

Benjamin Flück[1,2]✉, Laëtitia Mathon[3], Stéphanie Manel[3], Alice Valentini[4], Tony Dejean[4], Camille Albouy[5], David Mouillot[6,7], Wilfried Thuiller[8], Jérôme Murienne[9], Sébastien Brosse[9] & Loïc Pellissier[1,2]✉

High-throughput DNA sequencing is becoming an increasingly important tool to monitor and better understand biodiversity responses to environmental changes in a standardized and reproducible way. Environmental DNA (eDNA) from organisms can be captured in ecosystem samples and sequenced using metabarcoding, but processing large volumes of eDNA data and annotating sequences to recognized taxa remains computationally expensive. Speed and accuracy are two major bottlenecks in this critical step. Here, we evaluated the ability of convolutional neural networks (CNNs) to process short eDNA sequences and associate them with taxonomic labels. Using a unique eDNA data set collected in highly diverse Tropical South America, we compared the speed and accuracy of CNNs with that of a well-known bioinformatic pipeline (OBITools) in processing a small region (60 bp) of the 12S ribosomal DNA targeting freshwater fishes. We found that the taxonomic labels from the CNNs were comparable to those from OBITools, with high correlation levels for the composition of the regional fish fauna. The CNNs enabled the processing of raw fastq files at a rate of approximately 1 million sequences per minute, which was about 150 times faster than with OBITools. Given the good performance of CNNs in the highly diverse ecosystem considered here, the development of more elaborate CNNs promises fast deployment for future biodiversity inventories using eDNA.

Effective ecosystem governance and management require an increase in speed, accuracy and ease of collecting and processing of biodiversity data[31,49]. Biodiversity data collection requires a shift in focus from expert monitoring towards high-throughput data acquisition technology[24]. Conventional biodiversity monitoring approaches are labor intensive, depend on expert knowledge—resulting in long delays between sampling and results[53], and miss many species that are either small, rare, cryptic or elusive[41], which in turn hinders accurate ecological interpretations. Fortunately, our ability to rapidly generate inventories of whole species communities is improving with the emergence of environmental genomics, specifically environmental DNA (eDNA)[6,25,27,75]. All organisms living in an ecosystem shed tissue material, which can be detected through eDNA metabarcoding[74], offering an integrative view of the ecosystem composition[27,33]. Coupled with high-throughput DNA sequencing methods, eDNA metabarcoding can help with the rapid assessment and monitoring of biodiversity across all levels of life, from prokaryotes to eukaryotes[40], with a higher detection capacity and cost-effectiveness than traditional methods[59]. The reads from high-throughput amplicon sequencing of eDNA can be compared with reference barcode libraries, enabling the establishment of taxonomic lists directly from environment samples[74]. Ultimately, these lists can be used to assess ecosystem functioning and health status[25]. With an increasing number of initiatives proposing

[1]Department of Environmental System Science, ETH Zürich, 8092 Zurich, Switzerland. [2]Swiss Federal Research Institute WSL, 8903 Birmensdorf, Switzerland. [3]CEFE, Univ. Montpellier, CNRS, EPHE-PSL University, IRD, Montpellier, France. [4]SPYGEN, Le Bourget-du-Lac, France. [5]DECOD (Ecosystem Dynamics and Sustainability), IFREMER, INRAE, Institut Agro - Agrocampus Ouest, Rue de l'Ile d'Yeu, BP21105, 44311 Nantes Cedex 3, France. [6]MARBEC, Univ. Montpellier,CNRS, IRD, Ifremer, Montpellier, France. [7]Institut Universitaire de France, IUF, 75231 Paris, France. [8]CNRS, LECA, Laboratoire d'Écologie Alpine, Univ. Grenoble Alpes, Univ. Savoie Mont Blanc, 38000 Grenoble, France. [9]Laboratoire Evolution et Diversité Biologique (UMR5174), CNRS, IRD, Université Paul Sabatier, Toulouse, France. ✉email: benjamin.flueck@usys.ethz.ch; loic.pellissier@usys.ethz.ch

the use of eDNA metabarcoding routinely and globally to monitor ecosystems[5], the processing of such massive sequencing data will require novel automated bioinformatic solutions that are both fast and accurate.

As the laboratory molecular steps of eDNA metabarcoding have gained in efficiency[69,75], the major bottleneck and technical challenge has shifted from the development of efficient sampling and laboratory protocols to the processing of the produced large set of raw sequencing data into taxonomic lists[32]. In particular, eDNA metabarcoding amplifies small DNA sequences ('barcodes'), typically 60–300 bp long, from the mitochondrial genome for use with Illumina sequencing technology[71]. This sequencing process generates a huge quantity of small sequence reads that require fast and accurate bioinformatic processing to be interpreted[34,35]. The bioinformatic processing includes several steps (merging the forward and reverse reads, demultiplexing, dereplicating, filtering by quality, removing errors), after which the retained and cleaned sequences are assigned to a taxonomic label[32,50,56]. Taxonomic labelling then involves transforming sequence reads from eDNA into lists of taxa that can be used by experts and scientists to understand biodiversity patterns, structures and dynamics of assemblages. They can additionally be used for management decisions[68], based on the detection of rare[9,63], endangered[37], or invasive species[68]. Given that most existing pipelines are time consuming to apply[52], efficient algorithms transforming eDNA reads into accurate taxonomic lists using machine learning could potentially enable efficient and parallel automatization on cloud infrastructure for a broad application of eDNA technology[65].

Compared with traditional bioinformatic approaches[52], machine learning could increase the efficiency and capacity of the taxonomic labelling of eDNA reads[55]. Deep learning has revolutionized object classifications in various biological applications, from identifying species on images[38] to modelling species distributions in habitats[28]. Taxonomic groups represent discrete classes that can be related to sequence features, including the composition and distribution of nucleobases within DNA sequences[14,39]. For example, k-mer summarizes the counts of nucleotides within sub-sequences of length k and, in combination with machine classifications, have been used to label sequences from bacteria, archaea, fungi and viruses[58]. The association between k-mer features and taxonomic labelling can be trained in a neural network from a reference genetic database[58,60] to predict the label of any new sequence. Alternatively, a convolutional neural network (CNN) can self-learn a broader range of spatially organized DNA base-motif features existing in the DNA sequences[39]. The neural structure subsets signals from a restricted region of the input data known as the receptive field and responds to localized patterns in the sequence data. The numeric encoding of the four DNA bases makes it possible for the spatial placements of nucleotides to be interpreted by the CNN. In particular, Busia et al. developed a CNN[14] which trains a deep neural network to predict database-derived taxonomic labels directly from query sequences. Hence, preliminary use of machine learning with DNA sequence data shows the potential of this approach for taxonomic labelling[14,44], but so far it has mainly been used to label relatively long amplicons such as the full 16S gene, in fragments up to 250 bp long[14]. It remains to be determined how it performs in the taxonomic labelling of short sequences from eDNA metabarcoding.

The most computationally costly step in the processing of eDNA metabarcoding is data cleaning[52], and a large computational gain from machine learning could be achieved if a CNN can be applied directly on raw sequencing data that can contain many errors, including PCR substitutions or insertions or deletions of bases[66,67,73]. Existing eDNA bioinformatic pipelines apply a computationally demanding process of sequence processing and cleaning[52], conserving only high-quality reads[10] before the taxonomic labelling of DNA reads. To circumvent this data cleaning procedure, CNNs should be able to either identify low-quality sequences or accommodate noisy data in the taxonomic labelling. CNNs with data augmentation have been used to render networks more robust to noisy data, for example by adding random variation in the training data[70]. Busia et al.[14] artificially introduced variation into sequences within the reference database to build a more robust CNN, adding between 0.5 and 16% of mutations by switching DNA bases randomly[14]. While the authors found that moderate artificial noise rendered the network more robust to potential sequencing errors, setting an excessive value decreased the CNN performance. Furthermore, the CNN should be trained to tolerate the library tags and the PCR primers present in raw metabarcoding data, but these aspects have remained largely unexplored. The CNN could then be used to process and identify the sequences from raw metabarcoding files, independently of the processing step in which they are demultiplexed to each sample. If reliable, a CNN pipeline serves as a revolutionary tool to process the exponentially growing quantity of eDNA metabarcoding data used to characterize ecosystems.

