A Principled Benchmark for Seismic Data Segmentation

Gabriel L. Canguçu, Leonardo M. S. Jorge, Gabriela T. Barreto, Thales H. Silva, Luiz A. Lima, Walace S. Caldas, Carlos G. S. Tavares, William Robson Schwartz, Pedro O. S. Vaz-de-Melo, Alexei M. C. Machado

Abstract—**In recent years, several deep learning techniques have been applied to the problem of seismic facies segmentation. However, there is a lack of an authoritative protocol for evaluating such models, so that the comparison between results becomes compromised. This paper proposes a principled benchmark for lithofacies segmentation based on the public seismic volumes from the F3 Netherlands, Penobscot and Parihaka datasets, along with standard metrics for performance assessment. The utility of the benchmark is illustrated by the evaluation of the U-Net DeconvNet and SegNet encoder-decoder architectures. The goal is to offer a framework which will enable researchers to compare different methods and to develop more effective strategies for the segmentation problem.**

Keywords: seismic segmentation, deep learning, benchmark, lithofacies.

I. Introduction

Seismic data segmentation plays a pivotal role in the field of geophysics, enabling the identification and delineation of subsurface structures that are critical for resource exploration, hazard assessment, and infrastructure planning. Over the years, advances in deep learning techniques have shown promising results in automating seismic data segmentation tasks. However, as the scientific community strives to harness the full potential of these cutting-edge technologies, a pressing need arises for standardized evaluation metrics and protocols that can effectively compare and benchmark performance across different datasets and neural network architectures.

The motivation for such standardization stems from the increasing diversity of seismic datasets, each with unique challenges and characteristics. These data sets cover a wide range of acquisition geometries, sensor types, geological contexts, and levels of data quality. Furthermore, the continuous evolution of neural network architectures, loss functions, and training strategies demands a unified framework for assessing their efficacy and robustness in the domain of seismic data segmentation.

Advancements in deep learning within the field of geophysics are emerging rapidly. One significant challenge associated with this rapid progress is the difficulty for the community to validate and assess the usefulness and correctness of each new contribution, especially in fields where methods are validated using different experimental setups. A natural solution to this problem is the design of unified benchmarks that ensure that all methods are evaluated under consistent conditions.

In fact, in a recent review paper, Bergen *et al.* [1] argued that accelerating the use of machine learning in Geosciences requires open-science principles, benchmark datasets for rigorous evaluation, and close collaboration between geoscientists and Machine Learning (ML) researchers. More generally, other studies [2–5] emphasized that empirical rigor has not kept pace with advances in ML, encouraging many researchers to reassess the current state of the art in various ML domains through rigorous and comprehensive empirical evaluations [6–12]. A notable and consistent finding across these evaluations is that hyper-parameter tuning and testing on multiple datasets can result in traditional methods performing on par with more recent approaches. This phenomenon has been observed in generative adversarial networks [6], language models [7], information retrieval [8], networked point process models [10], data imputation methods [11] and binary classification tasks [12].

In light of these findings, the design of rigorous benchmarks for the evaluation of seismic segmentation models becomes a critical endeavor. Such benchmarks not only ensure fair comparisons but also foster the development of robust and generalizable models. By incorporating diverse datasets, a comprehensive set of metrics, and unified guidelines for model training and testing, the evaluation process can provide a clearer understanding of the strengths and weaknesses of different approaches while ensuring equitable comparisons among methods. This is essential for advancing the field in a manner that is both reliable and scientifically grounded.

This work aims at proposing a benchmark to provide a common ground for researchers, practitioners, and industry professionals to rigorously evaluate the performance of various algorithms and models. By establishing a benchmark with well-defined testing and training protocols, the community will be able to objectively assess the strengths and weaknesses of different approaches, fostering innovation and collaboration in the field of seismic data analysis.

The proposed benchmark is based on the following principles, some of which are antagonistic, so that a compromise proposal is sought. First, it should be *supported by the literature*, based on a systematic review of related works. Whenever possible, the most used experimental protocols are prioritized in the design. The second principle is *diversity*, in the sense that the benchmark should cover the many aspects of the

G. Canguçu, L. Jorge, G. Barreto, T. Silva, W. Schwartz, and P. Vaz-de-Melo are with the Department of Computer Science, Universidade Federal de Minas Gerais, Brazil.

L. Lima, W. Caldas and C. Tavares are with Petróleo Brasileiro S.A., Brazil.

A. Machado is with the Department of Computer Science, Pontifical Catholic University of Minas Gerais, Dom Jose Gaspar 500, Belo Horizonte, Brazil, 30130-100, +5531986349054 (e-mail: alexeimcmachado@gmail.com).

problem, including data and evaluation metrics. Diversity must be supported by *impartiality* so that any data splitting approach and performance evaluation produce unbiased assessments. It must, however, present *simplicity*, in order to favor easy analyses and be computationally efficient. Finally, any choices should be *meaningful*, technically sound and respecting the many geological aspects of the problem domain.

In the following sections, the key components of the proposed benchmark are outlined, including the diverse datasets considered, the evaluation metrics, and the guidelines for model training and testing. The design of the benchmark should enhance the reproducibility of deep learning research within the Geosciences domain. To this end, all associated materials have been made openly available.

II. Related works

The application of deep learning techniques in the interpretation of seismic facies reflecs a broader trend towards leveraging advanced computational methods in geosciences. This section reviews key contributions in this domain, with a particular focus on the attempts to stablish standadized experimental protocols to assess the efficacy of these methods.

One of the pioneers works towards this direction is the study of Alaudah *et al.* [13], which proposed the first benchmark for seismic facies classification. Their work stands out for open-sourcing a fully-annotated 3D geological model of the Netherlands F3 Block, and also for presenting two baseline models for facies classification based on a deconvolution network architecture. The models use either whole sections or 2D patches extracted from the inlines and crosslines, and the data is labeled into six classes.

Based on Alaudah's partition protocol, Guazzelli et al.[14], Trinidad et al.[15], and Tolstaya et al.[16] used the same testing sets established by the benchmark. However, instead of training on the whole training set, part of it was used for validation. Further modifications were introduced by Li et al.[17] and Wang et al.[18], who altered the partition protocol and used only the inline sections of the dataset.

A different publicly available interpretation of the Netherlands F3 Block was presented by Silva *et al.* [19] comprising nine horizons and approximately 190,000 labeled images. The interpretation was however not accompanied by a partition protocol, so that subsequent works on the dataset used diverse experimental designs. Wang et al. [20], for example, decided to use a random portion of 15% of the training set for validation and the remaining data for testing. On the other hand, Monteiro et al. [21] used a 60/20/20 split for training, validation and testing, with 5-fold cross-validation. Both papers used inlines and crosslines but while Monteiro chose a random/few shot approach, Wang used an equidistant sampling method to select slices at regular intervals.

