Automated seismic semantic segmentation using Attention U-Net

<table>
<thead>
<tr>
<th>Journal:</th>
<th>Geophysics</th>
</tr>
</thead>
<tbody>
<tr>
<td>Manuscript ID</td>
<td>GEO-2023-0149.R1</td>
</tr>
<tr>
<td>Manuscript Type:</td>
<td>Latest advancements in machine learning for geophysics</td>
</tr>
<tr>
<td>Keywords:</td>
<td>machine learning, facies, spectral analysis, frequency-domain, interpretation</td>
</tr>
<tr>
<td>Manuscript Focus Area:</td>
<td>Signal Processing, Seismic Amplitudes and Attributes</td>
</tr>
</tbody>
</table>

This paper presented here as accepted for publication in Geophysics prior to copyediting and composition. © 2023 Society of Exploration Geophysicists
Automated seismic semantic segmentation using Attention U-Net

Haifa AlSalmi$^{1,2}$ and Ahmed H. Elsheikh$^1$

$^1$ Heriot-Watt University, Edinburgh, UK.

$^2$ Imperial College London, Centre for Reservoir Geophysics, Resource Geophysics Academy, London, UK.

E-mail: h.alsalmi@hw.ac.uk (corresponding author).

(August 8, 2023)

GEO-2023-0149

Running head: Attention U-Net for seismic segmentation

ABSTRACT

Seismic facies mapping from a three-dimensional seismic cube is of significant value to various seismic interpretation and characterisation tasks. Traditional facies mapping is based on examining sedimentary environments and stratigraphic sequences, that provide distinct characteristics used for facies mapping. Given the complex nature of the task, manual facies mapping is typically time and labour consuming, and the quality of the decisions vary as a function of expertise. This complexity is further increased with the ever-increasing size of 3D seismic datasets. Deep learning methods
have shown a promising potential to perform fast, accurate and automated segmentation tasks. We investigate the application of machine learning techniques, particularly state-of-the-art deep convolutional neural networks (CNNs), as a framework to perform accurate automated seismic facies pixel-wise segmentation. The workflow consists of a CNN based U-Net architecture that adopts modern computer vision techniques. We propose three major changes to the standard U-Net to boost the performance for seismic semantic segmentation tasks: (1) using residual building blocks in the encoder (2) using transformer like attention gates after each residual block, and (3) using frequency spectrum data, in addition to seismic amplitude, as input to the network. We show that this implementation achieves higher accuracy metrics outperforming recently published state-of-the-art benchmarks. The performance of the proposed method is validated using two 3D seismic datasets, the F3 Netherlands dataset and the Penobscot dataset acquired offshore Nova Scotia, Canada. Experimentation involves training on a set of samples and tuning the hyper-parameters, followed by quantitative evaluation of the trained network. The proposed workflow produced high quality segmentation with significantly reduced artifacts, improved edge detection, and improved lateral consistency throughout the seismic survey.
INTRODUCTION

Seismic reflection data constitutes various characteristics that reflect the spatial distribution of geologic bodies. Seismic facies indicate the types of lithologic combination and sedimentary characteristics of strata (Sheriff, 1976; Brown, 2011). They can be identified by grouping seismic reflections whose patterns (such as amplitude, frequency, and geometry) are different from those of adjacent groups (West et al., 2002). Facies maps are fundamental for a myriad of tasks in exploration, production, and development of hydrocarbon fields. In seismic stratigraphy, facies importance lies in analysing subsurface depositional environments and lithofacies distribution (Dumay and Fournier, 1988). For reservoir characterisation it is essential for delineating reservoirs and mapping their extent, predicting reservoir properties and characterising its heterogeneity (Brown, 2011). In paleogeography reconstruction, facies are used to predict reservoir, seal, and source rock distribution.

Manually generated facies maps are generated through careful analysis of physical attributes, sedimentary environments and stratigraphic sequences. While the latter two focus on geologic frameworks that help us understand facie variation. Physical attributes of a seismic trace include: amplitude, phase and frequency. These properties are attributed to variations of physical parameters and are usually lithology induced. Changes in lithology and fluid content are often used to capture visual patterns that help delineate strata (Randen et al., 2000). Seismic facies analysis can then be described as examining reflection geometries and stratal architectures between sequence boundaries (Sieck and Self, 1977; Berg, 1982), to identify depositional sequences and
paleoenvironments in order to classify and infer seismic lithofacies (Mitchum Jr et al., 1977a,b; Roksandić, 1978).