Here, we used a comprehensive eDNA data set collected in tropical South America to evaluate the ability of CNNs to rapidly and accurately process eDNA metabarcoding files into taxonomic labels. We built CNNs that allow the processing of short sequences produced by eDNA metabarcoding and tested whether the accuracy and speed of CNNs are comparable to those of OBITools[10], a widely used pipeline to process eDNA data. As a case study we used one of the largest standardized eDNA data sets currently available for fishes, corresponding to a multi-year campaign effort to sample the tropical South American rivers of French Guiana[54] (Fig. 1). This eDNA data set is associated with a quasi-exhaustive reference database covering most of the known species of the region for the 'teleo' region of the 12S rRNA mitochondrial gene[22,26]. The raw data set contains nearly 700 million sequences, with about 205 million sequences belong to the samples of interest here. The freshwater ecosystems of French Guiana are among the most species-rich ecosystems for riverine fishes globally[3], and among the rivers the least impacted by humans[72]. Good performance of an approach in a complex ecosystem provides a robust proof of concept for further applications in any other ecosystem with a simpler species assemblage. Within this general processing framework and using this case study, we asked the following questions: (1) How does a CNN approach perform in the training of eDNA sequence classification for labels of the reference database? (2) How robust is the classification of a CNN applied directly to raw Illumina metabarcoding short sequences? (3) How do a classical metabarcoding pipeline and our CNN approach compare with the pre-existing information about biodiversity composition within two river catchments with a long history of traditional sampling?

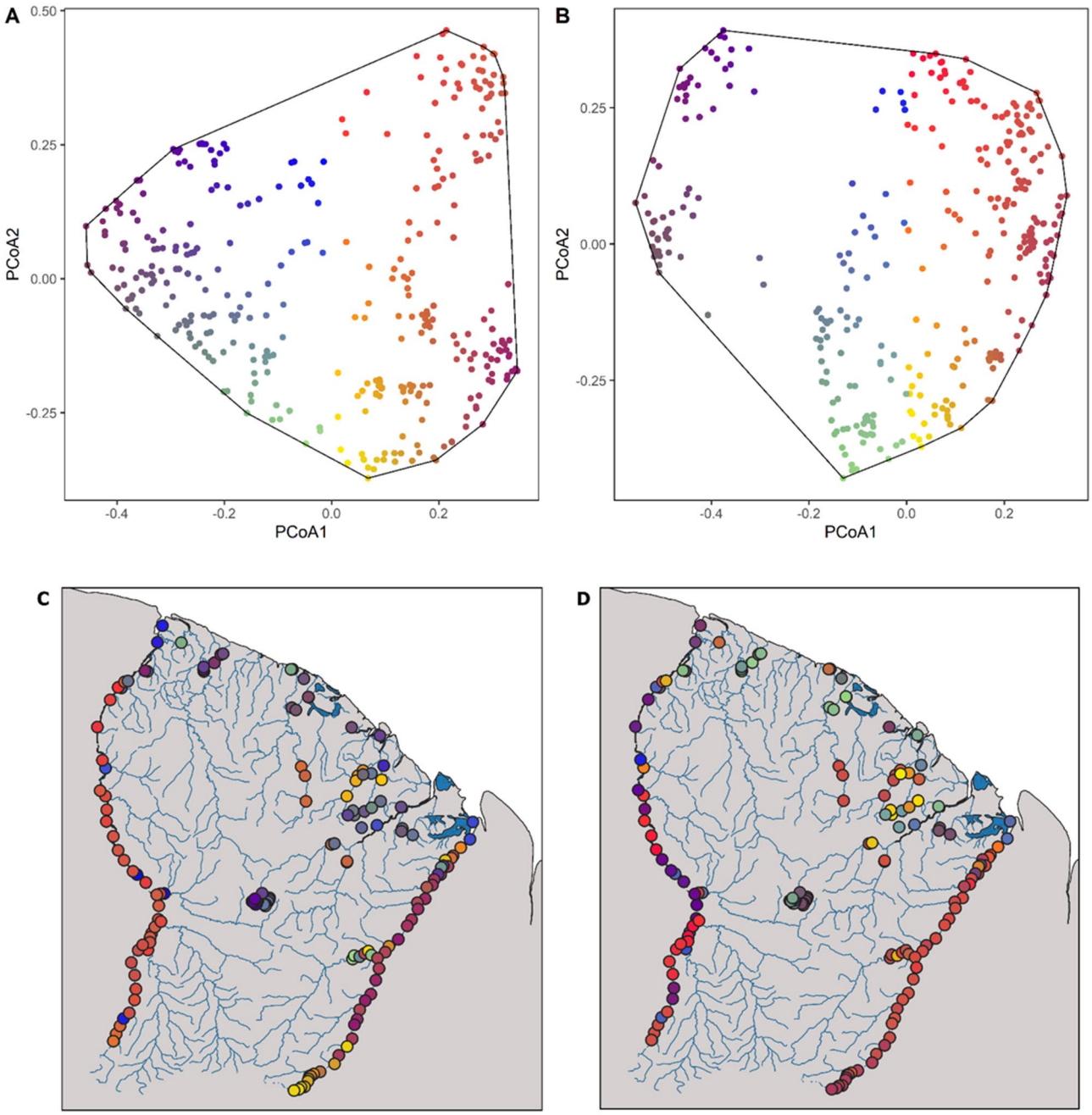

Figure 1. Principal coordinate analysis (PCoA) of species composition dissimilarity between filters. (**A**) Ordination of filter species composition dissimilarity in the outputs of OBITools. (**B**) Ordination of filter species composition dissimilarity in the outputs of the CNN applied to raw reads. Dissimilarity matrices were built with Bray–Curtis distances on read abundance per species per filter. (**C**) Maps of the filter locations, coloured according to the position of the filters in the PCoA space for OBITools outputs. (**D**) Maps of the filter locations, coloured according to the position of the filters in the PCoA space for the CNN applied to raw reads outputs. The maps were created with QGIS 3.6.1.

Results

CNN training and evaluation with split sampling. CNNs learned features of the 60 bp teleo sequence reads with good internal and external predictive power. Larger networks did not necessarily produce better results, indicating low overfitting. A CNN of moderate complexity learned the full structure contained in the training sequence data. The training and evaluation of the CNN with split sampling considered 156 species (out of 368) which had at least two unique sequences. The optimal CNN consisted of a 150×4 unit input layer, one convolutional layer of 4 filters with a 7×4 extent, 3 dense layers with 128 neurons each, and an output layer 156 neurons wide. On the training data, the networks achieved 92% accuracy, with small differences between the networks trained on the base reference data and those trained on the augmented reference data (i.e. with added tags, primers and reverse complements). When applying the CNN to the hold-out data (316 sequences

from the 156 species), we found an accuracy of 91% on the base data and 89% on the augmented data. When an optimized 0.9 binarization threshold was used with the F-beta metric, the accuracy rose to 98% for both CNNs, at the cost of 16–26% of the predictions being discarded for the base and augmented data, respectively. We then used the entire data set in the training process, using all 368 species, and repeated the analyses for the base and augmented data. The optimal CNN was similar to the previously chosen networks, with a single convolutional layer of 4 filters with a 7×4 extent, followed by 2 dense layers each 384 neurons wide. With these networks, training accuracy was similar to that from the split evaluation at 92%. Validating the networks on the reference sequences yielded higher accuracies of 96% and 94% for the base and augmented CNNs, respectively. With a binarization threshold of 0.9, the accuracy rose to 99% for both the base and augmented data sets, at the cost of rejecting 9–13% of all sequences evaluated (Supplementary Material Fig. 1). We used a binarization threshold of 0.9 for all further evaluations.

CNN application on the raw and cleaned eDNA data set. We found that there were limited differences in the output between the CNNs trained on the raw sequence data compared to those trained on cleaned data. To attenuate sequencing noise in the analysis, we considered a second threshold of the minimum number of reads required for a species to be retained. We compared the number of reads per species needed for each CNN with that needed for OBITools and observed that the median Kendall Tau-b correlation increased when a more stringent threshold on the minimum number of reads per species was applied to all levels of sample aggregation. An optimal threshold of 50 reads per species resulted a slightly better correlation for clean (median Kendall Tau-b = 0.77, range 0.22–0.94) than for raw reads (median Kendall Tau-b = 0.84, range = 0.2–1, Fig. 2) at the filter level. The same effect persisted on the PCR replicate and river levels. We considered only species with more than 50 reads within a PCR replicate in the following analyses. We repeated the analysis using the kappa similarity measurement (Fig. 3). The CNN applied to the clean reads (after assembling and demultiplexing) had a slightly higher composition similarity (median kappa value 0.96, range 0.83–1.0) than that applied to the raw reads directly from the Illumina outputs (median kappa value 0.93, range 0.79–0.99). The kappa values are based on the predicted presence and absence of species. Hence, the results were slightly better than those from the Kendall Tau b values, as those take the relative abundance of the predictions into account. All approaches recovered similar gradients of composition, differentiating between coastal and upstream assemblages (Fig. 1). The composition difference between methods resulted from a slightly larger number of species predicted by the CNN (median species number 63) than by OBITools (median species number 56). Furthermore, the CNNs still lacked feature parity with OBITools with regard to ambiguous sequences, which can result in more pronounced differences in the OBITools output.

