A semi-supervised learning approach to polymetallic nodule parameter modeling

Hari Vishnu *, Vincent Robert †, Bharath Kalyan *, and Mandar Chitre *
* Keppel-NUS Corporate Laboratory and Acoustic Research Laboratory,
National University of Singapore
e-mail: {hari, bharath, mandar}@arl.nus.edu.sg
† Telecom Paristech, e-mail: vincent.robert@telecom-paristech.fr

***Abstract*—We explore spatial modeling of polymetallic nodule parameters in the Clarion-Clipperton zone (CCZ) using a semi-supervised learning approach. Spatial models that utilize the factors affecting nodule formation or their proxy variables, are a useful tool for characterizing the CCZ nodule deposits and their economic potential. Some such models based on neural networks have been explored in the literature, but these approaches have employed supervised learning. These rely on hand-modeled features to incorporate the effect of topography, one of the key factors affecting nodule formation. We employ unsupervised learning via auto-encoders or principal component analysis to efficiently capture the effect of local topography around a point being modeled, and express it in terms of a few features. We then use these features for subsequent supervised learning-based nodule parameter modeling. Thus, this is overall a semi-supervised approach. We show that the efficient incorporation of bathymetric information into these features yields better modeling performance as compared to using hand-modeled topographic features.**

***Keywords*—polymetallic nodules, Clarion Clipperton zone, artificial neural network, semi-supervised learning, auto-encoder**

I. Introduction

The search for metal resources to meet industry demands is now slowly turning towards the oceans. In the northeast Pacific, the region between the Clarion and Clipperton fracture zones named Clarion-Clipperton zone (CCZ) offers hope in the form of polymetallic nodule (PMN) deposits [1]. These are found on the seabed at depths of several kilometers, and have been identified as a significant untapped resource for metals such as manganese, cobalt, nickel (Ni), lithium, titanium and even rare-earth metals [2], [3].

Though the CCZ is a vast area [1], the PMN deposits here have only been sparsely explored. In order to gauge the resources here for exploitation purposes, we require an approach that can spatially model the resource potential using the sparse data available [1], [4]. This can be done by going beyond interpolation of known nodule parameter data, and developing a model that incorporates information on the mechanism of nodules formation. This includes the effect of factors such as local topography, sediment, surface net primary productivity (NPP) and distance from the metal sources [1]. Topography has been shown to be a crucial factor in determining the quantity and quality of PMN at a site [5], [6]

Spatial modeling has been explored earlier in the literature [1], [4], [7], [8]. A report by the International seabed authority [1] presented a supervised learning (SL) based model using an artificial neural network (ANN) [9], which is a powerful tool for characterizing underlying variations encountered in data. It used topography as a feature by classifying it into categories such as abyssal seamounts and ridges. An approach introduced later represented the local topography in terms of numerical quantities representing the depth and depth gradients at the location being modeled [4]. In a follow-up work to this, we discussed the uncertainty associated with predictions made using this model [10].

SL-based modeling of the nodule parameters at a given location could be more effective if the bathymetric information available around that location is efficiently used as features. However, this bathymetric data consists of many points and hence would constitute a feature set of large dimensions if used in a raw form. An SL-based model cannot be satisfactorily trained with so many features using the scarce amount of labeled nodule parameter data available, which is why hand-modeled features were used in previous works to characterize topography. However, depth gradients capture the information on topographic variations only in a few directions, whereas there may be more uncaptured variations in other directions around the point of interest. Also, depth gradients can only capture information on relatively smooth features that are sufficiently described by gradients upto order 2. These are some limitations of using the gradient-based method.

An approach gaining attention in the deep learning literature is to shift from using hand-modeled features, to features modeled via unsupervised learning (UL) techniques instead. When abundant data is available on the raw features we wish to represent (in this case, topography), these can be compressed via UL into a fewer number of efficient features for use in SL. In this semi-SL approach, UL is used as a prior step in order to extract features via dimensionality-reduction.UL techniques such as principal component analysis (PCA) have conventionally been used for dimensionality-reduction in the machine learning literature [11]. Recent advancements in ANN literature have shown that UL architectures such as autoencoders (AE) are able to outperform PCA and yield even better performance in dimensionality-reduction [11], [12].

