Proceedings of the Institute of Acoustics

A DEEP SEGMENTATION APPROACH FOR MULTIBEAM ECHO SOUNDER BACKSCATTER DATA BASED ON SEAFLOOR TYPE

H. Moreau	Lab-STICC UMR CNRS 6285, ENSTA Bretagne, Brest, France
S. Homrani	Lab-STICC UMR CNRS 6285, ENSTA Bretagne, Brest, France
I. Mopin	Lab-STICC UMR CNRS 6285, ENSTA Bretagne, Brest, France
J. Le Deunf	Service Hydrographique et Océanographique de la Marine, Brest, France
J. Bignon	Service Hydrographique et Océanographique de la Marine, Brest, France
G. Le Chenadec	Lab-STICC UMR CNRS 6285, ENSTA Bretagne, Brest, France

1 INTRODUCTION

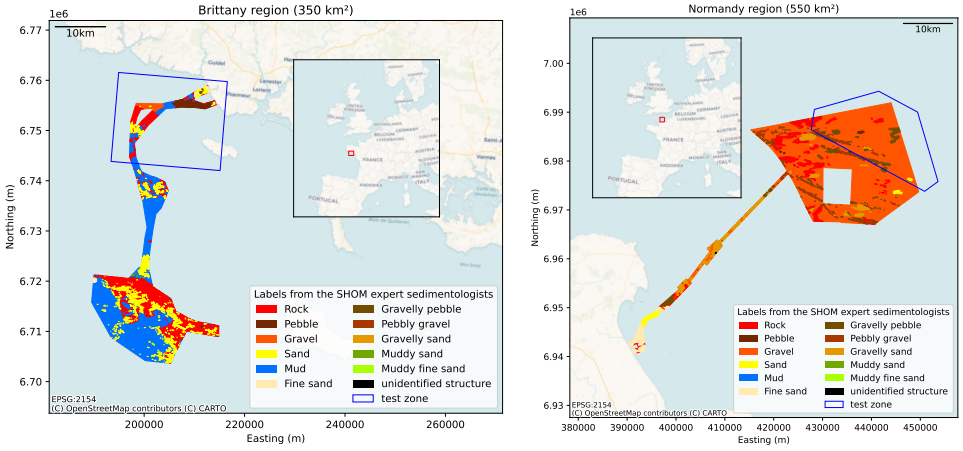
Characterising the seafloor has many applications, including geology, deep-sea mining, and benthic habitat habitat characterisation. To obtain large-scale information on the seabed, multi-beam echo-sounders (MBES) are often used. However, interpreting these sensor data is not trivial. Many publications tackle the problem of sediment type prediction using multibeam echosounders. Methods are diverse, including machine learning algorithms such as fully-connected neural networks[8], Random Forest[2,6] or explicit statistical modelization with Gaussian Mixture Models[7] or Conditional Random Fields[1]. They rely on a certain amount of handcrafted *features*, values which are representative of the problem to solve. When dealing with MBES measurements, features can be derived from soundings (depth, slope, roughness) or from the seabed acoustic responses, often called *backscattering strength* (BS), both measured simultaneously during a survey. However, the relationship between BS and sediment properties is complex[5] and despite recent advances[10], designing the right features to provide a machine learning model is not a solved problem.

One way to circumvent this hurdle is to use deep neural networks, a family of algorithms known for their ability to simultaneously produce features and classify the samples. However, deep neural networks, used as-is, require vast amounts of labelled data. A common way to obtain labels for the seafloor is to use either grab samplings, photo, or video evidence. These methods allow for a highly detailed characterisation of the seafloor, but they are limited to punctual measurements. In contrast, the areas covered by bathymetric or mosaic maps are much wider. This means that even by assigning a label to a set of nearby soundings[6], the amount of labelled data is bound to be much smaller than the number of soundings. One alternative to the problem chosen by Garone *et al*[4] is to use sedimentary maps drawn by experts (covering the whole study area) instead of using directly objective measurements like grab samplings. The authors trained a convolutional neural network to perform a two-class problem (bedrock/non-bedrock) using these labels. It resulted that the depth (along with its derivatives like the slope) was more efficient than the backscatter for a neural network to predict the seafloor sediment type from.