Validation with the known species list of the region. We found a major overlap between historical records and the species composition recovered from the CNN. The data synthesis across historical fish surveys yielded a total of 351 species in the Maroni and Oyapock rivers, 293 of which were present in the reference database and thus potentially detectable with eDNA. For both rivers combined, the CNN applied to raw reads assigned 319 species, 264 of which were known from the historical records, while 55 had never been recorded before (Fig. 4a). The CNN and OBITools detected 274 species in common, while the CNN retrieved 21 species known from the historical surveys in these rivers that were not retrieved with OBITools but identified 24 species not known from the survey synthesis or identified with OBITools. The species detected only with the CNN mainly belong to the Loricariidae, Cichlidae, Characidae and Callichthyidae families. The 23 species known from historical records and not detected by either eDNA method mainly belong to the Loricariidae, Characidae, Apteronotidae and Anostomidae families. The two species detected only with OBItools are from the Cichlidae and Aspredinidae families (Fig. 4b). The CNN applied to clean reads detected 293 species, 254 of which were present in the Maroni and Oyapock synthesis, 276 of which were also found in the outputs of OBITools, 9 of which were found only with the CNN and in the synthesis, and 8 of which were found only with the CNN. In the case of OBITools, 282 species were detected, 249 of which were included in the historical synthesis and 33 of which had never been recorded in the Maroni or Oyapock rivers (Fig. 4c). The species detected only with the CNN mainly belong to the Characidae family. The species known from historical records but not detected with either eDNA method belong to the Loricariidae, Characidae and Apteronotidae families. The two species detected only with OBITools are from the Loricariidae and Cichlidae families (Fig. 4d). The same analysis at the single river scale provided similar results (Supplementary Material Figs 4, 5). Hence, while both methods detected species not found in the historical records, the CNN generally recovered more species than OBITools, which could correspond to either new true observations or commission errors. The CNN applied to raw reads retrieved more species that were not in historical records nor found with OBITools. For the Maroni river, the CNN applied to raw reads and the CNN applied to clean reads retrieved 232 species in common, while 48 were found only with the raw reads and 16 only with the clean reads. For the Oyapock river, 185, 66 and 18 species were found in common, only with the raw reads, and only with the clean reads, respectively (Supplementary Material Fig. 6).

Computation time. Overall, the CNN processed approximately 1 million input sequences per minute, compared with 20,000 input sequences per minute for OBITools. For the CNN, we distinguished between two computational efforts, which were measured independently: (1) network training, which needed to be performed once per reference database, and (2) the application on field data. Training a network on the augmented and complete reference database currently took around 10 min on an Nvidia Titan RTX GPU. Training a network on the clean reference database was faster and takes 6 min on the same GPU. The training and application time is dependent on the size of the input data and the network size. A large part of the computational time for

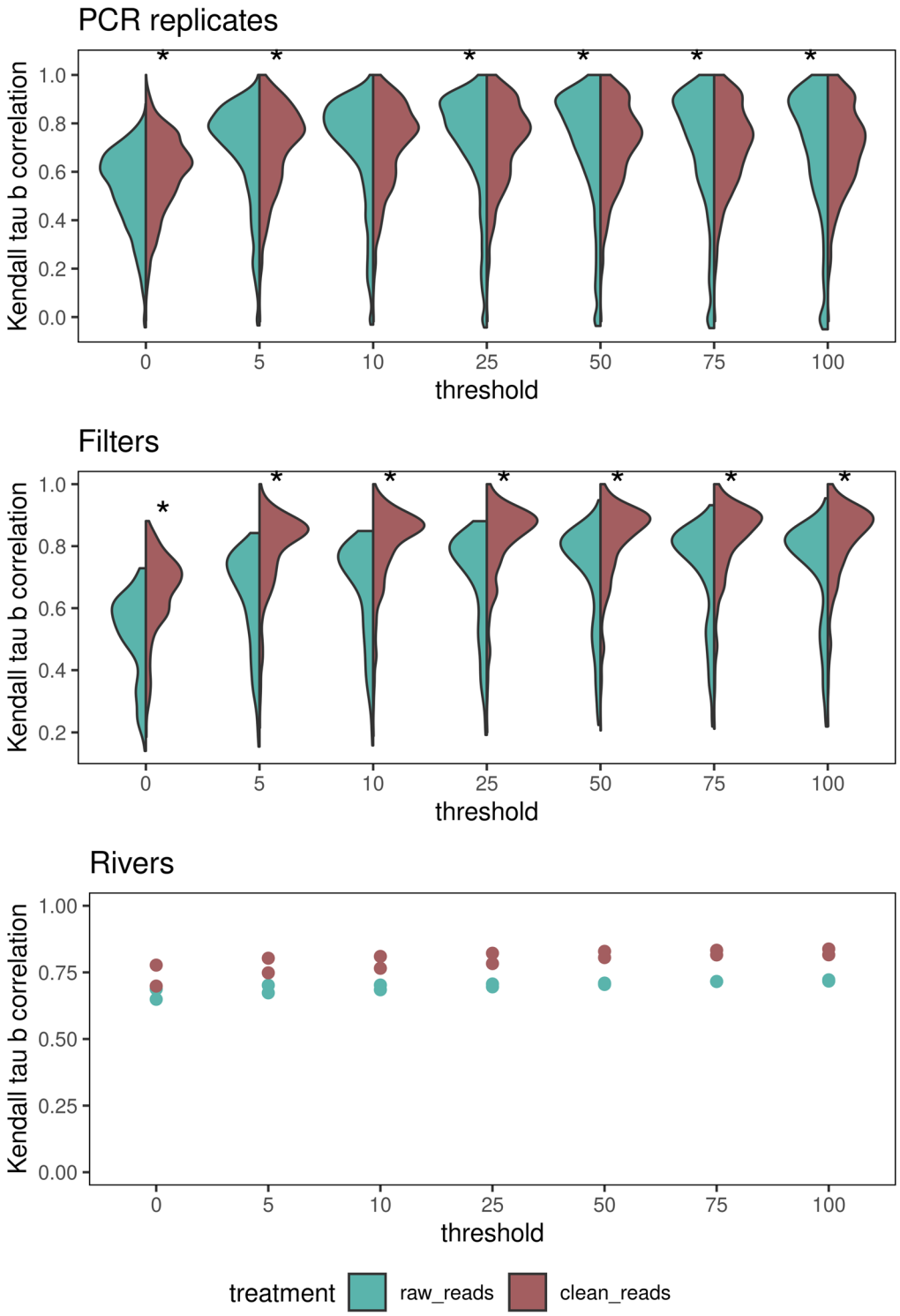

Figure 2. Kendall Tau-b correlation coefficient between the outputs of the CNN and OBITools. The left side of the violin plots (blue) displays correlation values between OBITools and the CNN applied to raw reads. The right side of the violin plots (red) displays correlation values between OBITools and the CNN applied to clean reads. The x-axis represents the threshold of the minimum read number per species for the species to be considered present. Stars represent a significant difference between the two correlations. The analysis was made at three levels: PCR replicates (top), eDNA filters (middle), and rivers (bottom).