Traditionally facies analysis requires both geologists and geophysicists to ensure both the dynamic and kinematic aspects of seismic are captured (Chopra and Marfurt, 2012). Given the elaborate nature of the processes involved, the task remains time and labour consuming. An elaborate overview of seismic facies analysis schemes is progressively detailed in Xu and Haq (2022). As an attempt to automate seismic facies analysis, recent research has proposed utilising deep learning based models. The advancement of deep learning (LeCun et al., 2015), and its successful use in computer vision tasks (Krizhevsky et al., 2017; Russakovsky et al., 2015), has led researchers to explore its application as an automated alternative to conventional seismic interpretation techniques. Applications of deep learning in tasks such as image classification, semantic segmentation, object recognition, motion tracking and image captioning, have been directed towards automated seismic interpretation.

Among the different types of deep neural networks, convolutional neural networks (CNNs) have been the most extensively studied (LeCun et al., 1989, 1998). Numerous applications of CNNs date back to the early 1990s. Hand written character detection is an example of early utilisation of CNNs developed by Microsoft. Since the early 2000s CNNs have been greatly successful in detection, segmentation and recognition of objects and regions in images. This success has brought about a revolution in computer vision; CNNs are now the dominant approach for almost all recognition and detection task and approach human performance on some tasks.
A traditional CNN architecture solves image classification tasks. In such scenarios, CNNs are used to provide an output class label for the whole image, this is also known as image-wise prediction. Other interpretation tasks require image segmentation which produces dense pixel-wise predictions. Attempting to use a CNN to achieve more dense predictions can be performed by using smaller input patches to infer a label to the central pixel (Waldeland et al., 2018; Zhao et al., 2018), this method is however imprecise and computationally time-consuming. Whilst effective for classification tasks, traditional CNNs are not suitable for image segmentation. Segmentation can be achieved by using a variation of the CNN referred to as the encoder-decoder architecture. The encoder similar to a CNN extracts features to obtain concentrated feature maps, and the decoder includes up-sampling operations that gradually recover the resolution of the feature maps and produces a predicted output. Such architectures include the U-Net (Ronneberger et al., 2015), DeconvNet (Noh et al., 2015) and SegNet (Badrinarayanan et al., 2017).

Some of the most successful applications in automated seismic interpretation include fault detection (Di and Gao, 2017; Di et al., 2018b; Guitton, 2018; Xiong et al., 2018; Wu et al., 2019b), log interpretation (Tschannen et al., 2017; Zhu et al., 2018), salt-body delineation (Waldeland et al., 2018; Shi et al., 2019; Di et al., 2018a; Di and AlRegib, 2020), horizon tracking (Peters et al., 2019; Wu et al., 2019a; Tschannen et al., 2020), property estimation (Alfarraj and AlRegib, 2019; Das et al., 2019; Wang et al., 2018; Di and Abubakar, 2022; Di et al., 2022), sequence analysis (Li et al., 2019; Di et al., 2020), integrated interpretation (Di et al., 2019; Guo et al., 2020), and facies
segmentation (Alaudah et al., 2019; Chen et al., 2022; Di et al., 2018a, 2019, 2020; Zhang et al., 2021; Zhao et al., 2018).

Most of these deep-learning aided applications in seismic interpretation, including the methods explored in this work, are supervised learning methods. This requires an interpreter to annotate a set of seismic sections as training data. The network then uses training data to map inputs to labelled outputs (Goodfellow et al., 2016). Two major challenges for supervised learning are: (1) the limited number of training examples to represent the complexities in seismic patterns well within the seismic survey and (2) imbalanced facies class distribution. In such cases, despite learning from the annotated sections, the CNNs generalisation capability is limited and the model tends to memorise rather than learn in what is called over-fitting the training data. Predictions on sections with patterns dissimilar to those in the training data are likely to be highly inaccurate. Similarly, imbalanced classes causes the accuracy of less frequent classes to be much lower than those with higher samples in the training dataset. Inaccurate predictions not only results in incorrect label assignment, but it also can appear in common forms of artifacts including blobs, zig-zags, and stripes at the interface between two facies.