We present an ANN-based approach to spatial modeling of nodule parameters which does not involve computation of depth gradients. Instead, we model the bathymetry around the point of interest by using UL, and compress the bathymetric information into a smaller number of features. We then use these features as inputs in SL modeling of the percentage of nickel by weight (Ni %), using an approach similar to ones

presented earlier [1], [4]. Ours is overall a semi-SL approach, described in the schematic in Fig.1 (b), and we compare it to the (a) purely SL approach presented in earlier works.

The paper is organized as follows. In section II, we discuss the features used for spatial modeling of the nodule parameters. In section III, we discuss how UL can be used for topography characterization and feature extraction. In section IV, we discuss the results obtained from our semi-SL approach based on this feature extraction, and finally conclude the paper in section V.

II. Features used for modeling

PMN are formed in deep-sea regions by entry of metal ions through a combination of diagenetic and hydrogenetic processes [13] around a nucleus. The metals reach the sea floor through a combination of terrigenous runoff, and biological and volcanogenic activity. Nodule formation at a location is affected by its depth, local topography, sedimentation rate, and biological activity at the sea surface by plankton which leads to supply of metals to the sea floor [6], [14], [15].

The features selected by us for SL modeling include the bathymetry and topography around the region being modeled, which will be discussed in detail in the next section. The other features we use for the modeling are discussed in the following. Table I lists out the open-domain sources from where we obtained data for the features and the nodule parameters modeled.

1) Distances from east Pacific rise and west coast of America: These indicate the distances of the location being modeled from the hypothesized sources of the metals. The metals from these sources undergo some dilution due to spreading as we move away from the sources [1]. We model this spreading as impacting the nodule parameters in a monotonic way. We follow the approach in [10] and force our ANN to learn a dependence of the form $c_i.t_i^{p_i}$, where c_i is a constant, $p_i < 0$ is a spreading-law exponent and t_i is the minimum distance from the source considered. The subscript i can take value of 1 indicating distance from the West coast of America, and a value of 2 indicating the distance from the east Pacific rise.

2) Net primary productivity: Net primary productivity (NPP) indicates the level of biological activity at the sea surface that contributes metals for nodule formation. It is defined as the flux rate of carbon in plankton due to biological processes [16]. NPP features for modeling were extracted from a carbon based productivity model [16].

3) Sediment type: We characterize the sediment-type at the location being modeled in terms of fractions of terrigenous material, pelagic clay, siliceous sediment and calcareous sediment. We assume that the sediment at any place can be fully characterized by these four types of content [4]. A limited amount of low-resolution data on sediment-type in CCZ is available on some online repositories, as mentioned in Table I. We interpret the information from these sources in terms of these 4 categories, and interpolate this information for use in regions where this data is unavailable.

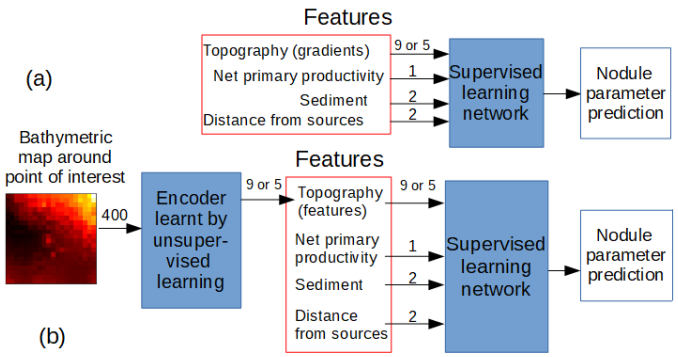

Fig. 1. Schematic showing (a) purely supervised approach used in previous works, (b) semi-supervised learning approach proposed by us. The numbers over the arrows connecting to the SL block indicate the number of features used in our SL modeling corresponding to each factor, such as topography or sediment type. The number over the arrow connecting to the encoder indicates the features used in the UL, and the one at its output indicates the number of features this information is compressed into (9 or 5).

III. Unsupervised learning for bathymetry characterization

We use UL to obtain a small number of features representing the bathymetric variation around a point. In our approach, we assume that the topographic variations affecting nodule genesis at a location are located within a 10 km $\times$ 10 km area surrounding it. Bathymetric information of any resolution available within this region is interpolated into a grid of 20×20 pixels. We aim to reduce the dimensionality of this grid (viz. 400) into a small set of features. One challenge in training a large ANN with many features is that it requires a large amount of data to yield effective results. Often, sufficient labeled data may not be available for this. For example, in our scenario we have obtained from the open domain a few hundred labeled data points on Ni % in CCZ nodules. This is not enough for learning a good dimensionality-reduction procedure for the 20×20 grid representing the surrounding bathymetry. However, we possess a large amount of unlabeled data on the features, viz. bathymetric information for the entire CCZ obtained from the open domain. We assume that a UL technique that can characterize bathymetric information from any grid within CCZ, would be useful in generating features for the SL for nodule parameter modeling within CCZ. We use this data to learn a good dimensionality-reduction encoder using UL.

