Deep Sea Explorers

Enzo Oukacha

ICHEP 2022

APC[1]

Jul 8th 2022

[1]Laboratoire Astroparticule et Cosmologie

REINFORCE

- REINFORCE is a project which aims at minimizing the knowledge gap between Large Research Infrastructures and Society through Citizen Science.
- The main tools are the demonstrators which are platforms on which the citizens can do science-related activities.
- These demonstrators are all present on the website Zooniverse.

KM3NeT

- KM3NeT is a 3D array of PMTs[2] placed at the bottom of the mediterranean sea, used to capture the Cherenkov light.
- ORCA[3] will study neutrino properties such as the Neutrino Mass Hierarchy ($E \sim GeV$)
- ARCA[4] will do neutrino and multi-messenger astronomy (E > TeV)

[2] Photomultiplier Tubes
[3] Oscillation Research with Cosmics in the Abyss
[4] Astronomy Research with Cosmics in the Abyss

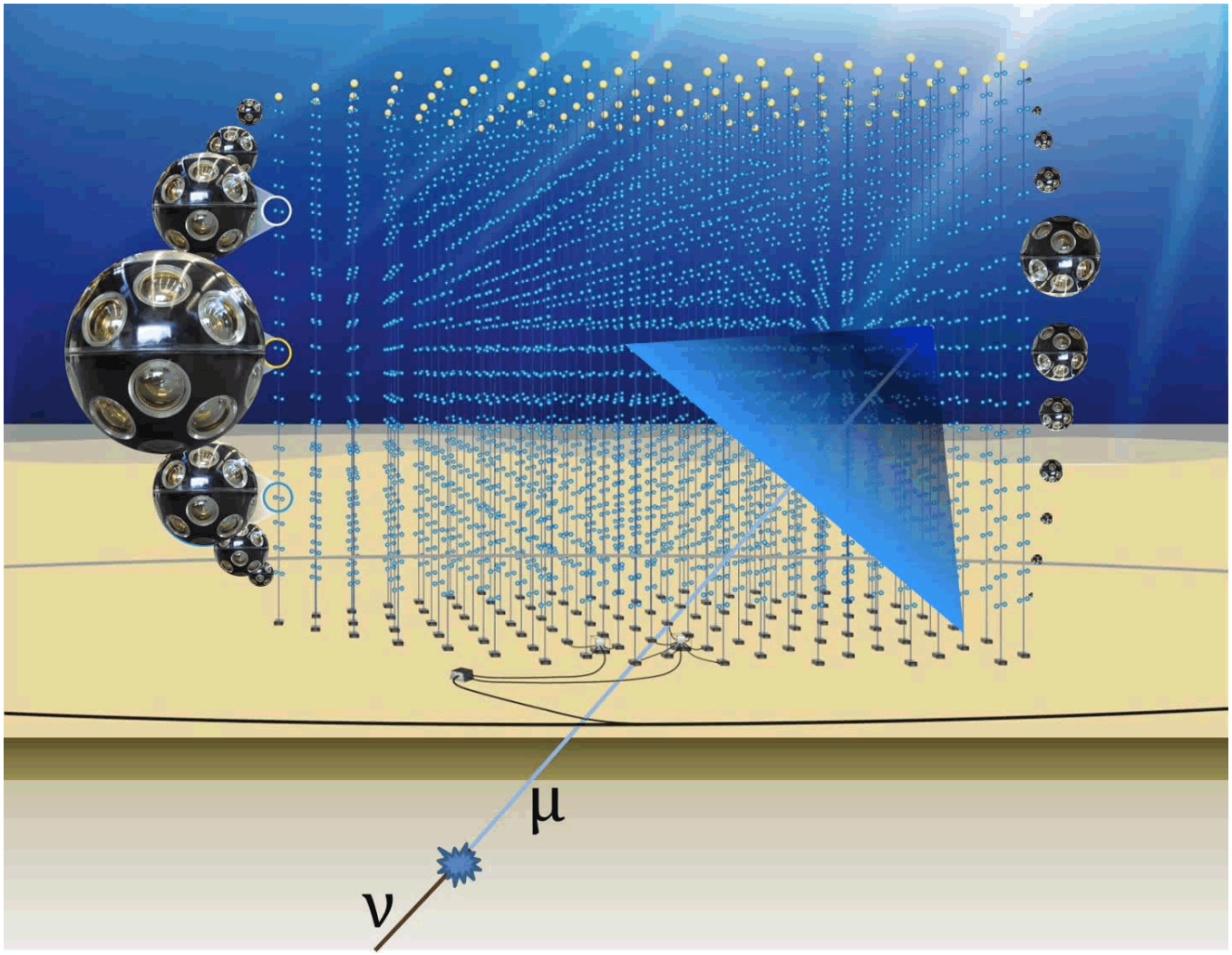

Figure: Illustration of a neutrino event

Deep Sea Explorers

- Main purpose is to be able to classify the bioluminescence events as well as the bioacoustic events at the bottom of the mediterranean sea and to have a better understanding of noises in KM3NeT data.
- Therefore 2 main activities :
  - Classification of the bioluminescence data to be able to distinguish the detection of Neutrinos from the light emitted by the bioluminescent marine organisms.
  - Classification of the ceteceans based on the sound they emit. (Hydrophones of KM3NeT)

Classification of species

Figure: Sperm whale (left) and Short finned pilot whale (right)

- Diving around 2000/3000m depth.
- Emitting very different signals.