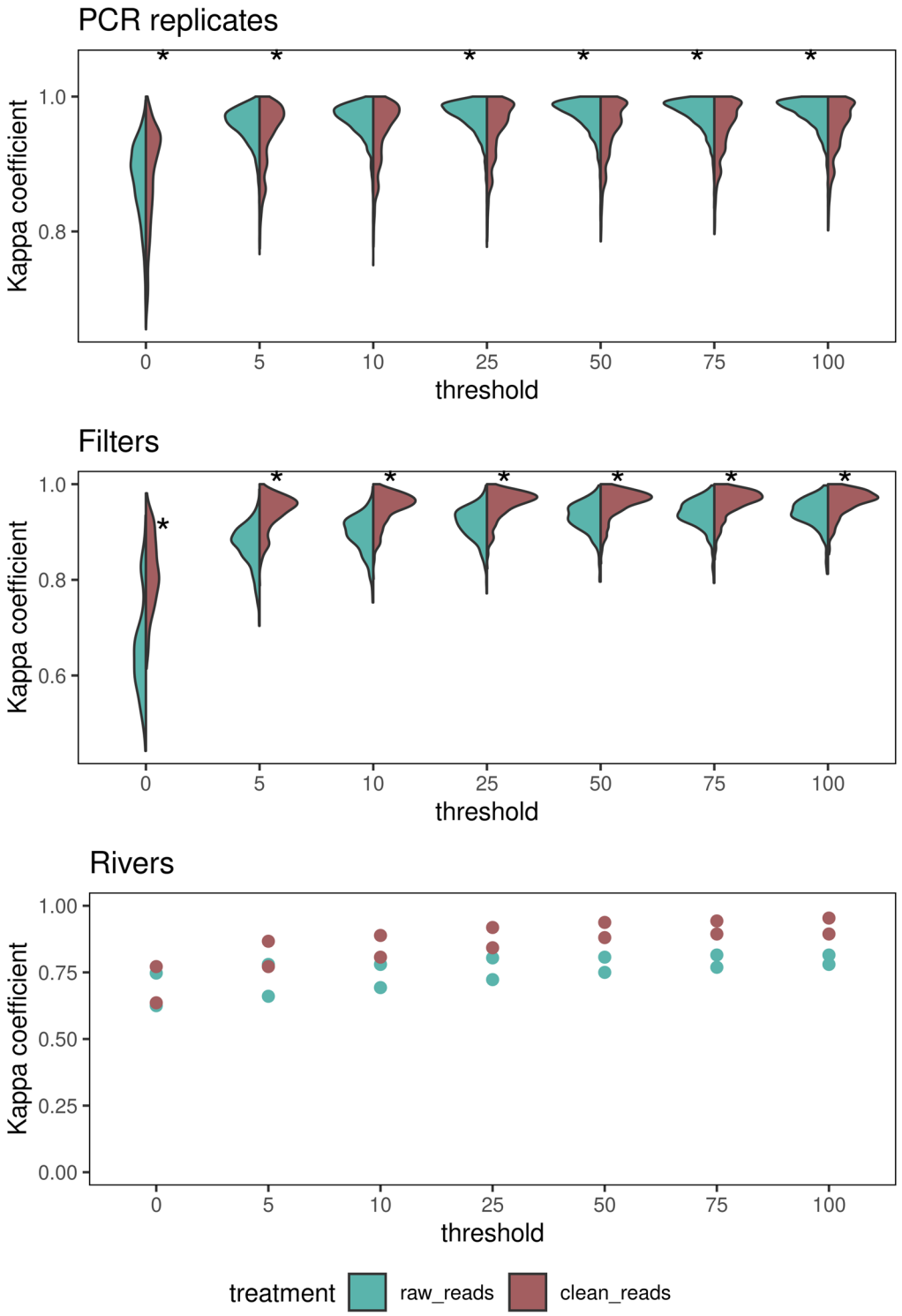

Figure 3. Kappa correlation coefficient between the outputs of the CNN and OBITools. The left side of the violin plots (blue) displays correlation values between OBITools and the CNN applied to raw reads. The right side of the violin plots (red) displays correlation values between OBITools and the CNN applied to clean reads. The x-axis represents the threshold of the minimum read number per species for the species to be considered present. Stars represent a significant difference between the two correlations.

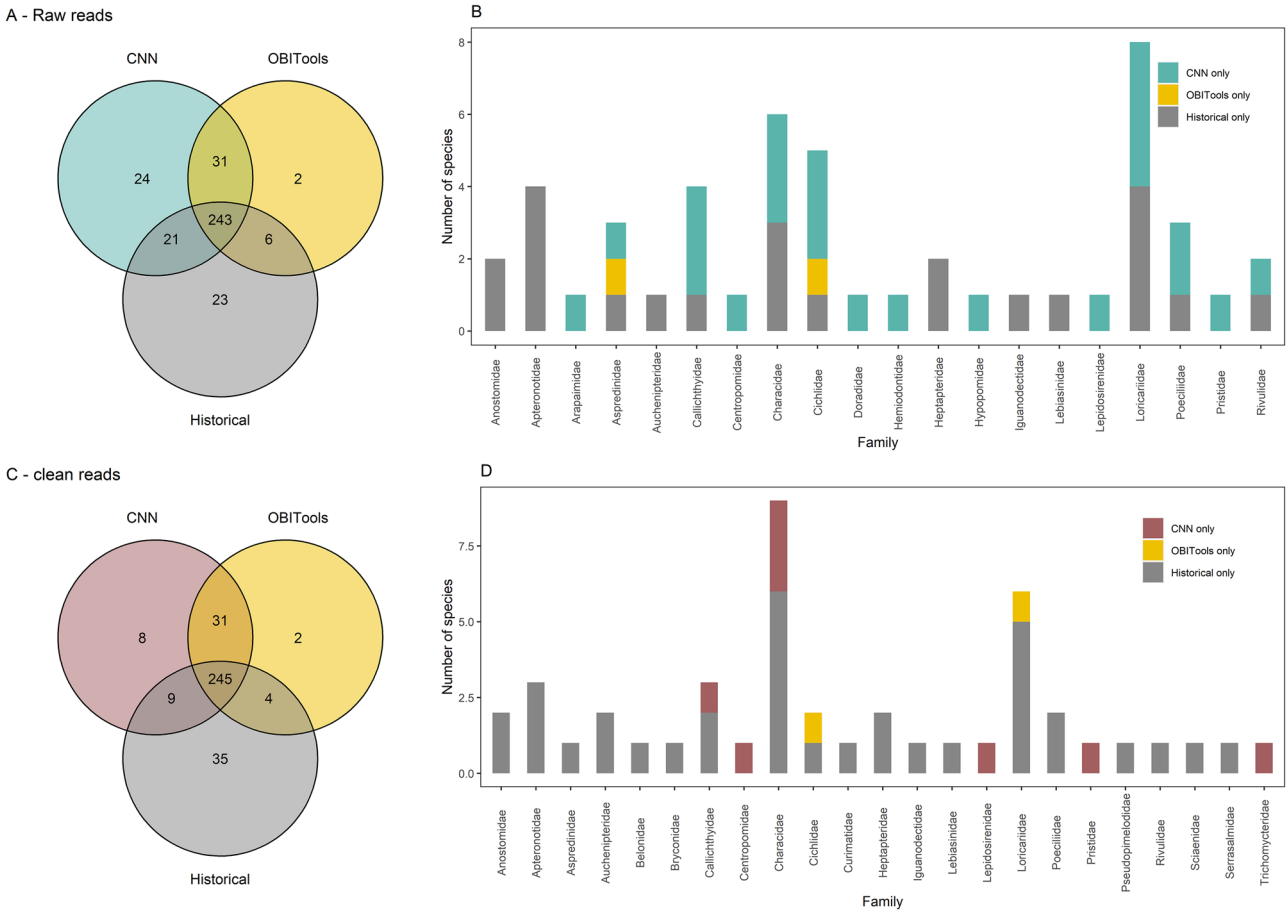

Figure 4. Species detections with the CNN approach, with OBITools, and in historical records in the combined Maroni and Oyapock rivers. (**A**) Overlap of species detections between the CNN applied to raw reads (blue), OBITools (yellow) and historical records (grey). (**B**) Number of species per family, detected with only one method (CNN applied to raw reads, OBITools or historical records). (**C**) Overlap of species detections between the CNN applied to clean reads (red), OBITools (yellow) and historical records (grey). (**D**) Number of species per family that were detected with only one method (CNN applied to clean reads, OBITools or historical records).

the OBITools pipeline is dedicated to the alignment (up to 80%) and demultiplexing (up to 15%) steps. By training and applying a convolutional neural network directly on raw reads, we could sidestep this issue completely and achieve significantly faster processing times and lower power consumption at the cost of more marked differences in the recovered compositions overall.

Discussion

The monitoring of biodiversity in highly species-rich ecosystems has generally been challenging, with gaps in biodiversity data existing in the tropics[23]. eDNA metabarcoding is a revolutionary method that can enhance the monitoring of species in complex ecosystems[48], but is associated with the challenge of rapidly processing large data sets. Our study demonstrates the application of a CNN to process short eDNA sequence reads directly from raw sequencing Illumina outputs. We show that the CNN approach delivers species compositions comparable to those from OBITools and historical records. Fish assemblages retrieved using OBITools and CNN were consistent with the current knowledge on Guianese fish fauna, with marked differences between coastal and inland sites[29]. Fish homogeneity in coastal areas was explained by a historical connectivity between the coastal basins during the Miocene[21], but also by the salt tolerance of a substantial number of the fishes inhabiting coastal streams[45]. Composition analysis further highlighted sites with a markedly different fauna, corresponding to the areas heavily disturbed by gold mining, forestry and agriculture[4]. In only a few minutes, the software transformed a raw fastq sequence data set into a species list associated with each eDNA sample collected in the field, which can serve further biodiversity analyses. Overall, our findings indicate that machine learning offers new possibilities for the taxonomic labelling of short DNA sequences and can transform rapidly collected eDNA data samples into interpretable taxonomy-based biodiversity indicators[25,30,68].

In classical bioinformatic pipelines, the processing from raw sequence reads to taxonomic identifications includes seven steps (paired-end read merging, demultiplexing, dereplication, quality filtering, removal and correction of PCR/sequencing errors, and taxonomic labelling) expected to be essential to generate high-quality results from metabarcoding studies, but which can be computationally demanding[8,16] and challenging to

articulate[50]. We show that a CNN can embed all these steps in a single process applied directly to the raw Illumina reads when the CNN is trained to handle noisy data. Moreover, for relatively short eDNA markers (e.g. 60 bp for the 'teleo' marker used here), merging paired-end reads is not necessary, which leads to a significant computational gain[52]. While still offering results roughly comparable to those of OBITools, the CNN decreases the processing time of the whole data set analysis by a factor of around 150. In a recent comparison, Barque (https://github.com/enormandeau/barque) combined with a fast demultiplexing module was able to process over 15 million reads in 30 min, while it took 17 h for OBITools V1[52]. Assuming the same rate as found for our CNN, i.e. 1 million read per minute, to this data set, the application of the CNN would be two times faster than the fastest existing bioinformatic pipeline in a single model[52]. Our study represents a first successful adaption of CNN to the processing of eDNA metabarcoding data, but we foresee several avenues of optimization to gain speed and accuracy, making it a promising tool for scaling-up biodiversity inventories via eDNA[5,64].