The third interpretation of the F3 dataset known as ConocoPhillips is synthetic and was generated by a neural network that annotates seismic facies. Although cited in works like the ones presented by Wang et. al. [22] and Zhang et. al [23], detailed information about this interpretation is scarce and no experimental protocol seems to be available. Similarly, other public datasets such as the Penobscot (terranubis.com/datalist/free/) and Parihaka (www.nzpam.govt.nz) lack of accompanied experimental guidelines.

In conclusion, the literature review corroborates the urgency for a unified benchmark, as proposed here. Table I presents a summary of prominent works on seismic data segmentation where diverse experimental procedures are used, both in data splitting and performance evaluation, making the comparison of results impractical. The next sections outline the available data and evaluation metrics that can be used to formalize a simple, unbiased, and meaningful benchmark.

III. Public Seismic Datasets

Similarly to other problems in computer vision, seismic segmentation also suffers from a lack of large publicly available annotated datasets that are suitable for training and evaluating deep learning models [13]. As an attempt to provide open labeled data, however, some annotated datasets were made publicly available under a Creative Commons Attribution license. A summary of the open labeled datasets, with their number of classes, size and format is displayed in Table II and detailed in this section.

A. The F3 Netherlands dataset

The F3 dataset is widely used in stratigraphic and machine learning studies due to its good resolution and well-defined geological area, and is available at the Open Seismic Repository (terranubis.com/datalist/free). The dataset consists of 384 km^2 of time migrated 3D seismic data, with 651 inlines and 951 crosslines. Slice 100 is the first inline slice of the interpreted volume, and slice 300 is the first in crossline direction. It is likely that the starting slices in both directions were excluded due to noisy signals, sections with gaps, or data distortion.

In the interpretation provided by Silva *et al.* [19], nine distinct horizons and ten classes are arranged in descending order of geological age and kept at their original resolution of 651×951×462 voxels. Also in 2019, Alaudah *et al.* [13] published an interpretation of the F3 Netherlands dataset using both well-logs and 3D seismic data in which six lithostratigraphic units were labeled. It is important to notice that this is the only interpretation of a public seismic dataset in the literature that defines a protocol for training and test partition, where the test labels are also available. The training set was composed of inline slices 300 to 700 and crossline slices from 300 to 1000. Two splits for testing were defined: Set 1, including inline slices from 100 to 299 and crosslines 300 to 1000; and Set 2, with inlines 100 to 700 and crosslines 1001 to 1200. The F3 dataset was still annotated by ConocoPhillips [34] in an open project that focused on the segmentation of seismofacies using deep neural networks. The interpretation used the entire dataset at its original resolution of 651×951×462 voxels, although not much information was provided about the labeling process.

A comparison between the annotations of Alaudah, Silva and ConocoPhillips reveals that the first considers fewer slices compared to the last two. Furthermore, Alaudah included fewer timeslices (200 to 454), which correspond to a depth

Table I: Summary of partition protocols found in the literature search.

Reference	Dataset	Interpr.	Dir.	Format	Position	Training	Validation	Test	Model	Performance
Wang, 2019[22]	F3	Own	NA	2D slices	Equally spaced	80%	10%	10%	U-Net	IOU(0.81:0.88)
Alaudah, 2019[13]	F3	Own	I+X	2D slices/ Patches	Adjacent/ random	I[300:700], C[300:1000]	NA	I[100:700], C[300:1200]	Encoder-decoder CNN	PA(0.79:0.91), MCA(0.57:0.82), FWIU(0.64:0.84)
Zhang, 2020[23]	Subset of F3	Conoco-Phillips	I	3D cubes	Random	23 slices	7 slices	10 slices	SegNet and U-Net based	PA(0.91:0.96)
Guazzelli, 2020[14]	F3, Stanford-VI-E	Alaudah	I+X	3D cubes	Adjacent	80% of train set	20% of train set	I[100:700], C[300:1200] of sets 1,2	TOP-CNN	PA(0.79:0.88), MCA(0.57:0.73), FWIOU(0.64:0.79)
Liu, 2020[24]	F3	Conoco-Philips	I	3D cubes/ Patches	Random	19,550 masked samples	NA	20,000 unmasked samples	VGGNet and GAN	A(0.82), F1(0.81), APS(0.89)
Li, 2021[25]	F3	Own	I+X	2D slices	Spaced	70%	10%	20% (10-fold c.v.)	ADDCNN	IOU(0.30:0.88)
Wang, 2021[20]	F3	Silva	I+X	2D slices	Equally spaced	32 and 65 sections	random 15% of train set	Remaining data	U-Net and Bayesian neural network	PA(0.96:0.98), MIOU(0.94)
Zhang, 2021[26]	F3	Conoco-Phillips	I	2D slices/ cubes	Random	25 inlines	NA	5 inlines	CNN, SegNet, DeepLabv3+	MCA(0.90:0.92), MIOU(0.36:0.93)
Trinidad, 2021[15]	F3	Alaudah	I+X	2D slices	Adjacent	70% of train set	30% of train set	I[100:700], C[300:1200] of sets 1,2	Atrous Bidirectional U-Net ConvLSTM	PA(0.90:0.94), CA(0.59:0.98), MCA(0.81:0.87), FWIU(0.83:0.89)
Li, 2022[27]	F3	Alaudah	I	2D slices	Adjacent	1%	NA	99%	DeepLabv3+	PA(0.96:0.98), MCA(0.90:0.95), FWIOU(0.93:0.95), MIOU(0.84:0.90), F1(0.91:0.95)
Chen, 2022[28]	F3	Alaudah	I+X	2D slices	Adjacent	varied samples	NA	varied samples	Hrnetv2-W32	PA(0.88:0.93), MCA(0.78:0.86), MIOU(0.64:0.74), FWIOU(0.82:0.87)
Tolstaya, 2022[16]	F3	Alaudah	I+X	2D slices	Adjacent	90% of train set	10% of train set	I[100:700], C[300:1200] of sets 1,2	U-Net and EfficienteNet B1	PA(0.93:0.94), MCA(0.78:0.95), FWIOU(0.77:0.91), MIOU(0.62:0.85)
Abid, 2022[29]	F3	Alaudah	I+X	2D patches	Random	60%	20%	20%	DeepLabv3+ and Seg-Net	MCA(0.94:0.97), MIOU(0.88:0.93)
Wang, 2022[30]	F3	Alaudah	I+X	2D slices	Random	10 to 100	NA	NA	U-Net	F1(0.82)
Wang, 2023[18]	F3	Alaudah	I	2D slices	Spaced	2, 4, 8 or 16 slices	NA	Remaining inlines	U-Net	PA(0.98:0.99), MCA(0.98:0.99)
Chevitarese, 2018[31]	Penobscot	Baroni	I+X	2D slices	Adjacent	60%	NA	30% (10% skipped)	Danet-FCN	CA(0.76:0.97)
Nasim, 2022[32]	Penobscot, F3	Baroni, Alaudah	I+X	2D patches	Adjacent	80% of F3	20% of F3	Penobscot	EarthAdaptNet	PA(0.73:0.83), MCA(0.58:0.78), FWIOU(0.57:0.77), MIOU(0.43:0.62), F1(0.91:0.95)
Monteiro, 2022[21]	Parihaka, F3	Chevron, Silva	I+X	2D slices	Random/Few shot	60%	20%	20% (5-fold c.v.)	ResNet50	MIOU(NA)
Su, 2022[33]	Parihaka	Chevron	I+X	3D patches	Spaced	75% of 590 profiles	25%	Same profiles used for training	U-Net	A(0.85:0.97)
Li, 2022[27]	Parihaka	Chevron	I	2D slices	Adjacent	1%	NA	99%	DeepLabv3+	PA(0.94:0.97), MCA(0.84:0.93), FWIOU(0.88:0.94), MIOU(0.81:0.88), F1(0.87:0.93)
Wang, 2022[30]	Parihaka, F3	Chevron, Alaudah	I+X	2D slices	Adjacent	464 slices	116 slices	251 slices	U-Net	A(0.95)
Tolstaya, 2022[16]	Parihaka	Chevron	X	2D slices	Adjacent	582 slices	NA	200 slices	U-Net and EfficienteNet B1	PA(0.93:0.94), MCA0.93:0.96), FWIOU(0.87:0.90), MIOU(0.72:0.77), F1(0.81:0.86)