Zooniverse

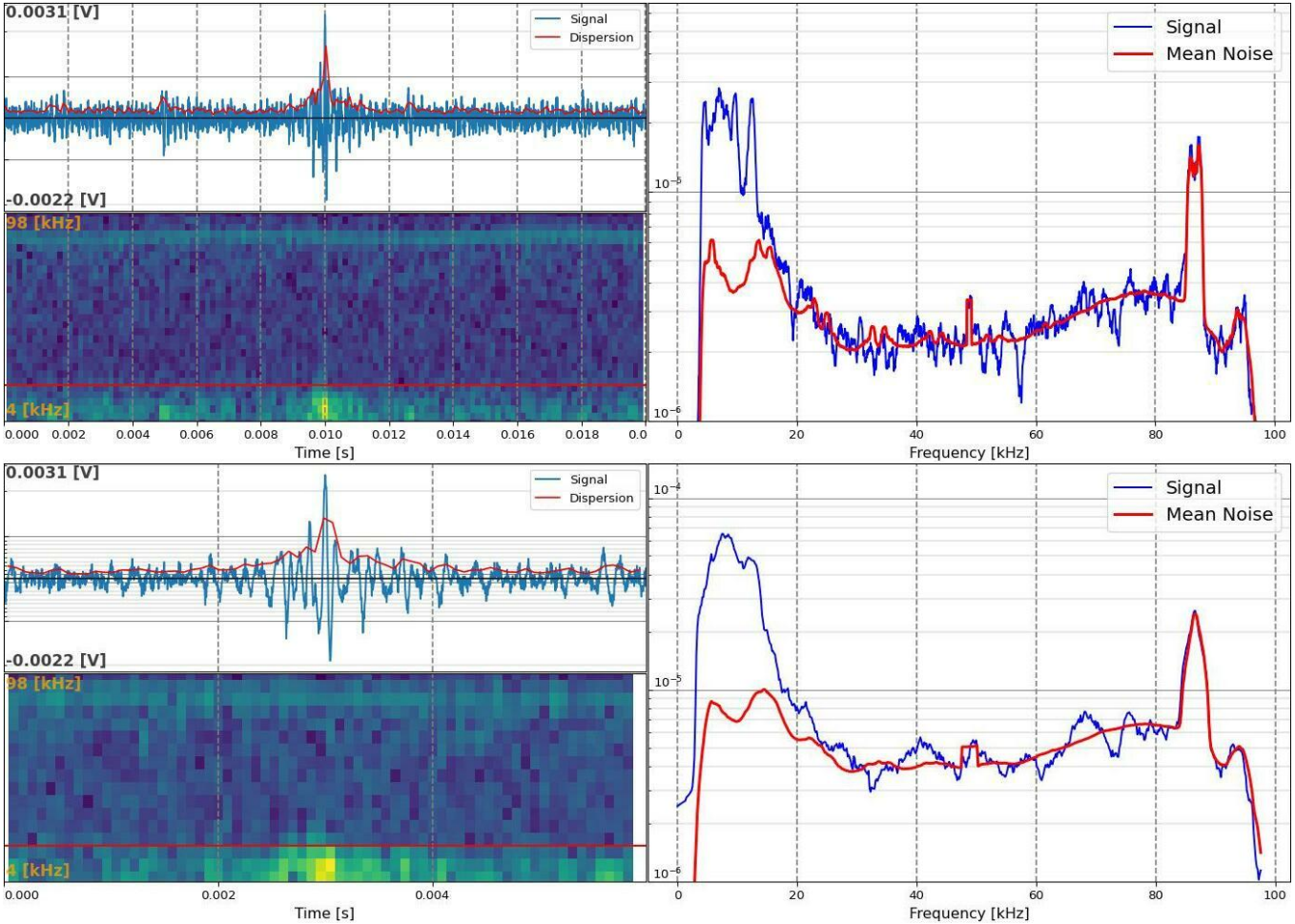

Figure: Example of an event to be classified (here a click[2] emitted from a Sperm whale)

[2] short sound wave

Why do we need the help of volunteers?

- The data that have been uploaded on Zooniverse are sounds that have been classified with less than 70% accuracy by the neural network.
- We need the help of the citizens to classify theses events in order to enhance the training of a neural network.
- Purpose : make a citizen-machine learning comparative study.

CNN model - UpDimV2

Layer name	Input size	Kernel	Strides	Out features
Conv-1D	$N * 4096 * 1$	3	1	32
Conv-1D	$N * 4096 * 32$	3	2	32
Skip	$N * 4096 * 1$	1	2	32
Conv-1D	$N * 2048 * 32$	3	2	64
Conv-1D	$N * 1024 * 64$	3	2	128
Skip	$N * 2048 * 32$	1	4	128
Conv-2D	$N * 1024 * 128 * 1$	3*3	1*1	32
Conv-2D	$N * 1024 * 128 * 32$	3*3	2*2	32
Skip	$N * 1024 * 128 * 1$	1*1	2*2	32
Conv-2D	$N * 512 * 64 * 32$	3*3	2*2	64
Conv-2D	$N * 256 * 32 * 64$	3*3	2*2	128
Skip	$N * 512 * 64 * 32$	1*1	4*4	128

Figure: Architecture of the CNN model (1st part)[2]

[2] M.Ferrari, H.Glotin et al. (2020)

CNN model - UpDimV2

Conv-3D	$N * 128 * 16 * 128 * 1$	3*3*3	1*2*1	32
Conv-3D	$N * 128 * 8 * 128 * 32$	3*3*3	2*2*2	64
Skip	$N * 128 * 8 * 128 * 1$	1*1*1	2*4*2	64
Conv-3D	$N * 64 * 4 * 64 * 64$	3*3*3	2*2*2	128
Conv-3D	$N * 32 * 2 * 32 * 128$	3*3*3	2*2*2	256
Skip	$N * 64 * 8 * 64 * 64$	1*1*1	4*4*4	256
Softmax	$N * 16 * 1 * 16 * 256$	16*1*1		
MaxPool	$N * 16 * 1 * 16 * 256$	16*1*1		
Flatten	$N * 1 * 1 * 16 * 256$			
Dense	$N * 4096$			1024
Dense	$N * 1024$			512
Dense	$N * 512$			7

Figure: Architecture of the CNN model (2nd part)[2]

[2] M.Ferrari, H.Glotin et al. (2020)