The present article aims at exploring the use of supervised deep neural networks to predict a seafloor type using backscatter and/or depth data provided by a MBES. We work on data organised as a [ping, beam] matrix. This publication is organised as follows: Section 2 describes the data, the area of study and the process employed to create the labels. Then, section 3 describes the practical implementation of the neural network, before section 4 presents the results. We conclude with some ways our work can be improved in Section 5.

2 DATA DESCRIPTION

The study area consists of two zones located off France, on the continental shelf in the east of the Atlantic Ocean and in the west of the English Channel (see Fig. 1). The survey was carried out by the French hydrographic and oceanographic service SHOM (*Service Hydrographique et Océanographique de la Marine*). Raw MBES data is a series of pings, measured at each position of the vessel. Each ping contains a fan of several beams containing one sounding and one backscatter value per beam. Even though the entire area was surveyed using multiple echosounders, we focus on the lines where the Kongsberg EM710 is used which represents 89.6% of the total number of soundings. With this MBES, data was recorded using a dual-swath mode, i.e. different frequencies are transmitted between central and lateral sectors and between swaths. They alternate between $70-90-70kHz$ (for the left, central, and right sectors, respectively) and $80-100-80kHz$. Even though the seafloor acoustic response is known to depend on transmitted frequency and incident angle of the acoustic transmitted pulse on the seafloor, we concatenate all data into a single image, and let the neural network learn to deal with this data. MBES data used are not calibrated in amplitude[3] which explains the sharp differences between transmit sectors on backscatter in figure 2.



(a) (b)

Figure 1: Maps of labels provided by expert sedimentologists of SHOM. We delimit two test zones to make sure the results are evaluated using data from a different location than the training data.

Because MBES raw data are directly used, outliers are present in soundings and backscatter measurements. Thus, before integrating data to the network, a filter on the depth of the soundings is applied: depths larger than $300\,\mathrm{m}$ (three times the maximal depth observed in both regions) are removed. As neural networks do not accept missing values, the sounding depth data is interpolated using a median filter applied in the [ping, beam] geometry. No filter is applied to backscatter measurements.

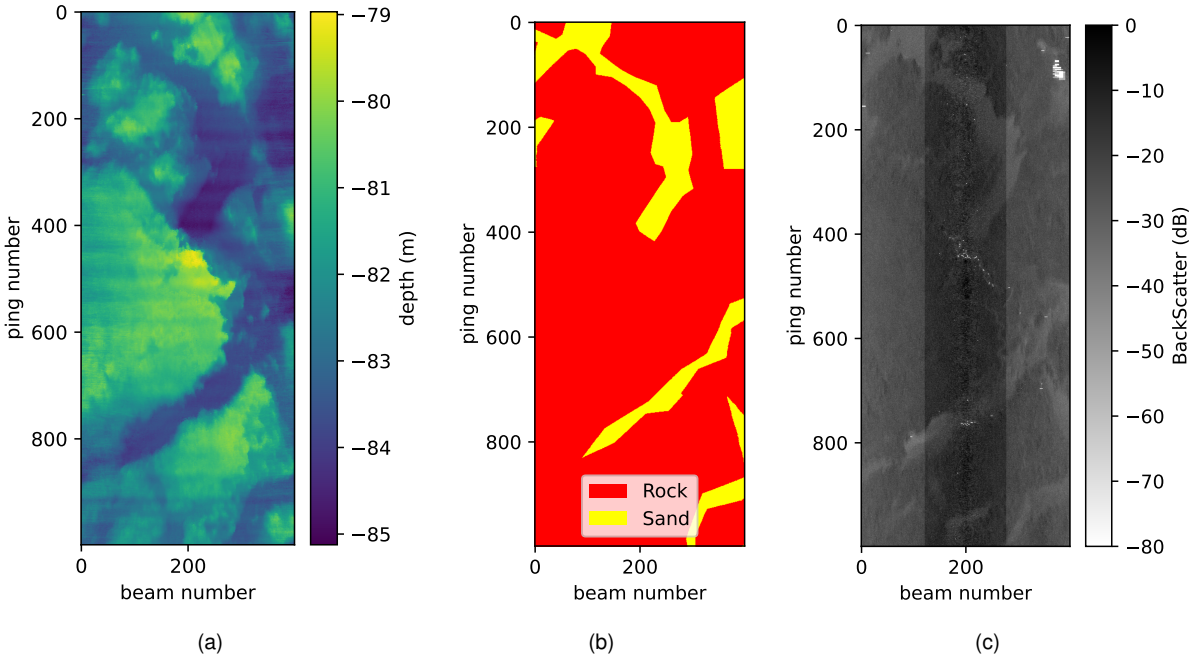


Figure 2: Example of raw data in [ping, beam] geometry: (a) sounding depth, (b) label, (c) and backscatter from a single survey line with Kongsberg EM710.