In this paper, we use supervised deep learning to predict seismic facies using a U-Net based architecture. To solve the interpretation limitations discussed above we propose three main modification to the standard U-Net architecture: (1) using residual blocks as building block in the encoder, this improves on tradition convolutional blocks by using identity mapping skip shortcuts (2) using transformer like attention
blocks after each residual block, and (3) using the frequency spectrum, in addition to seismic amplitude, as an input to the network. It is worth mentioning that, other efforts to combine attention to U-Nets have been made by Oktay et al. (2018). However, we propose to use a transformer type attention (Vaswani et al., 2017), unlike the additive attention previously used in (Oktay et al., 2018). In our experiments we observed that Transformer-like attention outperforms additive attention, and to the best of our knowledge, this is the first application of this type of attention module within a U-Net architecture for seismic image segmentation.

We showcase the performance of the proposed model by comparing it to published benchmarks. The first model developed by Alaudah et al. (2019) utilised a deconvolution network architecture and compares results using patch based versus section based training, they also demonstrate the advantage of adding skip connections and augmentations. The workflow presented by Alaudah et al. (2019) introduces a great starting point for a segmentation baseline, but fails to produce highly accurate segmentation results especially for datasets with complex geology. The second model introduced by Chen et al. (2022), utilised an encoder decoder architecture with three additions to enhance the segmentation performance. First, spatial pyramid sampling (SPS) for training sample selection. Second, multimodal fusion (M2F) module to build multimodal representations. Finally, a local-to-global (L2G) module, to help improve the recognition power and enhance the segmentation. The model developed by Chen et al. (2022) does improve on the results previously presented in Alaudah et al. (2019) but failed to produce high quality predictions in complex areas of the
seismic profile and segmentation artifacts could still be observed in the predictions. In the results section, we show in the results section that the frequency constrained attention U-Net renders better performance in seismic facies classification tasks in comparison to the aforementioned benchmarks.

The remainder of this article is organised as follows: in Section II, we introduce the proposed model where we list the details of residual building blocks, the utilised attention mechanism and spectral decomposition for feature engineering. We then demonstrate the effectiveness of the model using field applications in Section III. The first dataset we conduct experiments on is the Netherlands F3 block dataset (dGB Earth Sciences, 1987), and the second dataset is the Penobscot dataset acquired offshore Nova Scotia, Canada (dGB Earth Sciences, 2017). In section IV, we analyse the performance of the proposed model quantitatively and compare the results with published benchmarks. Finally, the main conclusions are drawn in Section V.

METHODS

We adopt a deep-learning model that incorporates important concepts from other methods, and it is these concepts that seem to be central to the success of the model on realistic datasets. Although the genesis of our approach originally lay elsewhere, we discuss these concepts and how they were infused to produce an original scheme derived from the standard approach.
Convolutional Neural Networks (CNNs)

Deep convolutional neural networks are a subset of deep neural networks. The architecture of CNNs consist of a series of convolutional layers that filter the original image (Fukushima, 1980; LeCun et al., 1989). The first part of the network focuses on feature extraction, this is performed using stacked convolutional layers followed by nonlinear activation functions and pooling. The role of convolutional layers is to detect local conjunctions of features from previous layers such as object shapes and boundaries (Walderland et al., 2018). Activation functions add nonlinearity to help the network fit more complex problems. The role of pooling is to merge semantically similar features into one condensed feature that preserves the most characteristic features. The combination of these elements supports the network to effectively detect characteristics of input data. The second part of the network works on classification, this is done using fully connected layers, these layers combine features derived from previous layers into vectors that are mapped to classification outputs.

The gist of CNNs takes advantage of a signals compositional hierarchies, in which higher level features are obtained by composing lower level ones (LeCun et al., 2015). In images local combinations of edges form motifs, motifs assemble into parts, and parts into objects. Similar hierarchies exist in sound, such as speech and text that are formed from syllables, words and sentences. This makes CNNs ideal for analysing signals such as seismic signals.
U-Net architecture