The training of CNNs leads to an efficient adjustment to the reference database, avoiding the need to explore a large number of parameters and arbitrary thresholds, as required in classical bioinformatic pipelines. Existing bioinformatic pipelines contain a variety of modules (i.e. QIIME2, DADA2, Vsearch), each with its own set of parameters[7,17,62]. Selecting the appropriate modules and parameters requires advanced knowledge of the functioning of the program, since changes in those parameters can considerably modify the outputs[8,13,36]. The absence of an appropriate and automated method for parameter optimization[2] often limits the use of those pipelines by non-specialists. In contrast, the application of a CNN only includes a first step of training, where the optimization of the network is nearly automated, and two independent steps for applying the CNN and demultiplexing the reads to reach to final taxonomic outputs per sample. During the learning step of a CNN, only three parameters have to be set by the user: the network size (number of layers, filters and units), the learning rate, and the augmentation values. During the application step, two parameters are optimized, the binarization threshold and the minimum number of reads per sample to be considered. We expect that these steps can be nearly automated within a user-friendly software, as developed for other machine learning applications[76]. Given the relative ease of the training process and application of CNN, the approach could be transformed into an application with a user-friendly interface demanding only a minimum amount of interaction. Hence, CNNs could make eDNA metabarcoding data processing accessible even to less trained users and provide an overview of biodiversity more rapidly.

A CNN trained on a complete reference database produced species composition outputs congruent with the outputs of a popular bioinformatic pipeline, but showed a tendency to predict more species than those of OBITools and historical records. Compositional differences in the outputs of pipelines have already been highlighted (e.g.[11]) and have mainly resulted from the detection of several false positives and false negatives[52]. With a binarization threshold of 0.9 optimized during the training phase, we found congruent but slightly divergent results between OBITools and the CNN applied to either the raw or clean reads. While the CNN and OBITools shared most of their recovered species, each method detected a few species not detected by the other approach (Fig. 4). However, the CNN showed a general tendency of overprediction compared with OBITools and the historical records, especially when it was applied directly to the raw sequencing data. Using the historical records as a baseline, the CNN applied to clean reads reduced the detection of species only found with the CNN, without decreasing the number of species shared with OBITools or historical records, suggesting false positives resulting from noisy inputs. Specifically, the CNN applied to raw reads detected more species from the Loricariidae, Cichlidae and Characidae families that were not found with OBITools, which may have been the result of sequencing errors that were not denoised by the CNN. In the case of the Cichlidae family, the short barcode we used is known to be poorly resolved[73], with many species sharing the same sequence[61], and our CNN did not perform well in this situation, like all other pipelines. Moreover, Loricaridae and Characidae are the two most speciose families of the Guianese fish fauna, with more than 50 species per family[45] and with several new species occurrences recorded each year in Guianese rivers (e.g.[12]). These two families, together with Cichlidae, are also known to host cryptic and still unnamed species, as shown by Papa et al. for the Maroni river[57]. This could also contribute to species misdetections. Finally, we found that the correlation between OBITools and CNN was lower at the sample level than at the level of the PCR replicates when the CNN was applied to raw reads. Hence, appropriately combining the PCR replicates could confer more robustness to the final outputs of the CNN. Refinement of the network could be added, so that the detection across multiple PCR replicates could be used to compute the final likelihood.

In our study we proposed a novel application of a CNN approach to eDNA metabarcoding data, but several improvements are required before broad-scale applications to large eDNA data sets can be considered. The CNN trained in this study learns from the species class and is forced to assign the sequences to that taxonomic level. Thus, when presented with conflicting sequences, the network might assign all of them to a single species, or may split the probabilities across several species, which might then be discarded given the use of the 0.9 binarization threshold. In contrast, in the case of a conflict, OBITools can assign sequences to higher taxonomic levels, thus keeping information related to these species with identical sequences. In this case study, we had an ideal situation where the reference database was almost complete for the territory. The CNN could be improved to handle incomplete reference databases and to be able to assign a read to another taxonomic level or to an unknown class, rather that forcing a species-level identification and relying on the binarization threshold to reject unknown sequences. Further, we expect that it is possible to improve the CNN by implementing more stringent filters that would reduce the number of false detection and prediction errors. For instance, a filter for tag-jump handling, included in previous pipelines for eDNA metabarcoding for fish (e.g.[22]) could be considered. Finally, while the computational speed was already faster than existing traditional pipelines, specific optimizations, such as network pruning or lower precision computations, could improve the performance further, making this approach even more attractive for applications in future broad-scale eDNA projects.

Conclusion and perspectives. We have demonstrated that we can use deep learning to increase the speed and decrease the energy consumption required for processing eDNA metabarcoding data, with a high accuracy when applied to clean reads and a slightly lower accuracy for raw reads. The largest part of the computation time for the CNN is for the training phase; once trained, the CNN can be used as a computationally efficient tool for applications in the cloud, facilitating analyses of the mass of eDNA data expected to be collected in future biodiversity surveys. eDNA data are being collected at an exponentially increasing rate. Owing to its easy application—due to the reduced number of processing steps and the automated learning of best-suited parameters, a CNN approach contrasts with other widely-used bioinformatic pipelines. Our work paves the way towards computationally efficient and user-friendly online processing pipelines that will contribute to the democratization of bioinformatic analyses of eDNA samples. Our work is a major complement to the recent development and standardization of eDNA in the laboratory; together, they will make it possible to extend the use of eDNA in community ecology and biogeography, even for poorly understood ecosystems or lineages[43], and they will help to install eDNA as a standard monitoring tool[42]. Our findings also reinforce the initial goal of quick and efficient eDNA application for biodiversity monitoring. We expect that the results from this study will be scaled up to help CNNs become a major toolkit for ecological analyses of eDNA data, possibly associated with a cloud infrastructure and parallel computation on GPUs.

Material and methods

eDNA data collection and reference database. As a test data set we used data collected in French Guiana, a *c*. 80,000 km^2 South American territory almost entirely covered by dense primary forest (Supplementary Material Fig. 2). The equatorial climate, associated with abundant rainfall, has created a dense hydrographic network consisting of six major watersheds and several coastal rivers that host a highly diverse fish fauna with at least 368 strictly freshwater fish species[45]. eDNA field collection was initiated in 2014 and continued until 2020. We sampled over 200 sites (see Murienne et al.[54] for details), where we filtered 30 litres of river water across a flow filtration capsule using a peristaltic pump. For the purposes of this study, we analysed only the filters collected in both the Maroni and Oyapock rivers.

At each site we collected one to ten filtration capsules, but at most sites two capsules were used (2×34 l), using a previously established protocol[20,26]. We used a peristaltic pump (Vampire sampler, Burlke, Germany) and disposable sterile tubing to pump the water through the encapsulated filtering cartridges (VigiDNA 0.45 µM, SPYGEN, France). We held the input part of the tube a few centimetres below the surface in rapid hydromorphologic units to facilitate homogenization of DNA in the water column. When the filters began to clog, we decreased the pump speed to avoid material damage. To minimize DNA contamination, the operators remained downstream from the filtration site, either on the boat or on emerging rocks. After filtration, we filled the capsules with a preservation buffer and stored them in the dark at room temperature for less than 1 month before DNA extraction. We applied the 12S rRNA 'teleo' gene fragment[77] using PCR and sequenced it on an Illumina platform, generating an average of 500,000 paired-end sequence reads per sample. The DNA extraction, amplification and sequencing protocol have been described previously[19].

We generated an eDNA reference database by combining fish specimens caught using various types of fishing gear. These data were complemented by fish collections carried out by environmental management agencies (DGTM Guyane, Office de l'eau Guyane, Hydreco laboratory), fish hobbyists (Guyane Wild Fish), and Museum tissue collections (MHN Geneva). Although rare for Guianese fishes, we also included existing sequence data from online databases (Genbank, Mitofish). We extracted and sequenced the 12S ribosomal gene from the collected species. The local reference database has improved over the years[20,22] and now covers over 368 species out of 380 estimated to occur in the region. This almost full coverage is exceptional considering the many gaps globally[51]. Sample collection was authorized by both the French Ministry of Environment (DEAL) and the Guyanese National Park (PAG). The samples comply with the international rules of the Nagoya protocol for access and benefit sharing (project refs ABSCH-IRCC-FR-246820-1 and ABSCH-IRCC-FR-245902-1).