CNN model : UpDimV2

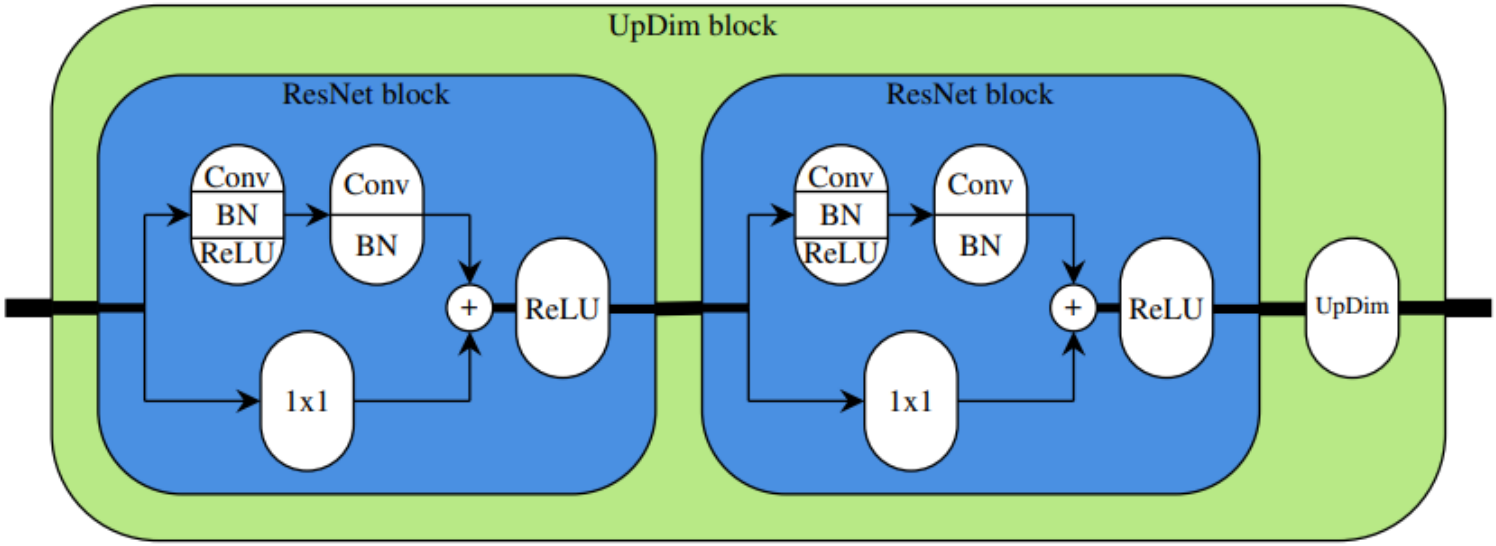

Figure: Architecture of an UpDim block[2]

[2] M.Ferrari, H.Glotin et al. (2020)

UpDimV2 : Data for the training

- Need to have a sample of data/labels (one hot encoded) with a high accuracy for the training (>90% of accuracy)
- Use of the data of the Challenge DOCC10[5] organised by the University of Toulouse the CNRS and ENS.[6]
- Data for which the predictions were made with an accuracy of 95% (clustering method)