NA: Not applicable, not available, or unclear; Interpr.: Source of interpretation; Dir.: Direction; I: Inline section; X: Crossline section; C.v.: Cross-validation; A: Accuracy, APS: Average Precision Score; PA: Pixel Accuracy; CA: Class Accuracy; MCA: Mean Class Accuracy; IOU: Intersection over Union; MIOU: Mean Intersection over Union; FWIOU: Frequency Weighted Intersection over Union; F1: Mean F1 Score.

Table II: Summary of the available public labeled datasets for seismic image segmentation.

Dataset	**Interpretation**	**Classes**	**Resolution** (I×X×T)	**Range of the dataset slices**	Size (GB)	Format
F3 Netherlands	Alaudah [13]	6	601×901×255	I:100..700, X:300..1200, T:200..454	1.1	npy
F3 Netherlands	Silva [19]	10	651×951×462	I:100..750, X:300..1250, T:1..462	1.1	tiff
F3 Netherlands	ConocoPhillips [34]	9	651×951×462	I:100..750, X:300..1250, T:1..462	1.1	sgy
Penobscot	Baroni [35]	8	601×481×1501	I:1000..1600, X:1000..1480, T:1..1501	2.2	h5
Parihaka	Chevron [36]	6	841×1116×1006	I:1..841, X:1..1116, T:1..1006	1.8	segy

I:Inlines; X:Crosslines; T:Timeslices.

between 1,005 and 1,877 meters. In this interpretation, only six classes were labeled, compared to ten in the study by Silva and nine in the ConocoPhillips classification. Figure 1 shows the inline slice 300 of the F3 dataset, as used in the three interpretations, where differences in resolution, content and labeling can be observed. The motivation to discard slices, as performed by Alaudah, was probably to avoid artifacts and missing data on the limits of the seismic volume caused by instrumentation or acquisition procedures. The interpretation provided by Silva, however, used the whole volume with all noisy sections and data gaps. Another study that mentioned this caveat was carried out by Wang *et al.* [20], which used 908 crosslines and 589 inlines of Silva *et al.* interpretation in their experiments.

B. The Penobscot dataset

The data used for the Penobscot set were provided by the Nova Scotia Department of Energy and the Canada Nova Scotia Offshore Petroleum Board, and managed by the Earth Sciences Open Seismic Repository (terranubis.com/datalist/free). It is useful for structural trap studies and integration with well data, with a focus on hydrocarbon prospects. The dataset consists of 87 km^2 time migrated 3D seismic data, with 601 inlines and 481 crosslines. Baroni *et al.* interpreted the inline slices between 1000 and 1600, crosslines between 1000 and 1480 and timeslices between 1 and 1501. The interpretation primarily comprises seven horizons and eight labels, as exemplified in Figure 1d.

C. The Parihaka dataset

The survey to acquire the Parihaka dataset was carried out at offshore Taranaki, New Zealand and was made publicly available by New Zealand Petroleum and Minerals (www.nzpam.govt.nz), with labels provided by Chevron U.S.A. (www.chevron.com). It is a high-quality dataset that includes complex geological structures, such as faults and gas hydrate deposits, and is often used in structural interpretations. The 3D volume is made of 841 inlines, 1116 crosslines and 1006 timeslices, containing six classes. The training set is composed of inlines 1 to 590 and crosslines from 1 to 782. Two test sets are suggested but their labels are unavailable. Figure 1e shows the inline slice 10 of the Parihaka dataset.

IV. Data Benchmark Proposal

Following the design principles of *support*, *diversity*, *impartiality*, *simplicity* and *meaningfulness*, this section describes the proposed data benchmark for seismic segmentation. The experimental protocol is composed of a set of seismic volumes with their respective interpretation, direction, position and range of the used slices, together with train/test splitting.

As public datasets are scarce, a decision was made to include subsets of the three available sets, namely the F3 Netherlands, Penobscot and Parihaka volumes. This choice bestows variety and meaningfulness to the benchmark, as different geological sites are represented in three continents. It also favors impartiality as long as each method using the benchmark will need to prove to be effective on more than a single dataset.

For the F3 dataset, the interpretation of Silva *et al.* [19] was preferred over the one presented by Alaudah *et al.* [13]. The choice was motivated by the fact that the former is larger and more diverse than the latter. As for the annotation provided by ConocoPhillips [34], it is rather focused on seismofacies (different from the another interpretations that focus on lithofacies) and limited details about the interpretation process are available.