A widely known method of dimensionality-reduction is PCA, which finds directions of greatest variance in the data, and represents each data point by its coordinates along each of these directions [11]. PCA's limitation comes from its linearity, which allows representation of data only via linear projections onto a few reduced dimensions. Recently, a nonlinear generalization of PCA using ANN known as an AE has been developed, which allows efficient representation of data that lies on a lower-dimensional non-linear manifold [11].

An autoencoder is an ANN that aims to reproduce its input at its output. When used for dimensionality-reduction, an AE consists of two building blocks: an encoder and a decoder. The encoder accepts a large-dimensional vector at its input and has a few output units. The decoder accepts these units

TABLE I. SOURCES OF DATA ON FEATURES OBTAINED FROM OPEN-DOMAIN

Feature	**Sources**
Bathymetry	General bathymetric grid of the ocean [17], [18], Online databases at NCEI [19]
Sediment-type and Ni %	Online databases at NCEI [19], primarily the Seadas database, consisting of: The NOAA and MMS Marine Minerals Bibliography, Archive of Core and Site/Hole Data and Photographs from the Ocean Drilling Program, NOAA/NOS and USCGS Seabed Descriptions from Hydrographic Surveys, Index to Marine and Lacustrine Geological Samples, Archive of Core and Site/Hole Data and Photographs from the Integrated Ocean Drilling Program, Archive of Core and Site/Hole Data and Photographs from the Deep Sea Drilling Project and ISA Central Data Repository [20]
NPP (mg C/m^2/day)	Oregon State University [21]

as input and outputs a vector of the same dimensions as the encoder's input. When the AE is trained to match its output and input, the bottleneck of few units bridging the encoder and decoder known as the latent space, holds a compressed representation of information. Once this is done, the encoder part of the trained AE can be used to effectively perform dimensionality-reduction on large-dimensional input data of similar type and dimensions. We utilize an AE architecture known as a convolutional AE [22]. This architecture taps into the power of convolutional neural networks in order to perform dimensionality-reduction [22]. Convolutional neural networks have been shown to be effective ANN architectures for image processing and characterization applications.

Previous works on ANN-based spatial modeling [4], [10] utilize a depth-gradient approach to characterize the topography in terms of a vector of 5 components: depth, and first and second order gradients in the cardinal directions. We can expand this feature set to 9 by including the first and second order gradients in the inter-cardinal directions, in order to try to capture more topographic information around the point.

We use a convolutional AE with the encoder consisting of 3 convolutional layers using filters of size 3×3 and of stride 1×1, each followed by max-pooling layers with similar sizes and stride except for the last one where the stride is of 2×2.[1] The convolutional layers are followed by 2 fully connected layers consisting of 4000 units each. The latent space is set to be of dimension 5 or 9, so that the feature size is comparable to that used in the depth-gradient approach. The decoder's architecture is the reverse of the encoder's. The overall architecture is described in Fig. 2. We use exponential linear unit activation functions [23] throughout the network, except for the last layer where a sigmoid activation is used. We use the Xavier initializer for the weights and the Adam optimizer for training [24]. Noise is added to the input images in order to prevent overfitting by the AE. The AE and PCA are trained using 119500 data samples consisting of 10 km$\times$10 km windows from around the CCZ. The validation set consists of 6635 samples, which are used to select the network and meta-parameters to use at the end of the training stage.

The cost functions that we select for the AE training are (1) maximum-absolute error and (2) mean-square error in reconstruction between the original and reconstructed bathymetric image pixels, averaged over all the images in the dataset considered. The maximum-absolute error is indicative of the worst case image reconstruction performance, whereas the mean-square error gauges the average-case performance. Note that PCA is designed to minimize the mean-square error using a linear projection, whereas for the AE we can choose any cost function. This versatility in choosing a wider range of cost functions to minimize apart from mean-square error alone, is one of the advantages of using an AE in addition to its nonlinearity. However, it is not obvious which of these metrics must be chosen in the UL stage in order to improve the subsequent SL modeling performance.