[5] Dyni Odontocete Click Classification

[6] https://challengedata.ens.fr/participants/challenges/32/

Data for the training

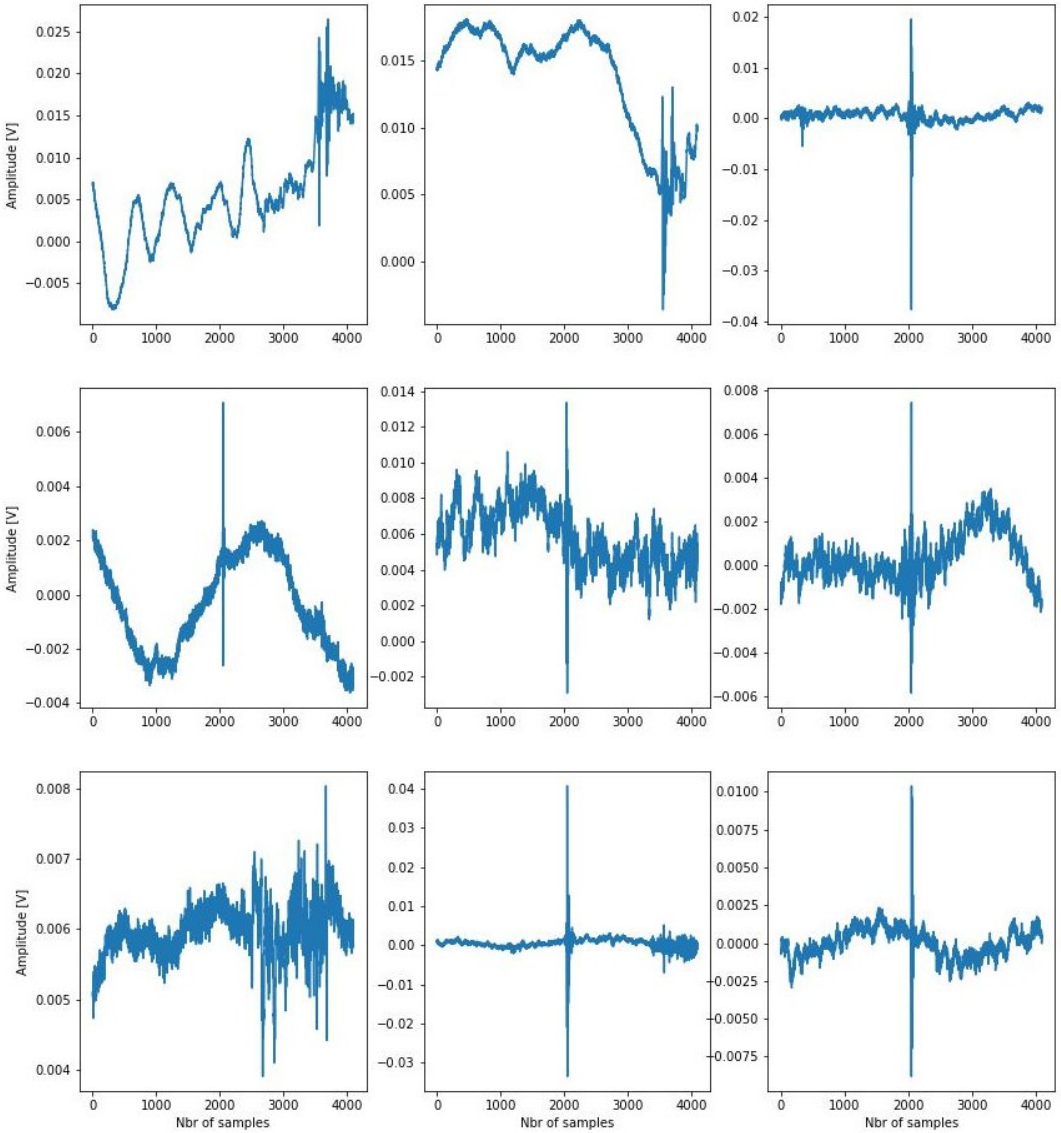

Figure: Example of data in the training set.