On of the most robust models for semantic segmentation tasks is the so called “U-Net” architecture (Ronneberger et al., 2015). Figure 1 displays the standard U-Net architecture that was first introduced to solve image segmentation for biomedical applications (Ronneberger et al., 2015). The architecture consists of a contractive path and an expansive path linked via skip connections. The contractive path, or the encoder, extracts features from the input image, whilst the expansive path, the decoder, reconstructs the image using upsampling operations such as bilinear interpolation and transposed convolution. The network is organised into building blocks made up of convolutional layers followed by activations and sampling operations. The contractive path consists of four groups, each building block consists of two $3 \times 3$ convolutions followed by a ReLU activation function and a $2 \times 2$ max-pooling operation with a stride of 2 for down-sampling, in which the feature maps prior to the max-pool operator are passed via the skip connection to the decoder. At each down-sampling step the spatial dimension is halved whilst the number of feature channels is doubled. The expansive part is also made up of four groups, of one up-sampling layer and two $3 \times 3$ convolutions each followed by a ReLU activation. Up-sampling doubles the spatial dimension, and halves the feature channels before being concatenated to the feature map from the contractive part via the skip connection. The final layer is a $1 \times 1$ convolutional layer applied to map the number of features to the desired number of classes. In total the network has 23 convolutional layers. The outputs are then used for pixel-wise predictions.
Residual building blocks

Figure 2 compares a standard feed forward building block to a residual block with shortcuts “skip connections”. Standard blocks are fed with an input $x$ and produces an output $y = F(x, W_i)$. Instead residual blocks uses an additional identity shortcut to map the input to the output following $y = F(x, W_i) + x$. Skip connections lead to easier optimisation of the neural network and avoids the problem of vanishing gradients especially for deep networks. In the proposed model, we replace standard convolutional blocks of the U-Net encoder with residual building blocks. Unlike the standard U-Net that uses pooling operations to down-sample, residual blocks use strided convolution to down-sample. Furthermore, batch normalisation (BN) is adopted after each convolution layer and before activation to normalise the contributions to a layer for every mini-batch. Table 1 summarises the details of each layer of the proposed encoder for the proposed architecture. After adding residual building blocks the network now has 29 convolutional layers in total.

Attention blocks

A desirable trait in image processing is the networks ability to focus on input areas that are harder to segment. Attention in a U-Net could be achieved though the addition of attention blocks that trims unnecessary details that are not useful to the on-going task and focuses on important details. In the proposed architecture, each layer in the contractive path passes through an attention block before being
passed and concatenated to the up-sampled features in the expansive path via skip connections. We note that the memory and computational costs of attention modules grow quadratically with the input size. As we will demonstrate in the results section, adding attention after each block in the contractive path significantly improves the segmentation performance and justifies the added computational cost of including these modules in the proposed architecture.

In this paper, we specifically focus on the attention mechanism employed in the transformer model proposed by (Vaswani et al., 2017). Figure 3 presents a concise illustration of the attention block used in the transformer architecture, known as “Scaled Dot-Product Attention”. This attention function maps a query vector, along with a set of key-value pairs, to generate an output vector. The output vector is computed as a weighted sum of the values, where the weight assigned to each value is determined by the compatibility function between the query and its corresponding key. In our implementation, the input consists of keys and queries with a dimension of $d_k$, as well as values with a dimension of $d_v$. To obtain the weights for the values, we compute the dot product of each key and query, divide the result by $\sqrt{d_k}$, and subsequently apply a softmax function. For computational efficiency, we perform the attention function on a batch of queries simultaneously, which are packed together into a matrix $Q$. Similarly, the keys and values are packed into matrices $K$ and $V$, respectively. The resulting output matrix is computed as follows:

$$Attention(K, Q, V) = \text{softmax} \left( \frac{Q K^T}{\sqrt{d_k}} \right) V. \quad (1)$$

Two commonly used attention mechanisms are additive attention (Oktay et al.,
2018; Schlemper et al., 2019) and dot product (multiplicative) attention. While the attention used in this work is similar to multiplicative attention, the main distinction lies in the absence of the scaling factor \(1/\sqrt{d_k}\). Additive attention employs a feed-forward network with a single hidden layer to compute the compatibility function. In terms of theoretical complexity, both mechanisms exhibit similarities. However, in practice, dot product attention is significantly faster and more space-efficient due to its utilisation of highly optimized matrix multiplication algorithms. While the two mechanisms perform comparably for small values of \(d_k\), additive attention surpasses dot product attention without scaling for larger values of \(d_k\) (Britz et al., 2017). It is suspected by Vaswani et al. (2017) that, as \(d_k\) increases, the magnitudes of dot products grow large, which pushes the softmax function into regions with extremely small gradients. To mitigate this issue, the dot product is scaled by \(1/\sqrt{d_k}\).

Figure 4 demonstrates the U-Net architecture augmented with attention blocks positioned after each residual block in the encoder. Within these attention blocks, the keys, values, and queries originate from the same source, specifically the output of the preceding encoder layer. This design ensures that each position within the encoder attends to all positions within the previous layer of the encoder.