OBITools bioinformatic pipeline. As a standard processing pipeline we selected OBITools[10], which is commonly used in eDNA metabarcoding studies[15,47,78]. We processed the reads from the sequencing following Valentini et al.[77]. In short, we assembled the forward and reverse reads using `illuminapairedend` with a minimum score of 40, retrieving only joined sequences. We then assigned the reads to each sample using `ngsfilter`. We then created a separate data set for each eDNA sample by splitting the original data set into several files using `obisplit`. After this step, we analysed each sample individually before generating the taxonomic list. We clustered strictly identical sequences together using `obiuniq`. Further, we excluded sequences shorter than 20 bp using `obigrep`. We then ran `obiclean` within each PCR product for clustering. We discarded all sequences labelled as 'internal', corresponding most likely to PCR substitutions and indel errors. We performed taxonomic labelling of the remaining sequences using `ecotag` with the custom genetic reference database relevant for the eDNA samples. Finally, we applied an empirical threshold to account for tag-jumps and spurious errors.

Reference data augmentation and training data set. The reference database has a full species coverage, but the number of DNA replicate sequences for each species was limited because there were only 683 sequences for 368 species. This makes training a CNN challenging for several reasons. The number of sequences per species is not balanced, there are not enough sequences to capture the entire inter- and intraspecific variation, and the noise from the sequencing process is not accounted for. To balance the data set using data augmentation procedures, we oversampled the underrepresented species before training. To increase the sequence variation, we implemented an inline sequence mutation step similar to that applied by Busia et al.[14]. During each

training epoch all sequences were randomly mutated. We added between zero and two random insertions and deletions each, as well as noise in the form of a 5% mutation rate. This procedure further reduced overfitting, as no training sample was likely to be repeated twice. For the evaluations, we either added no augmentation or 2% noise and singular insertions and deletions, as we expected the PCR amplification and sequencing to be better than the 5% noise considered during the training phase.

For the direct application on the raw reads, another data transformation step was required. All sequences processed in an Illumina machine retain the selected primers, and were tagged with 8-bp-long tags. During the sequencing two bases from the plate attachment sequence were often read as well. We therefore pre- and appended the forward and reverse primers, and the combined tags and attachment bps to the sequences from the reference database. Specifically, we added 10 bp of unknown bases to each reference sequence, represented by the IUPAC code 'N'. This shifted the sequences to a position in the training input similar to where they would occur in the Illumina data. While there is a canonical read direction for DNA, the read direction during the sequencing randomly occurs on either DNA strand. Therefore, we added the reverse complement of all sequences to the final data set. As a last step we truncated all sequences to a read length of 150 bps, as fixed by the field metabarcoding data.

Convolutional networks. CNNs play a key role in modern computer vision applications and date back to the emergence of artificial neural networks in the 1950s and 1960s. Some of the first applications of CNNs and their training method include digit recognition for handwritten ZIP codes[46]. Each convolutional layer in a CNN consists of a number of convolutional kernels often called filters. These filters can be thought of as feature detectors each responding to a specific feature in the input data. Compared with fully connected dense layers, the small extent of these filters drastically reduces the number of free parameters to train. Intuitive examples in image processing are edge or corner detectors. By arranging the DNA sequences as two-dimensional inputs, the convolutional layer can learn and exploit abstract features in the sequences.

CNN training and evaluation using split sampling. We investigated the performance of a CNN approach trained on the reference database at the species level. To encode DNA sequence information, each canonical base (A, C, T, G) and each IUPAC ambiguity code was translated to an appropriate four-dimensional probability distribution over the four canonical bases (A, T, C, G), including uncertain base reads (e.g. W and S). For example 'A' became [1,0,0,0] and 'W' became [0.5, 0, 0, 0.5]. The neural network was designed and optimized through a series of tests that allow the optimal set of correct DNA features to be selected. In particular, we explored an exhaustive number of model sizes, including one to three layers of 2D (depth-wise separable) convolutions with 4–16 filters each, one to three fully-connected layers with varying numbers of neurons each, and a softmax activated output layer which produces a probability distribution over all possible taxonomic labels. We applied dropout regularization and used leaky rectified-linear activation for all but the last layers.

We used TensorFlow[1] to train the CNNs with all the aforementioned data augmentations. Due to the sparse data set, we characterized and evaluated the performance of the neural networks using three different methods. First, we applied random split-sampling from the reference database. This established a proper separation between the training and validation data, but less than half the species in the reference data set had two or more sequences, resulting in a reduced range of species that could be included. Specifically, only 156 out of 368 species possessed two or more unique sequences and were considered for the split data set. Next, we trained several networks on the full reference data set with all 368 species and validated them using the original non-augmented reference data. We derived more synthetic data from the reference sequences similar to the training augmentations and evaluated them with the chosen network. We evaluated whether there were systematic errors in the CNN performance. We further investigated whether a binarization threshold, requiring the probability of the most likely prediction to be above a certain value, improved the classification performance. As we prioritized the absence of errors, i.e. fewer false positives, over the presence of correct predictions, we evaluated the effects of such a binarization threshold using the F-beta measure, which uses a weighted trade-off between these errors. We chose a small beta value of 0.3 to heavily discourage false positives at the cost of discarding some correct results.

CNN application on demultiplexed and cleaned samples. We tested the best trained CNN on the curated eDNA reads after the application of the main cleaning steps of the OBITools pipeline. In particular, from the Illumina raw output, we assembled the forward and reverse reads using the `illuminapairedend` algorithm from the OBITools package, after which we kept only high-quality reads and demultiplexed them across the different eDNA samples. We applied the best-trained CNN at the species level to these curated eDNA samples. We compared the taxonomic labelling performed by the CNN to classic labelling using `ecotag` from OBITools. We evaluated and applied different thresholds for accepting species detection as a way to remove spurious errors and wrong assignments (0, 5, 10, 25, 50, 75 and 100 reads in at least one PCR replicate). For each eDNA PCR replicate and filter, and for the whole rivers, we ranked the taxonomic groups by the number of reads recovered by each method and performed a Kendall rank correlation. We ran one rank correlation per eDNA sample and reported the median rank correlation across all samples. In addition, we compared the presence–absence using the kappa statistic, which measured the general agreement between the methods for each sample. We calculated the median percentages and median kappa values across the samples. Then, across all eDNA samples, we correlated the species richness obtained via CNN with that obtained with OBITools. Each analysis was performed at three different scales: the PCR replicate, the filtration capsule and the river. Finally, we ordinated the species composition of each filtration capsule for both methods using a principal coordinate analysis (PCoA), to compare differences in recovered compositions among the methods.

CNN application on the raw illumina sequences. We applied the best CNN directly to the raw outputs from the Illumina sequencing, where we omitted all the preprocessing steps from OBITools. The CNN was expected to learn how to ignore the primer, as it was constant for all presented sequences. Furthermore, the output sequences from the Illumina sequencer were fixed in length (150 bp), so we fixed the input width of the CNN to this size. We systematically zero-padded or truncated the input sequences to this length during training, evaluation, and application. After the training phase with the reference database and the application on fastq, we developed a custom code for the fast demultiplexing of the reads. By focusing on the tag information in the first few positions of the sequence and not considering read errors in tags, we reduced the demultiplexing to a few simple look-ups in a hash table (currently 5), therefore reducing computation time with limited information loss. As in the previous test, we obtained a list of taxonomic labels for each eDNA sample, which could be compared with species composition information obtained with the OBITools pipeline. We further applied a threshold approach, obtaining a predicted composition per sample for any threshold tested. As done previously, for each eDNA sample, we ranked the taxonomic groups by the number of reads recovered with each method and performed a rank correlation. We calculated the median rank correlations across all the eDNA samples. In addition, we compared the presence–absence at the species level using the overall kappa statistic. We further evaluated whether differences between methods were more frequent in specific taxonomic families than others. Then, across all eDNA samples, we correlated the species richness obtained via CNN with that obtained with OBITools, and ordinated species composition of each filter for both methods on a PCoA. We evaluated the change in accuracy between the CNN applied to curated reads compared with the CNN applied to raw fastq files.

Validation with existing biodiversity knowledge on the region. We compared the species composition recovered in the eDNA samples by CNN and OBITools to the species, genus and family checklists of each river catchment. Species lists for each catchment were obtained from an updated version of the catchment-scale species lists[45] provided in Le Bail et al. From this list, we updated the taxonomy and added novel occurrences of known species based on fish catches by several research and management organizations (see 'Material and methods' section). Only collected specimens with a validated taxonomy were considered when updating this list, and detections using eDNA were not considered. We specifically quantified the number of matching species, false presences and false absences from each method, taking the checklists as references.

Data availability

Partial data is available through Cilleros et al.[22]. The full data set is available from the corresponding author upon request.

Code availability

The code is available in the supplementary material and released under the AGPLv3 license.