2.1 Projection of label georeference map in echosounder geometry

In order to label the data with sediment types, we use labels of expert sedimentologists shown in figure 1 as ground truth. These labels maps are derived following the SHOM usual procedure: 1) using complementary subbottom profiler acoustic data, rocky areas are first delimited, 2) a reflectivity map of the area is then produced using MBES data and the DTM validated by hydrographers, and 3) combining information from the reflectivity map, the DTM, grab samples and side-scan sonar data when available. A georeferenced map of the most likely sediment type is produced, which we show in Fig. 1. The sediment classification (see Table 1) corresponds to the SHOM nomenclature and is based on grain size. The proportions of different grain sizes yield mixture labels such as "Sand with Gravel". For instance, if a sediment sample is composed of at least half of grain sizes between 0.5 and 2 mm, and 25% or more of particles between 2 and 20 mm, it will be attributed the label "Sand with gravel". If, instead, there is a majority of sand and no other category of grain size accounts for more than 25% of the mass of the sample, it will bear the label "Sand". The complete inventory of the sediment classes that appear in our dataset is shown in the left-most column of Table 1.

Because we use raw MBES data as input to the network, labels should be given in the same geometry ([ping, beam]). To obtain a set of labelled lines, we use the position of each sounding and compare it to the georeferenced labels. The label polygon where the sounding impact is located indicates the class for this [ping, beam] pixel. An example of [ping, beam] label image resulting from this process is shown in figure 2.

fine-grained labels			merged labels		
Label	Proportion train+val	Proportion test	Label	Proportion train+val	Proportion test
Rock	8.3 %	9.8 %	Rock	13.8 %	10.9 %
Pebbles	0.02 %	3.8 %	Gravel & Pebbles	66.2 %	75.5 %
Pebbles with Gravel	1.1 %	2.9 %			
Gravel with Pebbles	1.1 %	2.4 %			
Gravel	37.9 %	58.2 %			
Sand with Gravel	2.8 %	0.2 %	Sand	6.7 %	2.4 %
Sand	1.0 %	1.9 %			
Sand with Mud	0.005 %	0 %			
Fine Sand	0.17 %	0 %			
Fine Sand with Mud	0.003 %	0 %			
Mud	8.0 %	10.0 %	Mud	13.2 %	11.1 %
unidentified structure	0.006 %	0 %	missing	/	/
missing	39.5 %	10.8 %			

Table 1: The proportion of each label before and after merging into five classes. The proportion is measured in number of soundings (the pixels of the survey lines).

2.2 Fusion of sediment categories

The left part of Table 1 shows the proportion of the number of soundings of each labels available in our dataset. We can observe that the labels are fine-grained, perhaps too detailed for a network to discriminate using backscatter data. This is why we use coarser labels: we merge all the classes containing gravel and/or pebbles into a single class named ”Gravel and pebbles”. Similarly, the absence of several labels from the test set pushes us to merge all classes containing a majority of sand into a single class we keep labelling ”Sand”. The right columns of Table 1 provide the resulting classification of this merge. These new classes are the objective the neural network is trained to predict, as described in the next section.

3 METHOD

3.1 Train and test split

Training a neural network requires to use different samples from the ones which serve to evaluate its performance on unseen data. Ideally, to ensure that the model’s performance is really representative of its generalisation on unseen contexts, one should select training and testing zones which are completely separated. However, in our case, there is a strong imbalance between the classes of each region (see Fig. 1). If we were to select one region for the training set and the other for the test set, these two subsets would not share the same classes. Hence, we choose to split the data by delimiting two sub-regions that we can easily split off the rest.