[1]We refer the reader to [22] for an understanding of the ANN terminology used here.

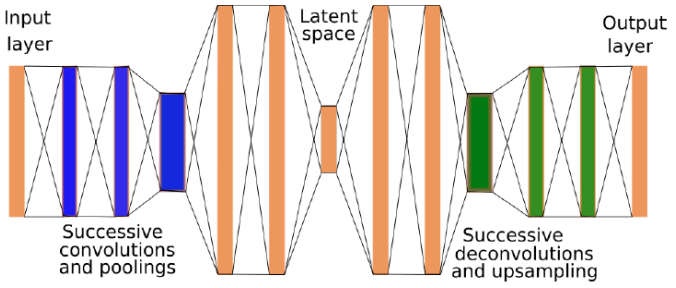

Fig. 2. Structure of the convolutional AE used. Blue blocks denote the convolutional layers in the encoder, green ones denote those used in the decoder, and orange blocks denote fully connected layers.

In Fig. 3 we plot images showing a comparison between the bathymetric variations around the location 12.25°N, 120.15°W, and a map reconstructed from it by PCA and AE, both using 9 principal components/latent units into which the information is encoded. Two AEs are considered: one trained using the mean-square-error cost function, and one using the maximum-absolute error. The reconstructed maps are all fairly similar to the original one, suggesting that both PCA and AE are able to encode enough information into the 9 features which captures the surrounding bathymetry fairly well. In the case of this particular image, the reconstructions provided by the two AEs seem to better capture some of the finer variations from the original.

IV. RESULTS

We explore the effectiveness of different types of topographic features in predicting nodule parameters via the earlier outlined SL approaches [4]. We model the conditional Ni % in PMN found in the CCZ, and compare the performance of our semi-SL approach against the existing purely SL framework. Conditional Ni % is defined as percentage of Ni by weight in the nodules conditional on the presence of nodules at a location. This is a useful indicator in determining commercial feasibility of extracting this metal via nodule harvesting at a given site. Ni in nodule deposits is considered to be of economic interest [2].

We perform modeling of the Ni % using a 2-layer feedforward SL network, with the feedforward architecture described in previous works [4]. Our network consists of 2 hidden layers

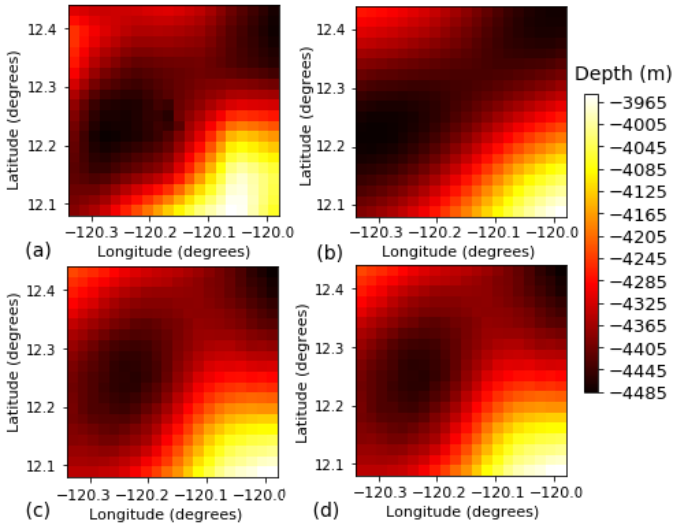

Fig. 3. Bathymetric maps around the location 12.25°N, 120.15°W: (a) original bathymetric map; encoded and reconstructed maps using (b) PCA, (c) AE trained based on mean-square error, and (d) AE trained based on maximum-absolute error.

TABLE II. COMPARISON OF NODULE-PARAMETER MODELING PERFORMANCE USING DIFFERENT APPROACHES TO CHARACTERIZE TOPOGRAPHY

Technique for characterizing topographic features	**Number of features**	**Regression coefficient on test data**
Autoencoder trained for MSE	9	0.562
Autoencoder trained for MSE	5	0.534
Autoencoder trained for MAE	9	0.517
PCA	9	0.533
PCA	5	0.531
Gradient features	9	0.510
Gradient features	5	0.525

with 2000 and 400 hidden units respectively, and one output neuron. We use rectified linear unit activation functions, Xavier initialization of the weights, and use the Adam optimizer for training. We perform training using 70 % of the available labeled data points. The validation dataset comprising of 15 % of the total data is used for selection of the final network to use at the end of the training stage. We impose L1 regularization of the weights, and the regularization factor is determined via a meta-parameter search across many candidate values. The topography features (5 or 9 in number) may be in the form of gradients or the encoded features obtained from PCA/AE. The number of input features is indicated in the schematic in Fig. 1. We compare the effectiveness of the SL and semi-SL approaches in terms of their regression coefficient in table II. This indicates the correlation of the trends in their predictions with the true nodule parameter values from a held-out test dataset. This test dataset comprises 15 % of the total dataset.