Received: 13 October 2021; Accepted: 24 May 2022
Published online: 17 June 2022

References

1. Abadi, M., Agarwal, A., Barham, P., Brevdo, E., Chen, Z., Citro, C., Corrado, G., Davis, A., Dean, J., Devin, M., Ghemawat, S. & Zheng, X. TensorFlow: Large-scale machine learning on heterogeneous systems. Software available from tensorflow.org. (2015).
2. Alberdi, A., Aizpurua, O., Gilbert, M. T. P. & Bohmann, K. Scrutinizing key steps for reliable metabarcoding of environmental samples. *Methods Ecol. Evol.* **9**, 134–147 (2018).
3. Albert, J. S. & Reis, R. E. One. Introduction to Neotropical freshwaters. In Historical biogeography of Neotropical freshwater fishes (pp. 3-20). University of California Press. (2011).
4. Allard, L., Popée, M., Vigouroux, R. & Brosse, S. Effect of reduced impact logging and small-scale mining disturbances on Neotropical stream fish assemblages. *Aquat. Sci.* **78**, 315–325 (2016).
5. Berry, O. *et al.* Making environmental DNA (eDNA) biodiversity records globally accessible. *Environ. DNA* **3**(4), 699–705 (2020).
6. Bohmann, K. *et al.* Environmental DNA for wildlife biology and biodiversity monitoring. *Trends Ecol. Evol.* **29**(6), 358–367 (2014).
7. Bolyen, E. *et al.* QIIME 2: Reproducible, interactive, scalable, and extensible microbiome data science. *Nat. Biotechnol.* **32**, 852–857 (2019).
8. Bonder, M. J., Abeln, S., Zaura, E. & Brandt, B. W. Comparing clustering and pre-processing in taxonomy analysis. *Bioinformatics* **28**(22), 2891–2897 (2012).
9. Boussarie, G. *et al.* Environmental DNA illuminates the dark diversity of sharks. *Sci. Adv.* **4**, eaap9661 (2018).
10. Boyer, F. *et al.* obitools: A unix-inspired software package for DNA metabarcoding. *Mol. Ecology Resour.* **16**(1), 176–182 (2016).
11. Brandt, M.I., Trouche, B., Quintric, L., Günther, B., Wincker, P., Poulain, J. & Arnaud-Haond, S. Bioinformatic pipelines combining denoising and clustering tools allow for more comprehensive prokaryotic and eukaryotic metabarcoding. Molecular Ecology Resources. Accepted (2021).
12. Brosse, S., Melki, F. & Vigouroux, R. Fishes from the Mitaraka mountains (French Guiana). *Zoosystema* **41**, 131–151 (2019).
13. Brown, E. A., Chain, F. J., Crease, T. J., MacIsaac, H. J. & Cristescu, M. E. Divergence thresholds and divergent biodiversity estimates: can metabarcoding reliably describe zooplankton communities?. *Ecol. Evol.* **5**(11), 2234–2251 (2015).
14. Busia, K., George, D. E., Fannjiang, C., Alexander, D.H., Dorfman, E., Poplin, R., Chang, P., & DePris, M. A deep learning approach to pattern recognition for short DNA sequences. BioRxiv (2020).
15. Bylemans, J., Gleeson, D. M., Hardy, C. M. & Furlan, E. Toward an ecoregion scale evaluation of eDNA metabarcoding primers: A case study for the freshwater fish biodiversity of the Murray-Darling Basin (Australia). *Ecol. Evol.* **8**(17), 8697–8712 (2018).
16. Calderón-Sanou, I., Münkemüller, T., Boyer, F., Zinger, L. & Thuiller, W. From environmental DNA sequences to ecological conclusions: How strong is the influence of methodological choices?. *J. Biogeogr.* **47**(1), 193–206 (2020).
17. Callahan, B. J. *et al.* DADA2: High-resolution sample inference from Illumina amplicon data. *Nat. Methods* **13**(7), 581–583 (2016).
18. Cantera, I., Coutant, O., Jézéuel, C., Decotte, J.B., Dejean, T., Vigouroux, R., Valentini, A. Murienne, J. & Brosse S. Slight deforestation causes harsh biodiversity decline in Amazonian rivers (submitted)
19. Cantera, I., Decotte, J. B., Dejean, T., Murienne, J., Vigouroux, R., Valentini, A., & Brosse, S. Characterizing the spatial signal of environmental DNA in river systems using a community ecology approach. BioRxiv (2020).

20. Cantera, I. *et al.* Optimizing environmental DNA sampling effort for fish inventories in tropical streams and rivers. *Sci. Rep.* **9**(1), 1–1 (2019).
21. Cardoso, Y. P. & Montoya-Burgos, J. I. Unexpected diversity in the catfish Pseudancistrus brevispinis reveals dispersal routes in a Neotropical center of endemism: The Guyanas Region. *Mol. Ecol.* **18**, 947–964 (2009).
22. Cilleros, K. *et al.* Unlocking biodiversity and conservation studies in high-diversity environments using environmental DNA (eDNA): A test with Guianese freshwater fishes. *Mol. Ecol. Resour.* **19**(1), 27–46 (2019).
23. Collen, B., Ram, M., Zamin, T. & McRae, L. The tropical biodiversity data gap: Addressing disparity in global monitoring. *Trop. Conserv. Sci.* **1**(2), 75–88 (2008).
24. Cordier, T., Lanzén, A., Apothéloz-Perret-Gentil, L., Stoeck, T. & Pawlowski, J. Embracing environmental genomics and machine learning for routine biomonitoring. *Trends Microbiol.* **27**(5), 387–397 (2019).
25. Cordier, T. *et al.* Ecosystems monitoring powered by environmental genomics: A review of current strategies with an implementation roadmap. *Mol. Ecol.* **30**(13), 2937–2958 (2020).
26. Coutant, O. *et al.* Detecting fish assemblages with environmental DNA: Does protocol matter? Testing eDNA metabarcoding method robustness. *Environ. DNA* **3**(3), 619–630 (2020).
27. Deiner, K. *et al.* Environmental DNA metabarcoding: Transforming how we survey animal and plant communities. *Mol. Ecol.* **26**(21), 5872–5895 (2017).
28. Deneu, B., Servajean, M., Bonnet, P., Botella, C., Munoz, F., & Joly, A. Convolutional neural networks improve species distribution modelling by capturing the spatial structure of the environment. PLoS Comput. Biol. (in press) (2021).
29. de Mérona, B., Tejerina-Garro, F. L. & Vigouroux, R. Fish-habitat relationships in French Guiana rivers: A review. *Cybium* **36**, 7–15 (2012).
30. DiBattista, J. D. *et al.* Environmental DNA can act as a biodiversity barometer of anthropogenic pressures in coastal ecosystems. *Sci. Rep.* **10**(1), 1–15 (2020).
31. Dornelas, M., Madin, E. M., Bunce, M., DiBattista, J. D., Johnson, M., Madin, J. S., Magurran, A. E., McGill, B. J., Pettorelli, N., Pizarro, O. & Williams, S. B. Towards a macroscope: Leveraging technology to transform the breadth, scale and resolution of macroecological data. Glob. Ecol. Biogeogr. (2019).
32. Dufresne, Y., Lejzerowicz, F., Perret-Gentil, L. A., Pawlowski, J. & Cordier, T. SLIM: A flexible web application for the reproducible processing of environmental DNA metabarcoding data. *BMC Bioinform.* **20**(1), 1–6 (2019).
33. Ficetola, G. F., Miaud, C., Pompanon, F. & Taberlet, P. Species detection using environmental DNA from water samples. *Biol. Lett.* **4**(4), 423–425 (2008).
34. Ficetola, G. F., Taberlet, P. & Coissac, E. How to limit false positives in environmental DNA and metabarcoding?. *Mol. Ecol. Resour.* **16**(3), 604–607 (2016).
35. Ficetola, G. F. *et al.* Replication levels, false presences and the estimation of the presence/absence from eDNA metabarcoding data. *Mol. Ecology Resour.* **15**(3), 543–556 (2015).
36. Flynn, J. M., Brown, E. A., Chain, F. J., MacIsaac, H. J. & Cristescu, M. E. Toward accurate molecular identification of species in complex environmental samples: Testing the performance of sequence filtering and clustering methods. *Ecol. Evol.* **5**(11), 2252–2266 (2015).
37. Gold, Z. *et al.* eDNA metabarcoding bioassessment of endangered fairy shrimp (Branchinecta spp.). *Conserv. Genet. Resour.* **12**, 685–690 (2020).
38. Grünig, M., Razavi, E., Calanca, P., Mazzi, D., Wegner, J. D., & Pellissier, L. Applying deep neural networks to predict incidence and phenology of plant pests and diseases. Ecosphere (accepted) (2021).
39. Helaly, M. A., Rady, S., & Aref, M. M. Convolutional neural networks for biological sequence taxonomic classification: A comparative study. In *International Conference on Advanced Intelligent Systems and Informatics* (pp. 523–533). Springer, Cham (2019).
40. Holman, L. E. *et al.* Animals, protists and bacteria share marine biogeographic patterns. *Nat. Ecol. Evol.* **5**(6), 738–746 (2021).
41. Iknayan, K. J., Tingley, M. W., Furnas, B. J. & Beissinger, S. R. Detecting diversity: Emerging methods to estimate species diversity. *Trends Ecol. Evol.* **29**(2), 97–106 (2014).
42. Jarman, S. N., Berry, O. & Bunce, M. The value of environmental DNA biobanking for long-term biomonitoring. *Nat. Ecol. Evol.* **2**(8), 1192–1193 (2018).
43. Juhel, J. B., Utama, R. S., Marques, V., Vimono, I. B., Sugeha, H. Y., Kadarusman, Pouyaud, L., Dejean, T., Mouillot, D. & Hocdé, R. Accumulation curves of environmental DNA sequences predict coastal fish diversity in the coral triangle. *Proc. R. Soc. B* **287**(1930), 20200248 (2020).
44. Kopp, W., Monti, R., Tamburrini, A., Ohler, U. & Akalin, A. Deep learning for genomics using Janggu. *Nat. Commun.* **11**(1), 1–7 (2020).
45. Le Bail, P. Y. *et al.* Updated checklist of the freshwater and estuarine fishes of French Guiana. *Cybium* **36**(1), 293–319 (2012).
46. LeCun, Y. *et al.* Backpropagation applied to handwritten zip code recognition. *Neural Comput.* **1**(4), 541–551 (1989).
47. Li, W. *et al.* Validating eDNA measurements of the richness and abundance of anurans at a large scale. *J. Anim. Ecol.* **90**(6), 1466–1479 (2021).
48. Lopes, C. M. *et al.* eDNA metabarcoding: A promising method for anuran surveys in highly diverse tropical forests. *Mol. Ecol. Resour.* **17**(5), 904–914 (2017).
49. Makiola, A. *et al.* Key questions for next-generation biomonitoring. *Front. Environ. Sci.* **7**, 197 (2020).
50. Marques, V. *et al.* Blind assessment of vertebrate taxonomic diversity across spatial scales by clustering environmental DNA metabarcoding sequences. *Ecography* **43**(12), 1779–1790 (2020).
51. Marques, V. *et al.* GAPeDNA: Assessing and mapping global species gaps in genetic databases for eDNA metabarcoding. *Divers. Distrib.* **27**(10), 1880–1892 (2020).
52. Mathon, L. *et al.* Benchmarking bioinformatic tools for fast and accurate eDNA metabarcoding species identification. *Mol. Ecol. Resour.* **21**(7), 2565–2579 (2021).
53. McGee, K. M., Robinson, C. & Hajibabaei, M. Gaps in DNA-based biomonitoring across the globe. *Front. Ecol. Evol.* **7**, 337 (2019).
54. Murienne, J. *et al.* Aquatic eDNA for monitoring French Guiana biodiversity. *Biodivers. Data J.* **7**, e37518 (2019).
55. Nugent, C. M. & Adamowicz, S. J. Alignment-free classification of COI DNA barcode data with the Python package Alfie. *Metabarcoding Metagenomics* **4**, e55815 (2020).
56. Pagni, M. *et al.* Density-based hierarchical clustering of pyro-sequences on a large scale-the case of fungal ITS1. *Bioinformatics* **29**(10), 1268–1274 (2013).
57. Papa, Y., Le Bail, P. Y. & Covain, R. Genetic landscape clustering of a large DNA barcoding dataset reveals shared patterns of genetic divergence among freshwater fishes of the Maroni Basin. Authorea Preprints (2020).
58. Piro, V. C., Dadi, T. H., Seiler, E., Reinert, K. & Renard, B. Y. ganon: Precise metagenomics classification against large and up-to-date sets of reference sequences. *Bioinformatics* **36**(Supplement 1), i12–i20 (2020).
59. Polanco Fernández, A., Marques, V., Fopp, F., Juhel, J. B., Borrero-Pérez, G. H., Cheutin, M. C., Eme, D. & Pellissier, L. Comparing environmental DNA metabarcoding and underwater visual census to monitor tropical reef fishes. *Environ. DNA* **3**, 142–156 (2021).
60. Polanco, A. *et al.* Comparing the performance of 12S mitochondrial primers for fish environmental DNA across ecosystems. *Environ. DNA* **3**(6), 1113–1127 (2021).