The incorporation of the Scaled Dot-Product Attention mechanism into the U-Net architecture offers several advantages in the context of segmentation tasks. Firstly, it enables the model to effectively capture long-range dependencies among spatially distant locations, thereby improving the understanding of global context and enhancing segmentation decision-making (Vaswani et al., 2017). Secondly, the attention mecha-
nism allows the model to dynamically prioritise informative regions, leading to sharper and more precise segmentation boundaries through the allocation of higher attention weights to relevant areas, as demonstrated in Appendix B (Rodriguez et al., 2019). Thirdly, in scenarios involving complex and cluttered scenes, the attention mechanism selectively attends to pertinent regions while suppressing distracting features, resulting in improved discrimination between objects and background (Rodriguez et al., 2019). Lastly, the attention mechanism ensures adaptability to varying object scales, empowering the model to handle objects of different sizes and capture both local details and global context, which are critical for accurate segmentation (Oktay et al., 2018). The motivation for incorporating attention into our network is to achieve higher accuracies, particularly in challenging segmentation tasks involving low-quality datasets. By integrating the Scaled Dot-Product Attention mechanism, the segmentation performance of the U-Net model is significantly enhanced. This improvement is attributed to the mechanism’s effectiveness in capturing dependencies, directing attention, managing complexity, and accommodating variations in object scale.

Spectral decomposition

Wavelet transform is a mathematical tool, which has become an indispensable part of numerous signal processing algorithms. Transforming a signal in the time domain into another form (e.g. time-scale, time-frequency) makes the signal more amenable to study and enables the signal features to be described more succinctly (Alsalmi and
Wavelet transform (WT) has been found to be particularly useful for analysing signals which can best be described as aperiodic, noisy, intermittent, transient and so on (Vrhel et al., 1997). The continuous wavelet transform (CWT) is defined by the inner product between a family of wavelets $\psi \left( \frac{t-b}{a} \right)$ and a signal $x(t)$. The CWT of a continuous signal $x(t)$ can then be expressed as (Daubechies, 1988, 1992):

$$W(a, b) = \frac{1}{\sqrt{a}} \int x(t) \psi^* \left( \frac{t-b}{a} \right) dt,$$

(2)

the asterisk ($^*$) indicates the complex conjugate of the wavelet, and $1/\sqrt{a}$ is used for energy conservation. In our implementation we use the Morlet wavelet as the mother wavelet in running CWT spectral decomposition. The Morlet wavelet is complex, this results in the wavelet transform $W(a, b)$ solution to also be complex. We take the power spectrum of the solution as $|W(a, b)|^2$.

It is important to note that the CWT produces a scalogram and not a time-frequency map. To transform the scalogram into a time-frequency map the simplest approach is to use the relation between scale ($s$) and the central frequency of the wavelet ($f_c$). We then estimate the pseudo-frequency associated with the scale by $f = f_c/s$ as the scale is inversely-proportional to frequency. Another key point is that a single seismic trace produces a 2D time-frequency spectrum. For a 3D cube the result is a 4D cube with the extra dimension being frequency. We use this 4D cube to extract a single frequency or the sum of a frequency range to generate a 3D spectral cube.
The rationale behind incorporating additional frequency channels into the model’s input is to exploit the frequency-related information present in the image, enabling the model to better capture nuanced details and improve overall performance. The integration of additional frequency decomposition channels into the input image enhances the model’s capacity to capture and represent diverse frequency components, effectively mitigating data discrepancies and variations. This approach entails decomposing the image into distinct frequency bands, thereby granting the model access to a more comprehensive representation. The incorporation of these frequency channels amplifies the model’s resilience against variations, facilitates efficient feature extraction, augments discriminative power, and fosters a comprehensive understanding of the image’s frequency content. Consequently, the inclusion of frequency decomposition channels significantly enhances the model’s performance by adeptly addressing frequency-based discrepancies, variations in texture and scale, and capturing crucial details (Alsalmi and Wang, 2021).

It is important to highlight that simply adding additional frequency channels to the input, without incorporating attention, will not lead to high-quality segmentation. This is because attention plays a crucial role in establishing spatial connections between these channels and assigning more weight to channels that contain more relevant information for the segmentation task. Conversely, utilising attention with only seismic and depth inputs is also suboptimal since the seismic section alone lacks sufficient information to enhance the discriminative ability of the networks.