We find that the semi-SL approaches presented here are able to outperform the purely SL method. Semi-SL modeling using the AE trained for mean-square error with 9 features performs the best amongst the approaches compared here. It outperforms SL using 9 gradient features by 10.2 % and yields a correlation coefficient of 0.562. PCA-based modeling using 9 features outperforms the 9-feature gradient-based modeling by 4.5 %, whereas the AE trained for maximum-absolute error performs only marginally better than the latter (1.37 %). Thus, the maximum-absolute error is not a suitable metric to choose in the unsupervised learning stage in this scenario, as it does not translate into significant gains during the SL modeling stage. The AE-based modeling shows improvement when the feature set is expanded, whereas PCA does not see much improvement. The gradient-based method degrades when this is done, possibly due to overfitting. Our results demonstrate the superiority of the semi-SL approach in enhancing the performance of spatial modeling of Ni %.

V. CONCLUSIONS

We outlined a semi-supervised learning based approach using an artificial neural network for spatial modeling of nodule parameters in the CCZ. This was built on existing work on spatial modeling, and made use of unsupervised learning to learn features representing the local topography. These features were then used in a supervised learning to model nodule parameters using labeled data.

We showed that it is possible to achieve better performance using the semi-supervised learning approach presented here, than with the purely supervised approach using hand-modeled gradient features. An unsupervised learning metric based on mean-square error works better than a maximum-absolute error metric in producing useful topographic features from bathymetric data. In the future, we can explore improvements in this approach by trying more metrics in the unsupervised learning. We can also attempt to boost the performance by following up the semi-supervised learning with a final end-to-end training step, where the AE/PCA network and the feedforward network are connected and supervised learning is performed to fine-tune the performance.

VI. ACKNOWLEDGMENT

The authors thank the National Research Foundation, National University of Singapore and Keppel Corporation for supporting this work done in the Keppel-NUS Corporate Laboratory. The conclusions put forward reflect the views of the authors alone, and not necessarily those of the institutions within the Corporate Laboratory. The authors would also like to thank Mr. Wong Liang Jie for his help in collating some of the data used in this paper.

REFERENCES

[1] ISA, "A geological model of polymetallic nodule deposits in the Clarion-Clipperton Fracture Zone," ISA, Tech. Rep., 2010. [Online]. Available: https://www.isa.org.jm/sites/default/files/files/documents/tstudy6.pdf

[2] J. Hein, K. Mizell, A. Koschinsky, and T. Conrad, "Deep-ocean mineral deposits as a source of critical metals for high- and green-technology applications: Comparison with land-based resources," *Ore Geology Reviews*, vol. 51, pp. 1–14, jun 2013.

[3] J. R. Hein, F. Spinardi, N. Okamoto, K. Mizell, D. Thorburn, and A. Tawake, "Critical metals in manganese nodules from the Cook Islands EEZ, abundances and distributions," *Ore Geology Reviews*, vol. 68, pp. 97–116, jul 2015.

[4] V. N. Hari, B. Kalyan, M. Chitre, and V. Ganesan, "Spatial Modeling of Deep-Sea Ferromanganese Nodules With Limited Data Using Neural Networks," *IEEE J. Ocean. Engg.*, pp. 1–18, 2017.

[5] A. Usui, A. Nishimura, M. Tanahashi, and S. Terashima, "Local variability of manganese nodule facies on small abyssal hills of the Central Pacific Basin," *Marine Geology*, vol. 74, no. 3-4, pp. 237–275, feb 1987.

[6] Y. Ko, S. Lee, J. Kim, K.-H. Kim, and M.-S. Jung, "Relationship between Mn nodule abundance and other geological factors in the northeastern pacific: Application of GIS and probability method," *Ocean Science Journal*, vol. 41, no. 3, pp. 149–161, sep 2006.