In order to find the best hyperparameters (learning rate, number and size of the convolution filters, *etc.*), we may not use the test set, but we use a *validation set*. This set is also distinct from the set of samples the neural network is trained with. Ideally, the separation of the validation from the training set should

	validation F1	test F1
backscatter	$72.5 \pm 1.1\%$	$51.8 \pm 1.5\%$
depth	$59.9 \pm 1.6\%$	$31.8 \pm 2.5\%$

Table 2: The average and standard deviation of models working with either the depth, the backscatter, computed over three runs.

follow the same rules as the creation of the test set. However, as we saw no other zone that could be separated from the rest without losing too much data, we decided against it. We create a validation set by sampling 20% of the survey lines at random. This means that the choice of hyperparameters may not be strictly the best when designing a model that generalises to other regions. However, as the performance of the model is computed on a test set rigorously distinct from the training and validation ones, the reported performances are quite representative of performances on unseen datasets. To summarize, we split the test set into train and validation sets using the zones in Figures 1(a) and 1(b), but the training and validation sets are obtained by sampling survey lines independently from each other.

3.2 Network architecture

Images are cut into patches of size 128×400 to feed the network and to fit into memory. During the training time, the network input is the whole beam, for a subset of 128 pings at once. We train the network to predict the type of sediment using an architecture inspired by U-net[9]. The network architecture consists of convolution layers with 7×7 filters. The number of filters is set to 64 for the first layer and doubles every time there is a downsampling operation. There are two downsampling steps (using average pooling), followed by upsampling operations using transpose convolutions. We train this network by updating the parameters using a CrossEntropy loss, where each class has been weighted proportionally to the inverse of the number of samples in the training set. We stop either after we ran through the training dataset 100 times or if the validation loss does not decrease after 20 iterations on the train set. We use a learning rate of 10^{-4}, with the Adam optimizer, with an L2 regularization parameter of 10^{-3}. We report the unweighted average of the per-class F1 scores, so that the metric is not biased in favour of the class with the most samples.

In order to measure the uncertainty, we repeat the training procedure three times (with three different random seeds) and provide the average and standard deviation of the scores and metrics we display.

4 RESULTS

Table 2 shows the results of the evaluation of the models using each type of input on the validation and test set. One may notice a severe performance drop between the validation and test sets. The difference between them comes from the fact that the validation set is located in the same region as the training set, despite using different survey lines. On the other hand, the test set is physically distinct from both the validation and training sets (see Fig. 1), ensuring that the performance on this set is fairly representative of results obtained on new zones.

To understand the reason for this difference, we also display the confusion matrix of one model in Figure 3. As this figure shows, the performance drop between the validation and test set is mainly due to the classes ”Pebbles and Gravel” and ”Sand”. The nature of the errors also changed between sets: for example, if the classes ”Pebbles and Gravel” and ”Mud” were fairly well-recognized in the validation set,

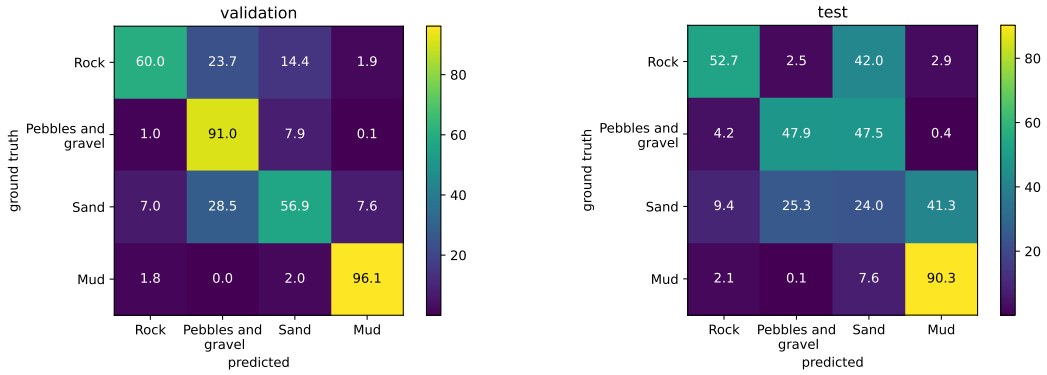


Figure 3: The confusion matrix of one model working on the backscatter, evaluated on the validation (a) and test (b) sets.

the class ”Pebbles and Gravel” is frequently confused with Sand in the test set. Similarly, the confusion of Sand for Mud is quite low in the validation set, while this error is frequent in the test set. Upon examining the exact labels of the test set (Table 1), we notice that in the training data, what constituted the label merged label ”Sand” is actually three quarters of ”Sand with gravel”, a category is almost absent from the test set. This means that the meaning of what we (and the model) call ”Sand” changed between the training and the test data. We hypothesize that this is the reason why most of the sand in the test set ended up being classified as the next class with lower grain size, ”Mud”.