To achieve high-quality segmentation results, it is crucial to combine both atten-
tion mechanisms and extra frequency channels. This synergistic combination forms a robust recipe that effectively leverages the spatial connections identified by attention and incorporates the supplementary information provided by the additional frequency channels. The effectiveness of this proposed approach will be demonstrated in detail in the forthcoming results section.

**EXPERIMENTAL DETAILS**

Two seismic datasets from subsurface reservoir exploration are chosen to demonstrate the effectiveness of the deep-learning model, each of which exemplifies different beneficial characteristics of the method. In this section, we briefly describe the survey areas and the characteristics of the identified facies within each area.

**Dataset 1**

The North sea continental shelf, located off-shore the Netherlands, is a hydrocarbon rich area divided into geographical zones. In 1987, a seismic survey to explore for hydrocarbon reservoirs was shot over a rectangular area of dimensions $16 \times 24$ km. This block, also known as the F3 block, became one of the most known and studied blocks after dGB Earth Sciences made the data obtained from the survey publicly available (dGB Earth Sciences, 1987).

Ten lithostratigraphic units within the North sea have been identified in literature (Van Adrichem Boogaert, 1993; Scheck-Wenderoth and Lamarche, 2005; Duin et al.,
In order to compare our results with the benchmarks presented in literature we used the annotated survey published by Alaudah et al. (2019). The dataset released is a fully annotated 3-D geological model of the F3 Block contains 6 annotated units. Figure 5 shows a 3D view of the fully annotated F3 dataset. The annotated units are (from newest to oldest): the Upper North Sea group, the Middle North Sea group, the Lower North Sea group, the Chalk and the Rijnland groups combined, the Scruff group, and the Zechstein group. Table 2 shows the percentage of pixels that belong to each class from our training set. The dataset is sampled at 4 ms and the seismic cube used in this study consists of 601 inlines (inline number 100-700), 901 crosslines (crossline number 300-1200) and 255 depth samples in the range of 0.828-1.848 s (or 1005-1877 m). The characteristic of the facies within the six regions can be described as follows (from shallow to deep) (Alaudah et al., 2019; Silva et al., 2019):

- Facie 1: comprises the Upper North Sea Group. Facies are characterised by mound shaped low amplitude reflectors.
- Facie 2: contains the Middle North sea group. Facies are semi-continuous and low amplitude.
- Facie 3: corresponds to the Lower North Sea group. Facies are semi-continuous and low amplitude.
- Facie 4: combines the Chalk group and the Rijnland group. The lower portion shows medium to low amplitude progradational sigmoidal features. The upper
portion is made of chalk and shows high amplitude reflectors.

- Facie 5: is made of the Scruff group. Facies are low amplitude and continuous with marked polygonal fault portions that create characteristic mound shaped discontinuities.

- Facie 6: comprises of the Zechstein group evaporites and carbonates of Zechstein. The top part shows reflectors with stratified facies, while no apparent reflections are visible on the lower portion.

For the test/train division we follow the same split introduced in the benchmark in order to fairly compare the results. Figure 6 shows the data divided into three sets as follows:

1. Training set: This includes all the data in the range of inlines [300,700] and crosslines [300,1000].

2. Test set-1: This set includes all the data in the range of inlines [100,299] and crosslines [300,1000]. This set was used for testing.

3. Test set-2: This sets includes all the data in the ranges of inlines [100,700] and crosslines [1001,1200]. This set includes a large Zechstein diapir in the NE of the survey that is never seen in the training set. This set was used for validation.

The dataset depicts features including faults, salt domes, and regional unconformities posing a challenge for the segmentation model.
Dataset 2

The second 3D dataset was acquired offshore Nova Scotia Canada located specifically in the Scotian basin on the Scotian shelf. This dataset is known as the Penobscot dataset and was acquired by the Nova Scotia Department of Energy and the Canada-Nova Scotia Offshore Petroleum Board, and is publicly available on the dGB Open Seismic Repository (dGB Earth Sciences, 2017). The basin was formed during the breakup of Pangea and covers a total area of 300000 km², with maximum sediment thickness of 18 km. The seismic survey consists of 87 km² post-stack time migrated 3D reflection seismic data. Figure 7a shows the seismic cube which is composed of 601 inlines and 482 crosslines, with bin size 12.5 m × 25.0 m. The total time range is 6000 ms sampled at 4 ms. The signal has a high SNR in the region above 3 ms, or 5 km depth, whereas the deeper region has lower resolution making it challenging to interpret.