A thorough examination on the F3 dataset revealed noisy sections and reverberation on the top 70 timeslices, in addition to the damaged inlines 1 to 99 and 701 onwards, already excluded by Alaudah. Although no references regarding the quality of the Penobscot data have been found in the reviewed literature, instances of damaged or low-quality images and data gaps in the seismic volume were observed from inline 1000 to 1069, as well as from inline 1530 to 1600. Gaps can be also observed in the upper portion along the crossline direction. Additionally, the crossline section 1000 has missing data for the most part, not just on the upper portion. For the Parihaka dataset, inlines 591 to 841 and crosslines 783 to 1116 were removed for the same reasons. Table III specifies the range of slices considered for the benchmark, after discarding slices that have corrupted and unlabeled data.

A decision was made to consider two-dimensional slices in both the inline and crossline directions, a practice supported by most of the literature. This approach is valuable for evaluating the segmentation methods' capacity to accurately track the geological features of the volumes in various directions.

In the reviewed literature, the data splitting scheme was the most varied aspect of the experimental design, as shown in Table I. The simplest and most frequent approach was to use 80% of the data in the form of 2D slices for training, from which a part could be optionally used for validation, and the remaining 20% of the slices for testing. Cross-validation was used in only two studies [21, 25], respectively with 10 and 5 folds.

The benchmark proposed in this paper adopted the 80/20 partitioning scheme with 5-fold cross-validation. Cross-

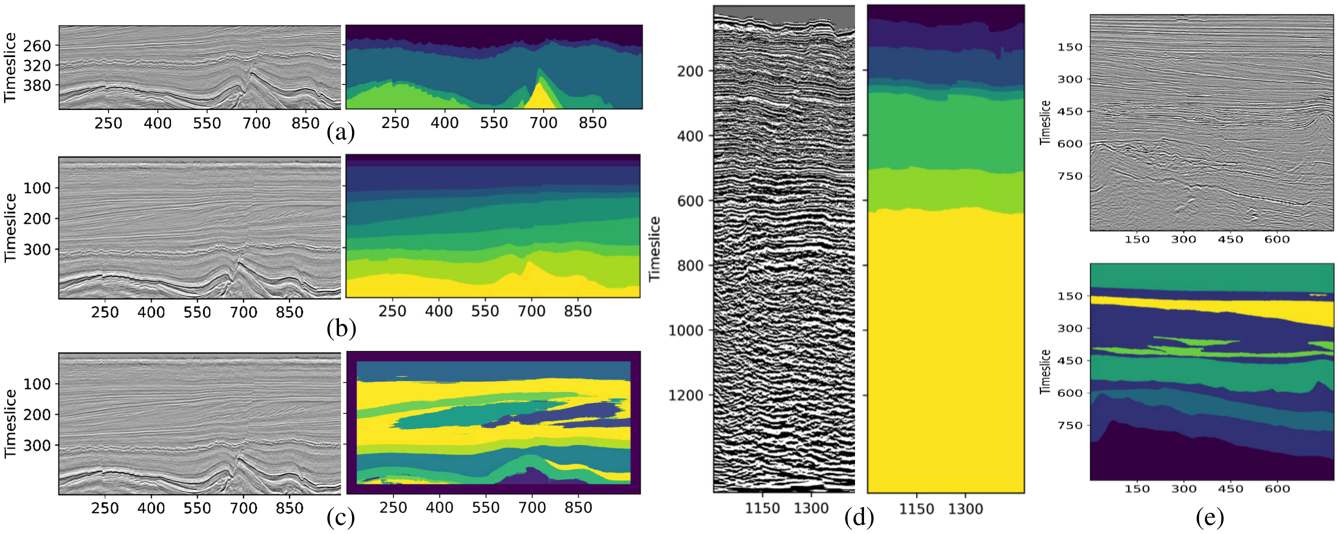


Figure 1: Inline section 300 from F3 Netherlands with interpretations provided by (a) Alaudah, (b) Silva and (c) ConocoPhillips; Inline 1300 from Penobscot with respective labels (d); Inline 10 from Parihaka with respective labels (e).

validation is an important step to impartiality, besides enabling statistical significance assessment through hypothesis testing. As a compromise between statistical significance and computational cost, 5 folds were preferred over the 10-fold scheme presented by Li *et al.* [25]. In addition to the 80/20 scheme, another one using only 20% of the data for training and the remaining for testing was included to evaluate the ability of the methods while dealing with small datasets. In both schemes, the volume was divided into five sections of adjacent slices for cross-validation. Adjacent-slice partitioning was used, as it is simpler and less prone to biasing than random selection. Spaced positioning in which the test block is separated from the training data by a gap as wide as 10% of the whole volume [31], but more frequently of 10 slices [17, 22], was considered not worthy of data waste, as discussed in Section VII. Table III describes the exact ranges of each block for the 5-fold cross-validation scheme, as proposed in this benchmark.

V. Benchmark Evaluation Metrics

In total, nine different evaluation metrics for the seismic segmentation task have been identified in the reviewed literature, with four of them being the most frequent: Pixel Accuracy (PA), Mean Intersection over Union (mIoU), Mean Class Accuracy (MCA), and Frequency-weighted Intersection over Union (FWIoU). Although these metrics are highly correlated, as shown in Table IV, a choice was made to include all of them in the benchmark proposal, so as to provide diversity, meaningfulness, impartiality, and a way to compare future methods with the past literature. The metrics are defined based on the number of true positives (TP), true negatives (TN), false positives (FP), false negatives (FN), the number of pixels in the image, n, and the number of classes, m.

A. Region-based analysis

Pixel Accuracy is a straightforward evaluation metric that quantifies the percentage of correctly classified pixels within an image:

$$\text{PA} = (\text{TP} + \text{TN})/\text{n}.$$

While it offers a fundamental measure of accuracy, it overlooks class imbalances.

The MCA evaluates the average accuracy among the classes in a dataset. It provides a more fine-grained evaluation compared to PA by considering individual class performances, although imbalance is still not accounted for:

$$\text{MCA} = (1/m)\textstyle\sum_{i=1}^{m} TP_i/(TP_i + FN_i).$$

The mIoU is a standard metric for evaluating image segmentation models that considers the overlap between the predicted and ground truth segmentation masks for each class:

$$\text{mIoU} = (1/m)\textstyle\sum_{i=1}^{m} TP_i/(TP_i + FP_i + FN_i).$$

The FWIoU extends the concept of region overlap between the predicted and ground truth segmentation masks by weighing the score of each class by their frequencies:

$$\text{FWIoU} = (1/n)\textstyle\sum_{i=1}^{m} n_i TP_i/(TP_i + FP_i + FN_i).$$

In many segmentation tasks, class imbalance is common, where some classes have significantly more or fewer pixels than others. Metrics like FWIoU help to mitigate the impact of imbalanced datasets.