We argue that removing the ’Sand and Gravel’ category is not a desirable solution. In fact, such discrepancies are the reason why separating the training region from the test region is paramount. For instance, there might be a situation where the test zone contains a region with a mix of a majority of sand and a minority of a sediment we did not see in the training data (for example, silt). A model that generalises correctly should, ideally, be able to keep calling this combination ”Sand”. Hence, this type of discrepancy is useful because it should be expected in a real environment. The performance mismatch is due to the fact that the classes in the training set are not representative of the types of seafloor we may find in any location. Finding solutions to this domain shift is one possible avenue for future work.

Table 2 also shows that using backscatter data reaches a better performance than using depth data, indicating that, in our specific area, there is more information related to the seabed type in the backscatter than in the bathymetric map. Garone *et al.*[4] obtained opposite results: they found that the depth fared better than the BS. There are several differences between the two series of works: the area of study is not the same, and the two networks are not equal. However, we believe that two of these differences are significant. Firstly, we use a [ping, beam] matrix that implicitly keeps the angular variations of the backscatter (which depends on the incidence angle of the transmitted pulse on the seafloor), whereas Garone *et al.*[4] use a reflectivity map where the influence of the angle of incidence is lost when data is aggregated. Secondly, we used four classes, while Garone *et al.* focused on distinguishing bedrock from non-bedrock. We believe that distinguishing between Sand, Mud, and Pebbles and Gravel from the depths (or the slopes) of the soundings will be harder than distinguishing the bedrock from the rest.

We select one model working with either the depth or the backscatter and examine their predictions in detail. Figure 4 shows the output of each network when asked to predict the seafloor type from data previously shown in Fig. 2. We may see that the network using the depth did not succeed in discrim-

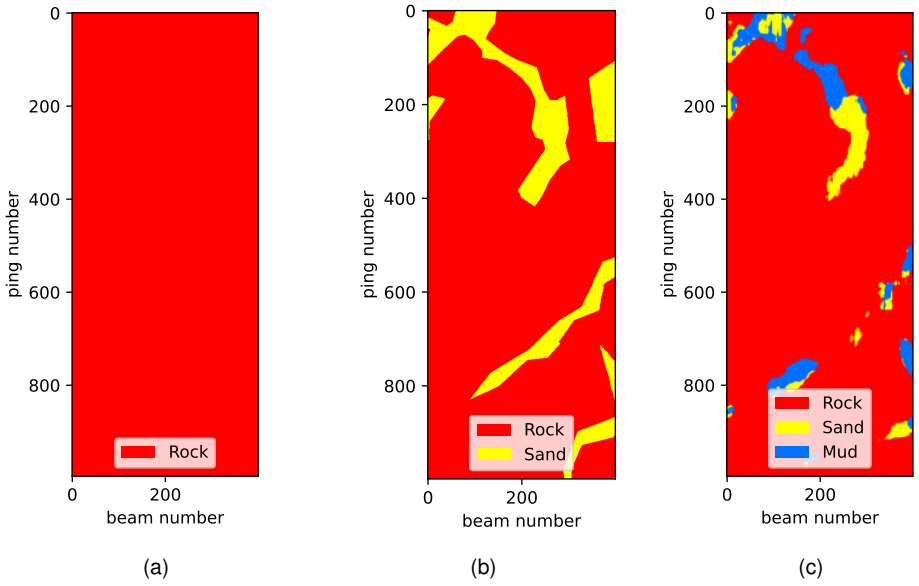


Figure 4: The predictions of the network using the depth (a) or BS (c), along with the labels from experts sedimentologists (b), for the survey line corresponding to Fig. 2. Contrary to others[4], we do not use any smoothing procedure.

inating the different sediment types. The network that used the backscatter succeeded to some extent in differentiating the rocks from the rest, but still confused sand for mud. Note that, despite working with non-calibrated data, the model working on the backscatter was not influenced in its predictions: it learnt to ignore the difference between sectors. This is likely due to the fact that all the data the network saw comes from a single model of sounder (Kongsberg EM710), meaning that the gain difference between sectors is constant across datasets.