Stratigraphy interpretation for the Penobscot dataset was produced based on the internal seismic characteristics and seismic reflector separating the boundaries. Seven horizons (H1-H7) were made publicly available (Baroni et al., 2019), we used these horizons to produce segmentation labels for the seismic data by labelling pixels within interval regions where the horizon intersects with the seismic data. In total eight classes of facies were identified, Figure 7 b shows the labelled data. The class labels are highly imbalanced, where some of the facies in the dataset have very few numbers of samples. Table 5 shows the percentage of pixels that belong to each class in the dataset. Facie 1-3 make up more than 80% of the dataset. These are the deepest
facies within the dataset where the seismic resolution is poor, resulting in these areas being challenging to train.

The characteristics of the facies within the eight regions can be described as follows (from deep to shallow) (Baroni et al., 2019):

- **Facie 1**: This region is mostly parallel, concordant, high-amplitude reflectors. Chaotic reflectors are also identified and are attributed to the decreased frequency content of the seismic wave with depth.

- **Facie 2**: is characterised by parallel to sub-parallel, continuous, high-amplitude reflectors. This configuration reflects the deposition of fluvio-deltaic sediments of the Mississauga Formation.

- **Facie 3**: consists of sub-parallel reflectors, but less continuous than the previous sedimentary package.

- **Facie 4**: is composed of low amplitude horizons corresponding to the package consisting of deep marine shales and limestone showing little lithological contrast.

- **Facie 5**: is characterised by sub-parallel reflectors of varying amplitude.

- **Facie 6**: mostly consists of parallel, high amplitude reflectors. A region of chaotic reflectors can be seen and is associated with marine slump deposits.

- **Facie 7**: is made up of high amplitude reflectors. Most reflectors are continuous whilst others are diving angles and truncated, indicating typical characteristics
of a high energy environment.

- Facie 8: identifies the region above the seabed.

The seismic cube was divided into three segments for the test-validation-train splits. Figure 8 shows the seismic cube with highlighted regions. Seismic inlines in the range of [0-300] are highlighted in red and used for training. The region highlighted in yellow (inline 300-400) are used for validation and the green highlighted region is used for testing (inline 400-600). To evaluate the performance of our model, we use the same set of evaluation metrics commonly used in computer vision literature. These metrics were also used by the benchmarks for evaluation. Details and equations for the accuracy measure are presented in Appendix A.

Machine learning implementation details

The model is trained using the Adam optimiser (Kingma and Ba, 2014), for a total of 200 epochs to optimise the cross-entropy loss function. To compensate for the highly imbalanced class representation, we use class-weights for the cross-entropy loss function this stimulates the network to obtain a higher accuracy for those classes. We use a cyclic learning rate scheduler for the optimiser that fluctuates between two boundaries 0.001 and 0.0001 after every batch (Smith, 2017). We use standard data augmentation techniques, including randomly rotating the patches by up to ±15°, adding random Gaussian noise with a mean of 0 and standard deviation of 0.02, randomly flipping the sections horizontally and vertically, taking random cutouts of...
size 8 × 8, and random distortions like elastic transform, grid distortion and optical
distortion. These augmentations were invoked at probabilities of 0.7, 0.3, 0.5, 0.5,
1, 0.8, 0.5, and 0.5, respectively. Figure 9 shows a seismic patch and some of the
augmentations used during the training process. To preserve the spatial and contex-
tual relationships between different channels, augmentations are applied uniformly to
all input channels. Specifically, in the case of vertical flipping, the order of geologic
layers is maintained by the depth channels, which effectively preserve the sequence
order. We emphasise that the validation set was not used to change hyperparameters
but to validate the convergence of the model. We implement all the experiments in
PyTorch (Paszke et al., 2019) on a single NVIDIA TITAN RTX graphics processing
unit (GPU) with 24 GB memory.

Workflow

In this section, we present details of the preprocessing techniques adopted to prepare
the seismic data for both the training and testing phases. Both test cases were split
into three sections that were used for training, validation and testing (see Figure 6
and Figure 8). The red highlighted area was used for training, yellow for validation
and green was used for testing. Each split was only used once. Both 3D seismic
cubes are normalised to a range of [-1,1] as a prepossessing step. For the training and
validation phase each seismic profile was divided into patches (128 × 128), whereas full
seismic profiles (unpatched) were used for testing. To ensure that depth information
is preserved when using patches we create a 3D depth cube. Each seismic patch
then has a corresponding depth patch which is added as an additional channel to the seismic amplitude patch.