B. Qualitative analysis

Conducting a qualitative analysis of the results is also crucial for assessing the efficacy of the models. This involves visually inspecting the segmentation maps generated by the deep learning algorithms and comparing them against the ground truth to evaluate the quality of facies boundaries and the overall region partitioning. The qualitative evaluation allows researchers to discern how well the model captures the intricacies of seismic data, such as the continuity of layers, the delineation of different facies, and the identification of subtle features that might be indicative of key geological structures.

Table III: Summary of the proposed benchmark.

Dataset	Range of the used slices	Fold size	Starting slice indices for 5-fold cross-validation
F3 Netherlands (Silva)	I:100..699, X:300..1249, T:71..462	I:120, X:190	I={100, 220, 340, 460, 580}; X={300, 490, 680, 870, 1060}
Penobscot	I:1070..1529, X:1000..1479, T:1..1501	I:92, X:96	I={1070, 1162, 1254, 1346, 1438}; X={1000, 1096, 1192, 1288, 1384}
Parihaka	I:1..590, X:1..780, T:1..1006	I:118, X:156	I={1, 119, 237, 355, 473}; X={1, 157, 313, 469, 625}
Metrics: PA, MCA, IoU, FWIoU			
Methods: U-Net, SegNet, DeconvNet			

I:Inlines; X:Crosslines; T:Timeslices; MCA:Mean class accuracy; IoU:Intersection over union; FWIoU:Frequency weighted IoU; PA:Pixel accuracy.

Table IV: Correlations between the performance metrics.

	PA	MCA	mIoU
MCA	0.90		
mIoU	0.94	0.98	
FWIoU	0.99	0.91	0.95

This process not only validates the quantitative metrics but also provides insights into the model's performance in real-world scenarios, where the complexity of seismic data often presents challenges not fully encapsulated by quantitative measures alone.

VI. Benchmark algorithms

Numerous deep learning architectures have been proposed to address seismic image segmentation, as shown in Table I. In order to illustrate the application of the proposed benchmark, three popular models were implemented as baselines: the U-Net [37], the SegNet [38] and the DeconvNet [39]. The hyperparameters used in the experiments are described in Section VII and in the code documentation that accompanies this benchmark. It was out of the scope of this work to optimize the network models used for segmentation, so that fine tuning was applied based on grid searching up to a certain level that would be competitive to the results reported in the literature.

The U-Net [37] was introduced in 2015 and is extensively used in semantic segmentation. Its "U" shape evidences the encoder-decoder structure incorporated with skip connections. This capability preserves both low-level and high-level features, making it well-suited for tasks requiring fine details, such as seismic image analysis.

The DeconvNet [39] employs the thirteen convolutional layers and five max-pooling layers of VGG16 as its encoder. Unlike VGG16, however, it replaces fully connected layers with a mirrored version of convolutional layers for the decoder, performing upsampling instead of downsampling. DeconvNet is robust and simple, being useful in tasks requiring precise pixel-level segmentation. However, due to the absence of skip connections, DeconvNet may struggle to preserve spatial information during upsampling, making it less accurate in segmenting small objects compared to U-Net.

Finally, the SegNet [38] was proposed in 2017 as a compromise between accurate pixel-wise segmentation and efficiency. Similarly to DeconvNet, it also employs VGG16's convolutional layers for the encoder, retaining maximum values and their corresponding indices during max pooling. This preserves spatial locations of maximum activations, helping with the upsampling in the decoder.

VII. Experimental Results

A set of 180 experiments was conducted to exemplify the application of the benchmark. The U-Net, DeconvNet and SegNet models were applied to the analyses of the F3 Netherlands, Parihaka, and Penobscot, in inline and crossline directions, using the 80/20 and 20/80 benchmark training/test partitions, in a 5-fold cross-validation strategy. The performance of these baseline models on the test sets for each of the four benchmark evaluation metrics is presented in Table V.

The choice of hyperparemeters was mostly based on the literature, as it provides a suitable baseline for this benchmark. Out of all the reviewed works, only two of them do not describe the used parameters [13, 14]. For the remaining works, the most popular choices were Adam for the optimizer [15, 20, 22–29, 32, 33], cross-entropy for the loss function [15, 16, 18, 20, 21, 24, 25, 27–29, 32, 33] and L2 regularization for weight decay [22, 24, 25, 27, 29–32]. Although there was no consensus regarding learning rates, the most common value was 10^{-4} [15, 20, 22, 24–28, 30–33]. Furthermore, the number of epochs was kept as 30 based on empirical tests, since the literature lacks a clear preference for this parameter.

All models were implemented in the Python programming language based on their original implementations [37–39], using the PyTorch library and run in an AMD EPYC 7742 64-Core processor with NVIDIA A100-SXM4 GPU (80 GB of VRAM), under the Ubuntu 22.04 LTS operating system. As for preprocessing the data, while encoder-decoder architectures like U-Net, SegNet, and DeconvNet typically require input sizes that are powers of 2, the provided implementations use padding to align encoder and decoder feature maps before concatenation, and therefore no additional preprocessing is required. Each model took at most 30 minutes to train. Finally, all the source code that is necessary to use the present benchmark is made available at *github.com/uai-ufmg/seismic-segmentation-benchmark* as a way of encouraging other researchers to use a standard experimental protocol to evaluate their own segmentation methods.

Table V highlights the differences between the metrics upon evaluating the models. PA and MCA yield the highest scores in almost every experiment, which may be explained by the fact that these metrics do not account for class imbalance. In turn, the mIoU and FWIoU measure how the predictions overlap with their respective annotation masks.