5 CONCLUSION

This study aimed at experimenting the prediction of seabed sediment type using a neural network on acoustic backscatter and bathymetric data from an MBES. We relied on sedimentary maps from expert sedimentologists to assign a label to each sounding. This volume of labelled data was used to train a U-net neural network to predict a label map. Results showed a performance drop-off when the network had to generalise to a location it was not trained on, which underlines the need for a test set which is well separated from the training set. In addition, the network did not seem to be impaired by the use of non-calibrated backscatter data in our specific example where only one type of MBES is employed. In addition, using backscatter data to infer sediment type seems to be more successful in training a network than depth data.

There are several avenues for future work. Firstly, we worked exclusively on data that has not been calibrated, thus we hope that using calibrated data to train a network yields a model which generalises better. Another possibility for future works is to better use our labels: in all the present publication, we considered the sedimentary map as a ground truth, to which we compared the prediction of the network. Most of the

literature prefers to work on objective, unquestionable labels, such as grab samplings. Although this limits considerably the amount of labels, we could propose to work in semi-supervised setting, or even use the sedimentary map in a weakly-supervised fashion.

ACKNOWLEDGMENTS

The authors would like to thank the French Service d'Hydrographie et d'Océanograhie de la Marine (SHOM) for collecting the data, with the support of the French General Direction of the Energy and Climate (Direction Générale de l'Énergie et du Climat, DGEC).

REFERENCES

1. D. Buscombe and P. E. Grams. Probabilistic Substrate Classification with Multispectral Acoustic Backscatter: A Comparison of Discriminative and Generative Models. *Geosciences*, 8(11):395, Nov. 2018. Number: 11 Publisher: Multidisciplinary Digital Publishing Institute.
2. M. Diesing, S. L. Green, D. Stephens, R. M. Lark, H. A. Stewart, and D. Dove. Mapping seabed sediments: Comparison of manual, geostatistical, object-based image analysis and machine learning approaches. *Continental Shelf Research*, 84:107–119, Aug. 2014.
3. D. Eleftherakis, L. Berger, N. Le Bouffant, A. Pacault, J.-M. Augustin, and X. Lurton. Backscatter calibration of high-frequency multibeam echosounder using a reference single-beam system, on natural seafloor. *Marine Geophysical Research*, 39(1):55–73, June 2018.
4. R. V. Garone, T. I. Birkenes Lønmo, A. C. G. Schimel, M. Diesing, T. Thorsnes, and L. Løvstakken. Seabed classification of multibeam echosounder data into bedrock/non-bedrock using deep learning. *Frontiers in Earth Science*, 11, 2023.
5. D. R. Jackson, D. P. Winebrenner, and A. Ishimaru. Application of the composite roughness model to high-frequency bottom backscattering. *The Journal of the Acoustical Society of America*, 79(5):1410–1422, May 1986.
6. X. Ji, B. Yang, and Q. Tang. Seabed sediment classification using multibeam backscatter data based on the selecting optimal random forest model. *Applied Acoustics*, 167:107387, Oct. 2020.
7. L. Koop, A. Amiri-Simkooei, K. J. van der Reijden, S. O'Flynn, M. Snellen, and D. G. Simons. Seafloor Classification in a Sand Wave Environment on the Dutch Continental Shelf Using Multibeam Echosounder Backscatter Data. *Geosciences*, 9(3):142, Mar. 2019. Number: 3 Publisher: Multidisciplinary Digital Publishing Institute.
8. V. Ntouskos, P. Mertikas, A. Mallios, and K. Karantzalos. Seabed Classification From Multispectral Multibeam Data. *IEEE Journal of Oceanic Engineering*, 48(3):874–887, July 2023.
9. O. Ronneberger, P. Fischer, and T. Brox. U-Net: Convolutional Networks for Biomedical Image Segmentation. In N. Navab, J. Hornegger, W. M. Wells, and A. F. Frangi, editors, *Medical Image Computing and Computer-Assisted Intervention – MICCAI 2015*, Lecture Notes in Computer Science, pages 234–241, Cham, 2015. Springer International Publishing.
10. T. Zhao, G. Montereale Gavazzi, S. Lazendić, Y. Zhao, and A. Pižurica. Acoustic Seafloor Classification Using the Weyl Transform of Multibeam Echosounder Backscatter Mosaic. *Remote Sensing*, 13(9):1760, Jan. 2021. Number: 9 Publisher: Multidisciplinary Digital Publishing Institute.