[7] A. Knobloch, T. Kuhn, C. Rühlemann, T. Hertwig, K.-O. Zeissler, and S. Noack, "Predictive Mapping of the Nodule Abundance and Mineral Resource Estimation in the Clarion-Clipperton Zone Using Artificial Neural Networks and Classical Geostatistical Methods," in *Deep-Sea Mining*, R. Sharma, Ed. Gewerbestrasse: Springer, 2017, ch. 6, pp. 189–212.

[8] R. Sharma, N. Khadge, and S. Jai Sankar, "Assessing the distribution and abundance of seabed minerals from seafloor photographic data in the Central Indian Ocean Basin," *International Journal of Remote Sensing*, vol. 34, no. 5, pp. 1691–1706, mar 2013.

[9] J. Schmidhuber, "Deep Learning in neural networks: An overview," *Neural Networks*, vol. 61, pp. 85–117, 2015.

[10] H. Vishnu, B. Kalyan, and M. Chitre, "Spatial modeling and uncertainty characterization of polymetallic nodules in the Clarion-Clipperton zone," in *Oceans 2017, Anchorage*, Anchorage, 2017, pp. 1–6.

[11] Hinton, G and R. Salakhutdinov, "Reducing the Dimensionality of Data with Neural Networks," *Science*, vol. 313, no. July, pp. 504–507, 2006.

[12] G. Hinton, L. Deng, D. Yu, G. Dahl, A.-R. Mohamed, N. Jaitly, V. Vanhoucke, P. Nguyen, T. Sainath, and B. Kingsbury, "Deep Neural Networks for Acoustic Modeling in Speech Recognition," *IEEE Signal Processing Magazine*, vol. 29, no. 6, pp. 82–97, 2012.

[13] P. Halbach, "Processes controlling the heavy metal distribution in Pacific ferromanganese nodules and crusts," *Geologische Rundschau*, vol. 75, no. 1, pp. 235–247, 1986.

[14] R. A. Kotlinski, "Relationships Between Nodule Genesis And Topography In The Eastern Area Of The C-C region - Meeting of Scientists for the Preparation of a Programme of Work for the Development of a Geological Model of the Clarion-Clipperton Fracture Zone," Nadi, Fiji, Tech. Rep., 2003.

[15] K. Mewes, J. M. Mogollón, A. Picard, C. Rühlemann, T. Kuhn, K. Nöthen, and S. Kasten, "Impact of depositional and biogeochemical processes on small scale variations in nodule abundance in the Clarion-Clipperton Fracture Zone," *Deep-Sea Research Part I: Oceanographic Research Papers*, vol. 91, pp. 125–141, 2014.

[16] M. Behrenfeld and P. Falkowski, "Photosynthetic rates derived from satellite-based chlorophyll concentration," *Limnology and Oceanography*, vol. 42, pp. 1–20, 1997.

[17] General Bathymetric Chart of the Oceans website. GEBCO 30 arc-second grid. [Online]. Available: http://www.gebco.net/data_and_products/gridded_bathymetry_data/gebco_30_second_grid/

[18] P. Weatherall, K. Marks, M. Jakobsson, T. Schmitt, S. Tani, J. Arndt, M. Rovere, D. Chayes, V. Ferrini, and R. Wigley, "A new digital bathymetric model of the world's oceans," *Earth and Space Science*, vol. 2, no. 8, pp. 331–345, aug 2015. [Online]. Available: http://doi.wiley.com/10.1002/2015EA000107

[19] NOAA National Centers for Environmental Information. Main page. [Online]. Available: http://www.ngdc.noaa.gov/

[20] International Seabed Authority. Central data repository. [Online]. Available: https://www.isa.org.jm/central-data-repository

[21] (2015, Nov.) Ocean Productivity. [Online]. Available: http://www.science.oregonstate.edu/ocean.productivity/index.php

[22] J. Masci, U. Meier, D. Cirean, and J. Schmidhuber, "Stacked convolutional auto-encoders for hierarchical feature extraction," *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, vol. 6791 LNCS, no. PART 1, pp. 52–59, 2011.

[23] D.-A. Clevert, T. Unterthiner, and S. Hochreiter, "Fast and Accurate Deep Network Learning by Exponential Linear Units (ELUs)," in *International Conference on Learning Representations*, Puerto Rico, 2016, pp. 1–14. [Online]. Available: http://arxiv.org/abs/1511.07289

[24] D. Kingma and J. Ba, "Adam: A Method for Stochastic Optimization," pp. 1–15, 2014. [Online]. Available: http://arxiv.org/abs/1412.6980