The discrepancy between the metrics is even more evidenced when models are trained using only 20% of the data. As expected, models underperform when trained with fewer data. Nonetheless, the drop in performance is not proportional to the decrease in the sample size. The relative good scores obtained with a fifth of the seismic volume used for training

Model	Dataset	Dir	80% training / 20% test				20% training / 80% test			
			PA	MCA	mIoU	FWIoU	PA	MCA	mIoU	FWIoU
U-Net	F3 - Silva	I	0.97 (0.03)	0.97 (0.02)	0.94 (0.05)	0.94 (0.05)	0.91 (0.03)	0.91 (0.03)	0.84 (0.04)	0.83 (0.04)
		X	0.93 (0.06)	0.93 (0.06)	0.87 (0.10)	0.88 (0.09)	0.81 (0.07)	0.80 (0.08)	0.69 (0.09)	0.70 (0.09)
	Parihaka	I	0.91 (0.04)	0.80 (0.09)	0.70 (0.10)	0.83 (0.06)	0.75 (0.03)	0.58 (0.09)	0.47 (0.06)	0.61 (0.04)
		X	0.93 (0.03)	0.88 (0.06)	0.81 (0.08)	0.88 (0.06)	0.78 (0.06)	0.68 (0.08)	0.55 (0.09)	0.66 (0.07)
	Penobscot	I	1.00 (0.00)	0.99 (0.00)	0.97 (0.01)	0.99 (0.01)	0.98 (0.01)	0.98 (0.01)	0.94 (0.01)	0.97 (0.01)
		X	0.94 (0.09)	0.91 (0.09)	0.84 (0.16)	0.91 (0.11)	0.97 (0.03)	0.95 (0.03)	0.90 (0.07)	0.95 (0.04)
SegNet	F3 - Silva	I	0.96 (0.03)	0.96 (0.03)	0.93 (0.05)	0.93 (0.05)	0.87 (0.04)	0.87 (0.03)	0.77 (0.06)	0.77 (0.06)
		X	0.93 (0.06)	0.93 (0.06)	0.87 (0.09)	0.88 (0.09)	0.78 (0.06)	0.77 (0.06)	0.64 (0.08)	0.65 (0.08)
	Parihaka	I	0.92 (0.04)	0.82 (0.09)	0.73 (0.11)	0.87 (0.06)	0.76 (0.02)	0.59 (0.06)	0.48 (0.03)	0.61 (0.03)
		X	0.92 (0.04)	0.88 (0.07)	0.77 (0.10)	0.86 (0.06)	0.75 (0.08)	0.63 (0.08)	0.51 (0.10)	0.63 (0.09)
	Penobscot	I	1.00 (0.00)	0.99 (0.00)	0.97 (0.01)	0.99 (0.00)	0.97 (0.02)	0.95 (0.03)	0.91 (0.04)	0.95 (0.03)
		X	0.99 (0.01)	0.97 (0.02)	0.94 (0.03)	0.98 (0.01)	0.96 (0.05)	0.91 (0.10)	0.84 (0.14)	0.93 (0.07)
DeconvNet	F3 - Silva	I	0.97 (0.03)	0.97 (0.03)	0.93 (0.05)	0.94 (0.05)	0.83 (0.05)	0.84 (0.04)	0.73 (0.06)	0.72 (0.06)
		X	0.92 (0.07)	0.92 (0.07)	0.85 (0.12)	0.86 (0.10)	0.77 (0.07)	0.75 (0.07)	0.62 (0.09)	0.64 (0.09)
	Parihaka	I	0.92 (0.03)	0.82 (0.08)	0.73 (0.10)	0.86 (0.05)	0.75 (0.04)	0.61 (0.08)	0.47 (0.04)	0.61 (0.06)
		X	0.92 (0.03)	0.89 (0.05)	0.77 (0.07)	0.85 (0.05)	0.75 (0.07)	0.66 (0.08)	0.52 (0.08)	0.63 (0.08)
	Penobscot	I	1.00 (0.00)	0.99 (0.01)	0.97 (0.01)	0.99 (0.00)	0.97 (0.02)	0.95 (0.04)	0.90 (0.03)	0.94 (0.03)
		X	0.97 (0.05)	0.97 (0.02)	0.92 (0.05)	0.96 (0.05)	0.87 (0.25)	0.87 (0.17)	0.78 (0.27)	0.84 (0.29)

Table V: Mean (standard deviation) for the pixel accuracy (PA), mean class accuracy (MCA), mean intersection over union (mIoU), and frequency weighted intersection over union (FWIoU) achieved by U-Net, SegNet, and DeconvNet. The results of 5-fold cross-validation is shown for the seismic volume splits of 80/20 and 20/80 in inline (I) and crossline (X) directions.

shows that applications with weakly supervised or unsupervised models are a promising research direction.

Qualitative examples of the predictions obtained with different train/test ratios are shown in Fig. 2. It can be seen that introducing more data can significantly mitigate the presence of discontinuities in the segmentation. Furthermore, the lower portions of the images seem to be where models are less accurate. This is a consequence of the data acquisition process, in which noise amplitude increases with depth. All the image volumes obtained in the experiments are available in the repository github.com/uai-ufmg/seismic-segmentation-benchmark (outputs/predictions.tar.gz), from which the outputs obtained with the methods can be compared plane by plane.

The experimental results shown in Table V corroborate the choice of this benchmark for multiple datasets and the analysis in both inline and crossline directions. In the 20/80 inline split, for example, the SegNet achieves a mIoU of 0.77 for the F3 dataset, 0.48 for the Parihaka and 0.91 for the Penobscot. One-way analysis of variance (ANOVA) reveals that these scores are significantly different at the level of α=0.001 (F(2,12)=118.27, p<0.001), where F is the F-statistics, p is the associated p-value, and α is the Type I error rate. As another example, the mIoU achieved by SegNet for the F3 dataset is 0.77 if taken in inline, but only 0.64 in the crossline direction. These values are different at the level of α=0.05 (F(1,8)=8.45, p=0.02).

An additional set of experiments investigated the choice for folds composed by adjacent slices against randomized sampling. For example, in the case of the 80/20 split of the Parihaka dataset in inline direction, the achieved mIoU with U-Net was 0.70 when the folds were defined as adjacent sections. The same method achieves an average(s.d.) mIoU of 0.95(0.01) if the folds are composed by random sampling. These values are different at α=0.001 (F(1,8)=30.94, p=0.001), showing that randomized sampling yields biased results.

A final set of experiments investigated the benefits of discarding sections around the test fold, in order to prevent data contamination by the training set. As shown before, the 80/20 split of the Parihaka dataset in inline direction achieved a mIoU of 0.7 using U-Net. If the dataset is now split in such a way that 10 slices between the train and test sets are removed, the average(s.d.) mIoU becomes 0.697(0.098). ANOVA reveals that the mIoU scores are equivalent (F(1,8)=0.002, p=0.963), showing that discarding slices produced no significant effects on the performance. Although this tendency should be investigated for other experimental configurations, a visual analysis of the datasets shows that the seismic patterns do not vary considerably at two slices that are 10 slices apart. The exclusion of these data is therefore of limited profit, which justifies the choice of the proposed benchmark to use the whole range of seismic sections.