First results : Loss plot

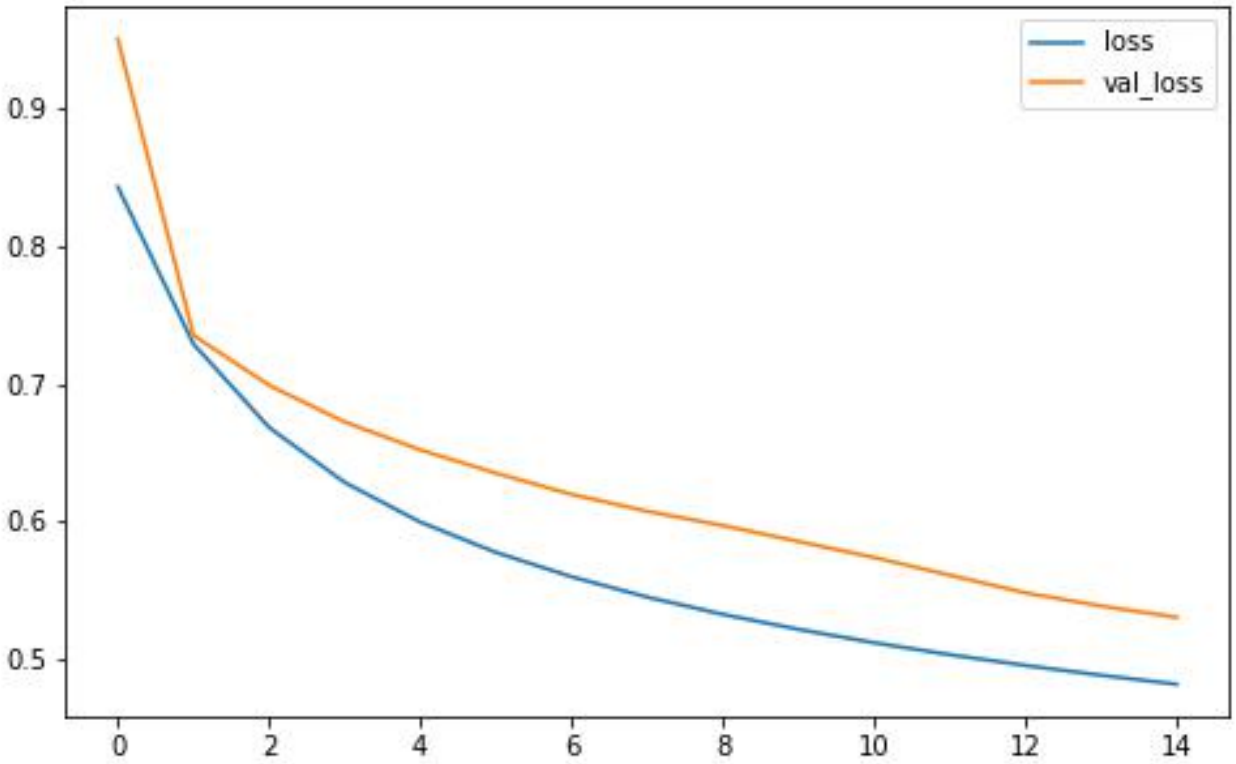

Figure: Training and Validation loss over 15 epochs.