61. Polanco Fernández, A., Martinezguerra, M. M., Marques, V., Francisco Villa-Navarro, Borrero-Pérez, G. H., Cheutin, M. C., Dejean, T., Hocdé, R., Juhel, J. B., Maire, E., Manel, S. & Pellissier, L. Recovering aquatic and terrestrial biodiversity in a tropical estuary using environmental DNA. *Biotropica* **53**(6), 1606–1619 (2021).
62. Rognes, T., Flouri, T., Nichols, B., Quince, C. & Mahé, F. VSEARCH: A versatile open source tool for metagenomics. *PeerJ* **4**, 1–22 (2016).
63. Rojahn, J., Gleeson, D. M., Furlan, E., Haeusler, T. & Bylemans, J. Improving the detection of rare native fish species in environmental DNA metabarcoding surveys. *Aquat. Conserv. Mar. Freshw. Ecosyst.* **31**(4), 990–997 (2021).
64. Ruppert, K. M., Kline, R. J. & Rahman, M. S. Past, present, and future perspectives of environmental DNA (eDNA) metabarcoding: A systematic review in methods, monitoring, and applications of global eDNA. *Glob. Ecol. Conserv.* **17**, e00547 (2019).
65. Sato, Y., Miya, M., Fukunaga, T., Sado, T. & Iwasaki, W. MitoFish and MiFish pipeline: A mitochondrial genome database of fish with an analysis pipeline for environmental DNA metabarcoding. *Mol. Biol. Evol.* **35**(6), 1553–1555 (2018).
66. Schirmer, M. *et al.* Insight into biases and sequencing errors for amplicon sequencing with the Illumina MiSeq platform. *Nucleic Acids Res.* **43**(6), e37 (2015).
67. Schnell, I. B., Bohmann, K. & Gilbert, M. T. P. Tag jumps illuminated–reducing sequence-to-sample misidentifications in metabarcoding studies. *Mol. Ecol. Resour.* **15**(6), 1289–1303 (2015).
68. Sepulveda, A. J., Nelson, N. M., Jerde, C. L. & Luikart, G. Are environmental DNA methods ready for aquatic invasive species management?. *Trends Ecol. Evol.* **35**, 668–678 (2020).
69. Shokralla, S., Spall, J. L., Gibson, J. F. & Hajibabaei, M. Next-generation sequencing technologies for environmental DNA research. *Mol. Ecol.* **21**(8), 1794–1805 (2012).
70. Shorten, C. & Khoshgoftaar, T. A survey on image data augmentation for deep learning. *J. Big Data* **6**, 60 (2019).
71. Singer, G. A. C., Fahner, N. A., Barnes, J. G., McCarthy, A. & Hajibabaei, M. Comprehensive biodiversity analysis via ultra-deep patterned flow cell technology: A case study of eDNA metabarcoding seawater. *Sci. Rep.* **9**(1), 1–12 (2019).
72. Su, G. *et al.* Human impacts on global freshwater fish biodiversity. *Science* **371**(6531), 835 (2021).
73. Taberlet, P., Bonin, A., Coissac, E. & Zinger, L. *Environmental DNA: For Biodiversity Research and Monitoring* (Oxford University Press, Oxford, 2018).
74. Taberlet, P., Coissac, E., Pompanon, F., Brochmann, C. & Willerslev, E. Towards next-generation biodiversity assessment using DNA metabarcoding. *Mol. Ecol.* **21**(8), 2045–2050 (2012).
75. Thomsen, P. F. & Willerslev, E. Environmental DNA-An emerging tool in conservation for monitoring past and present biodiversity. *Biol. Conserv.* **183**, 4–18 (2015).
76. Thuiller, W., Lafourcade, B., Engler, R. & Araújo, M. B. BIOMOD–A platform for ensemble forecasting of species distributions. *Ecography* **32**(3), 369–373 (2009).
77. Valentini, A. *et al.* Next-generation monitoring of aquatic biodiversity using environmental DNA metabarcoding. *Mol. Ecol.* **25**(4), 929–942 (2016).
78. West, K. *et al.* Large-scale eDNA metabarcoding survey reveals marine biogeographic break and transitions over tropical northwestern Australia. *Divers. Distrib.* **27**(10), 1942–1957 (2021).

Author contributions

B.F. and L.P. conceived the idea, study design, and analytic methods. B.F. developed the neural networks and ran the computational study. J.M. and S.B. collected the eDNA samples. A.V. and T.D. completed the eDNA laboratory and data preparation work. B.F. and L.M. analysed the results. L.P., L.M. and B.F. led the writing of the manuscript, with the support of S.M., A.V., T.D., C.A., D.M., W.T., J.M. and S.B.

Competing Interests

The authors declare no competing interests.

Additional information

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1038/s41598-022-13412-w.

Correspondence and requests for materials should be addressed to B.F. or L.P.

Reprints and permissions information is available at www.nature.com/reprints.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.