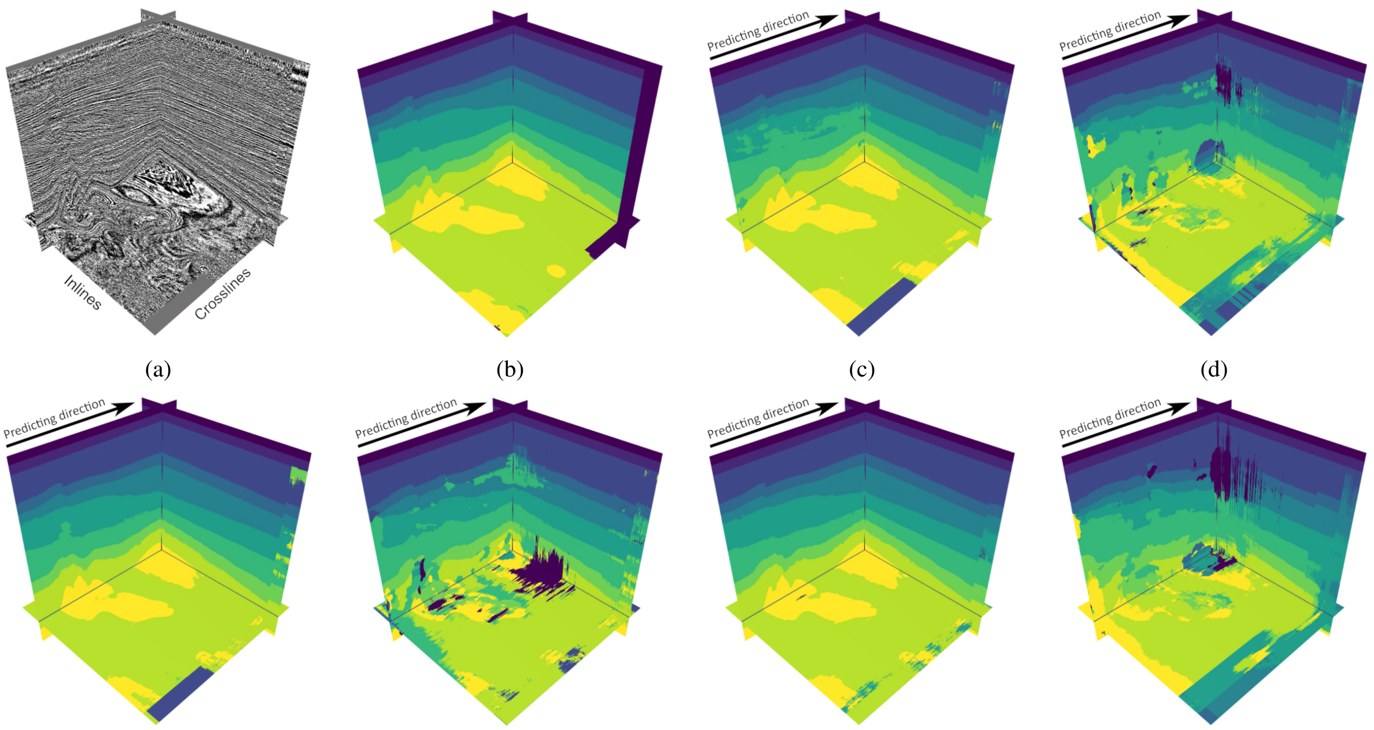
VIII. Conclusion

This paper proposed a protocol for evaluating seismic segmentation models in the form of a benchmark that encompasses three distinct public datasets and four performance metrics. Three encoder-decoder architectures were used to illustrate the application of the proposed data partition scheme. Quantitative and qualitative results, reported in a series of experiments, showed that the benchmark is a compromise between *diversity*, *impartiality*, *meaningfulness* and *simplicity*. We hope that this study and its accompanied source code will encourage researches to pursue more standardized experimentation that favor comparability and reproducibility while proposing new segmentation models.

Acknowledgements — The authors would like to thank Petróleo Brasileiro S.A. for the technical and financial support through its cooperation agreement with UFMG - Universidade Federal de Minas Gerais.

References

[1] K. J. Bergen, P. A. Johnson, M. V. de Hoop, and G. C. Beroza, "Machine learning for data-driven discovery in



(a) (b) (c) (d)

(e) (f) (g) (h)

Figure 2: Examples of qualitative results for the segmentation of the F3 dataset as seen from inline 100, crossline 100, and time slice 400. The first 2 volumes show (a) the seismic data and (b) the ground truth annotated by Silva. The remaining volumes show the segmentation obtained by SegNet using (c) 80/20 and (d) 20/80 splits, by U-Net using (e) 80/20 and (f) 20/80 splits, and by DeconvNet using (g) 80/20 and (h) 20/80 splits. Arrows above the output volumes indicate the prediction direction (crosslines).

solid Earth geoscience," *Science*, vol. 363, no. 6433, 2019.

[2] A. F. Cooper, Y. Lu, J. Forde, and C. M. De Sa, "Hyperparameter optimization is deceiving us, and how to stop it," *Advances in Neural Information Processing Systems*, vol. 34, pp. 3081–3095, 2021.

[3] B. Hutchinson, N. Rostamzadeh, C. Greer, K. Heller, and V. Prabhakaran, "Evaluation Gaps in Machine Learning Practice," in *2022 ACM Conference on Fairness, Accountability, and Transparency*, New York, NY, USA, 2022, pp. 1859–1876.

[4] T. Moreau, M. Massias, A. Gramfort, P. Ablin, P.-A. Bannier, B. Charlier, M. Dagréou, T. la Tour, G. Durif, C. F. Dantas, and Others, "Benchopt: Reproducible, efficient and collaborative optimization benchmarks," *Advances in Neural Information Processing Systems*, vol. 35, pp. 25 404–25 421, 2022.

[5] A. Simkó, A. Garpebring, J. Jonsson, T. Nyholm, and T. Löfstedt, "Reproducibility of the Methods in Medical Imaging with Deep Learning." in *Medical Imaging with Deep Learning*. PMLR, 2024, pp. 95–106.

[6] M. Lucic, K. Kurach, M. Michalski, O. Bousquet, and S. Gelly, "Are Gans created equal? A large-scale study," *Advances in Neural Information Processing Systems*, pp. 700–709, 2018.

[7] G. Melis, C. Dyer, and P. Blunsom, "On the State of the Art of Evaluation in Neural Language Models," in *ICLR*, 2018.

[8] W. Kryscinski, N. S. Keskar, B. McCann, C. Xiong, and R. Socher, "Neural Text Summarization: A Critical Evaluation," in *Proceedings of the 2019 EMNLP-IJCNLP*, Stroudsburg, PA, USA, 2019, pp. 540–551.

[9] A. Vaswani, G. Brain, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention Is All You Need," in *NeurIPS*, 2017, pp. 5998–6008.