First results : Accuracy plot

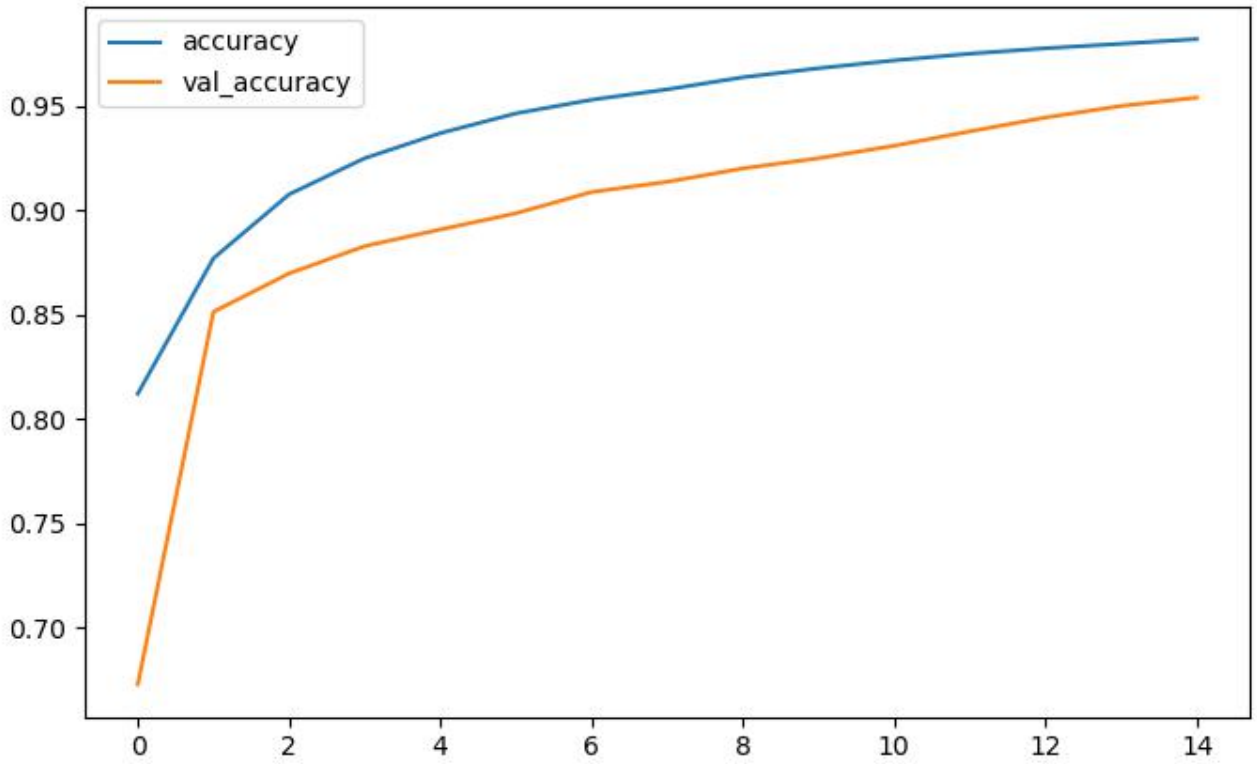

Figure: Training and Validation accuracy over 15 epochs.

First results : Normalized confusion matrix

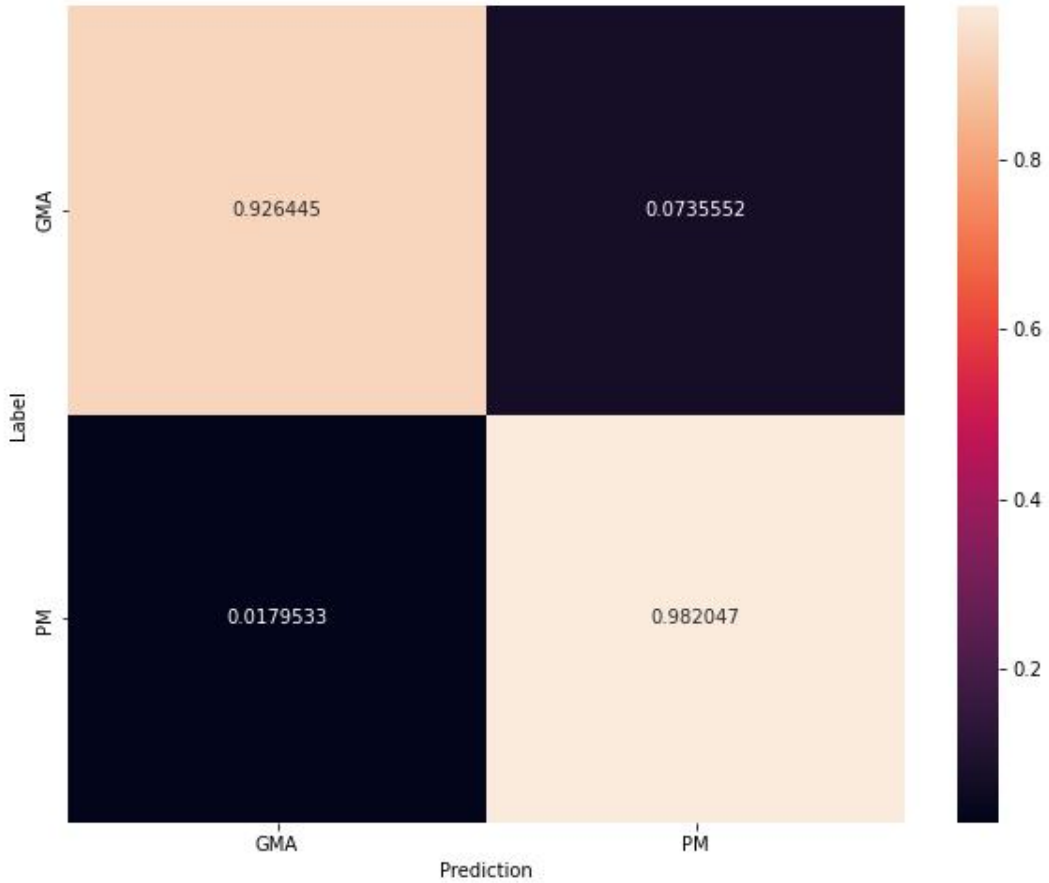

Figure: Normalized confusion matrix. GMA = Short finned pilot whale and PM = Sperm whale

Comparative study : Data for the test

- Data collected from the Zooniverse website.
- Selection of data/labels (predictions with an accuracy of at least 65 percent)