[10] G. Borges, P. O. Vaz-de Melo, F. Figueiredo, and R. Assuncao, "Networked Point Process Models Under the Lens of Scrutiny," in *ECML PKDD*, 2020.

[11] S. Jäger, A. Allhorn, and F. Bießmann, "A Benchmark for Data Imputation Methods," *Frontiers in Big Data*, vol. 4, 2021.

[12] J. Bowles, S. Ahmed, and M. Schuld, "Better than classical? The subtle art of benchmarking quantum machine learning models," *arXiv:2403.07059*, 2024.

[13] Y. Alaudah, P. Michałowicz, M. Alfarraj, and G. AlRegib, "A machine-learning benchmark for facies classification," *Interpretation*, vol. 7, no. 3, pp. 175–187, 2019.

[14] A. B. Guazzelli, M. Roisenberg, and B. B. Rodrigues, "Efficient 3D semantic segmentation of seismic images using orthogonal planes 2D convolutional neural networks," in *2020 IEEE IJCNN*, 2020, pp. 1–8.

[15] M. Trinidad, A. Canchumuni, R. Feitosa, and M. Pacheco, “Seismic facies segmentation using atrous convolutional-LSTM network,” in *2021 PANACM*, Rio de Janeiro, 2021, pp. 1–7.

[16] E. Tolstaya and A. Egorov, “Deep learning for automated seismic facies classification,” *Interpretation*, vol. 10, no. 2, pp. SC31–SC40, 2022.

[17] Y. Li, P. Hu, Z. Liu, D. Peng, J. T. Zhou, and X. Peng, “Contrastive clustering,” in *2021 AAAI*, vol. 35, 2021, pp. 8547–8555.

[18] L. Wang *et al.*, “Semi-supervised semantic segmentation for seismic interpretation,” *Geophysics*, vol. 88, no. 3, pp. 1–57, 2023.

[19] M. Silva, L. Baroni, R. Ferreira, D. Civitarese, D. Szwarcman, and E. Brazil, “Netherlands dataset: A new public dataset for machine learning in seismic interpretation,” *arXiv:1904.00770*, 2019.

[20] D. Wang and G. Chen, “Seismic stratum segmentation using an encoder–decoder convolutional neural network,” *Math. Geosci.*, vol. 53, no. 6, pp. 1355–1374, 2021.

[21] B. Monteiro, H. Oliveira, and J. dos Santos, “Self-supervised learning for seismic image segmentation from few-labeled samples,” *IEEE Geosci. Remote Sens. Lett.*, vol. 19, pp. 1–5, 2022.

[22] Z. Wang, F. Li, T. Taha, and H. Arabnia, “Improved automating seismic facies analysis using deep dilated attention autoencoders,” in *2019 CVPR Workshops*, 2019, pp. 511–513.

[23] Y. Zhang, Y. Liu, H. Zhang, and H. Xue, “Seismic facies analysis based on deep learning,” *IEEE Geosci. Remote Sens. Lett.*, vol. 17, no. 7, pp. 1119–1123, 2019.

[24] M. Liu, M. Jervis, W. Li, and P. Nivlet, “Seismic facies classification using supervised convolutional neural networks and semisupervised generative adversarial networks,” *Geophysics*, vol. 85, no. 4, pp. O47–O58, 2020.

[25] F. Li, H. Zhou, Z. Wang, and X. Wu, “ADDCNN: An attention-based deep dilated convolutional neural network for seismic facies analysis with interpretable spatial-spectral maps,” *IEEE T Geosci Remote*, vol. 59, no. 2, pp. 1733–1744, 2020.

[26] H. Zhang, T. Chen, Y. Liu, Y. Zhang, and J. Liu, “Automatic seismic facies interpretation using supervised deep learning,” *Geophysics*, vol. 86, no. 1, pp. IM15–IM33, 2021.

[27] K. Li, W. Liu, Y. Dou, Z. Xu, H. Duan, and R. Jing, “Contrastive learning approach for semi-supervised seismic facies identification using high-confidence representations,” *arXiv:2210.04776*, 2022.

[28] X. Chen, Q. Zou, X. Xu, and N. Wang, “A stronger baseline for seismic facies classification with less data,” *IEEE T Geosci Remote*, vol. 60, pp. 1–10, 2022.

[29] B. Abid, M. Khan, and A. Memon, “Seismic facies segmentation using ensemble of convolutional neural networks,” *Wirel Commun Mob Com*, vol. 2022, 2022.

[30] R. Wang, J. Stitt, and A. Shugar, “Identifying geologic facies through seismic dataset-to-dataset transfer learning using convolutional neural networks,” Stanford University, Tech. Rep., 2021.

[31] D. S. Chevitarese, D. Szwarcman, E. V. Brazil, and B. Zadrozny, “Efficient classification of seismic textures,” in *2018 IJCNN*, 2018, pp. 1–8.

[32] Q. Nasim, T. Maiti, A. Srivastava, T. Singh, and J. Mei, “Seismic facies analysis: a deep domain adaptation approach,” *IEEE T Geosci Remote*, vol. 60, pp. 1–16, 2022.

[33] H. Su-Mei, S. Zhao-Hui, Z. Meng-Ke, Y. San-Yi, and W. Shang-Xu, “Incremental semi-supervised learning for intelligent seismic facies identification,” *Appl Geophys*, vol. 19, no. 1, pp. 41–52, 2022.

[34] ConocoPhillips, “Data from the Machine learning of Voxels (MalenoV), a seismic interpretation project from ConocoPhillips Norge,” http://bit.ly/3SdzRyX, 2017.

[35] L. Baroni, R. Silva, R. Ferreira, D. Civitarese, D. Szwarcman, and E. Brazil, “Penobscot dataset: Fostering machine learning development for seismic interpretation,” *arXiv:1903.12060*, 2019.

[36] Chevron U.S.A. Inc., “2020 SEG annual meeting machine learning interpretation workshop,” https://public.3.basecamp.com, 2020.

[37] O. Ronneberger, P. Fischer, and T. Brox, “U-net: Convolutional networks for biomedical image segmentation,” in *2015 MICCAI*, vol. III, no. 18, 2015, pp. 234–241.

[38] V. Badrinarayanan, A. Kendall, and R. Cipolla, “Segnet: A deep convolutional encoder-decoder architecture for image segmentation,” *IEEE T Pattern Anal*, vol. 39, no. 12, pp. 2481–2495, 2017.

[39] H. Noh, S. Hong, and B. Han, “Learning deconvolution network for semantic segmentation,” in *2015 ICCV*, 2015, pp. 1520–1528.