Data for the test : Audio data

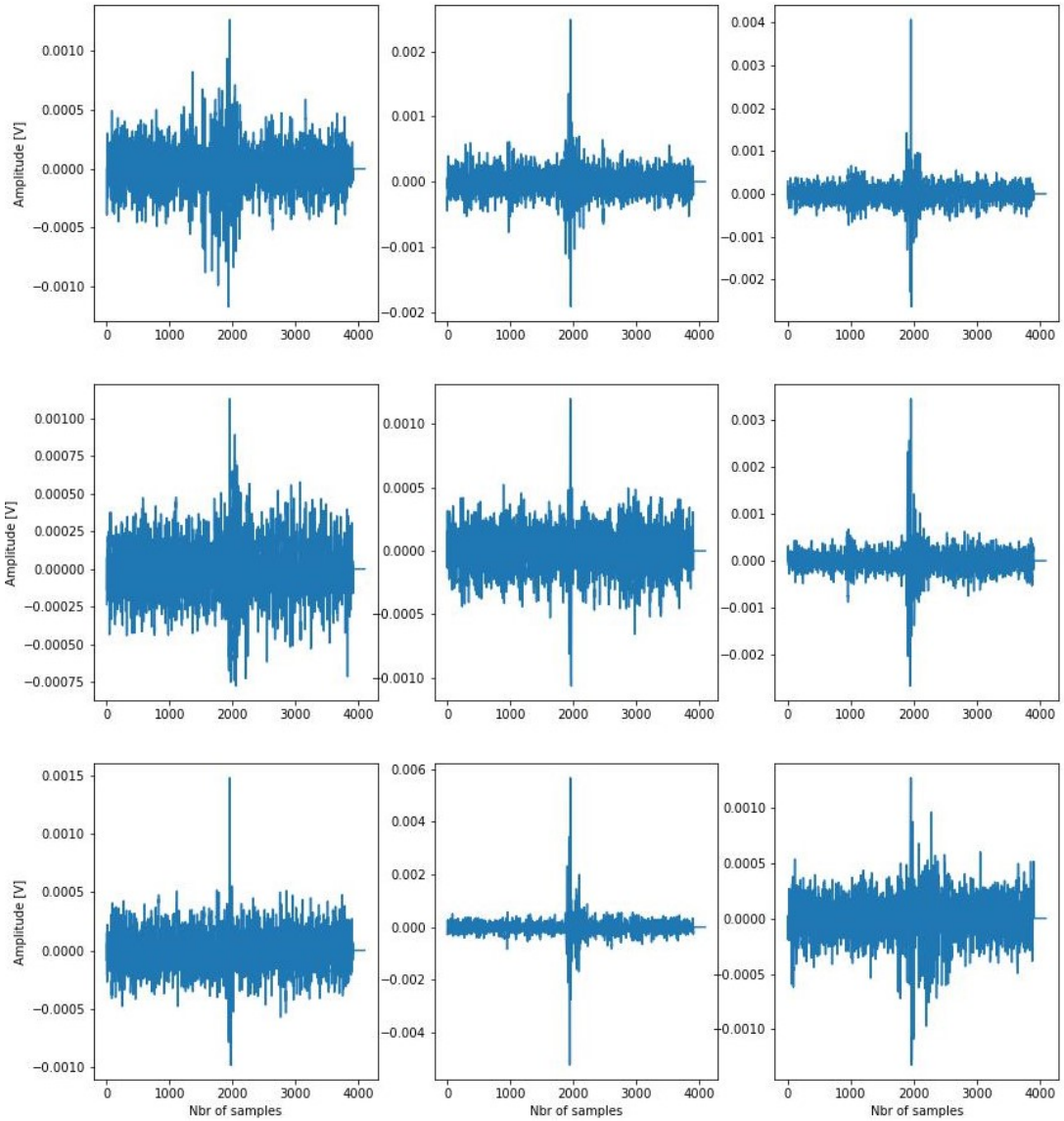

Figure: Example of data in the test set.

Final results of the comparative study

- Train the CNN and then make predictions on the test set.
- Get the confusion matrix to compare with the classifications of the citizens.
- See if the data collected from the citizens can actually help us in enhancing the accuracy of the neural network.