Next we add spectral features generated using the CWT. These additional features reveal information embedded in the seismic trace that are not readily apparent in the raw seismic trace. Spectral data are particularly chosen for their ability to capture transient features within each seismic trace. CWT operates by convolving a signal with a wavelet to measure similarity. By comparing a signal to a wavelet at various scales a function of two variables is obtained (time and scale). A 4D spectral cube is then produced for each 3D seismic cube.

To identify the frequency range with the dominant spectral energy, the first step is to generate a Fourier power spectrum. This is useful for the selection of the relevant pass-band cubes generated from the CWT. Figure 10 shows the power spectrum estimation of the F3 seismic dataset using the Fourier transform. The dominant spectral energy is identified to be within the range between 10-40 Hz. The same step is conducted for the Penobscot dataset. Figure 11 shows the Fourier power spectrum for the dataset. The range of 10-40 Hz is also selected as the peak range. Using the frequency range identified from the power spectrum we extract three frequency cubes representative of the low-frequency, mid-frequency and high-frequency segments for the range between 10-40 Hz. Spectral data corresponding to each range is then extracted from the 4D frequency cube forming three independent 3D cubes. Three additional channels are then added to the existing two-channel patch which contains seismic amplitude (channel 1) and depth (channel 2). The additional spectral channels...
represent bandlimited versions of the full-bandwidth seismic in channel 1.

Figure 12 summarises the workflow. Each dataset has a total of five 3D cubes, this includes seismic amplitude, depth, low-frequency, mid-frequency and high-frequency cubes. To generate an input for the model we take a single 2D profile from each cube and stack it together to form a 5 channel input. This is then processed as patches to fit in the memory of the GPU. The 5 channel patches are now fully processed and are fed to the proposed model. The output is then a single 2D segmented patch to represent the different class labels. This is then repeated for all the patches in a single profile and then repeated for all the profiles within the input cube to generate a fully segmented cube. In the training phase, the output is compared to the ground-truth and the loss is backpropagated to tune (optimise) the model weights and biases. The trained model is then used to predict class labels on the test set.

RESULTS AND DISCUSSION

After processing the input data, the proposed U-Net model is trained on data patches until the validation loss converges. Training is performed on patches while testing is performed on sections. The challenge in seismic image segmentation is attributed to features not being apparent from the seismic image. Additional features such as the spectral channels are used in this work to help unfold intricate features within the seismic image. Another challenge that occurs in many practical geoscience applications, such as facies segmentation in seismic images, is unbalanced class labels. This leads to difficulties in training a neural network (Buda et al., 2018). This does
not effect the total accuracy of the model because of the weighted sum nature of the accuracy calculation. However, a models accuracy in segmenting pixels from less frequent classes is considered highly favourable and improves the ability of the model to generalise. We use four ways to mitigate the effects of data imbalance. First, augmentations to help improve the models generalisation ability. Second adding class weights to the loss, these weights penalise the network more when misclassifying less frequent classes, this promotes improved performance on these classes. Third, adding extra input features to help the network delineate distinct features associated with each class. Last, adding attention blocks which implicitly focuses on different regions of the input and results in enhanced performance.

F3 dataset

Figure 13 shows test results for a series of inlines sampled equally to represent the full range of the test set of the F3 block. The predicted segmentation results highly correlate with the ground truth with a pixel accuracy of 95.6%. The accuracy of the trained model on the test dataset represents the ability to accurately segment pixels and determine boundaries between the facies on a portion of the dataset not seen before while training.

Figure 14 showcases a comprehensive overview of the confusion matrix, including accuracy metrics for our results. The rows correspond to the predicted classes, while the columns represent the true class. Correct observations are represented along the diagonal, while off-diagonal cells indicate incorrectly classified pixels. In each cell,
the displayed percentage represents the number of pixel observations relative to the total number of pixels in the dataset. It is evident from the results that the number of correctly classified pixels outweighs the number of incorrect classifications. In the confusion matrix, the rightmost column displays precision values in green and false positives in red, while the bottom row exhibits recall values in green and false negatives in red. Precision evaluates the accuracy of positive predictions by calculating the ratio of correctly predicted positive instances to all predicted positives, whereas recall gauges the inclusiveness of actual positive instances by calculating the ratio of correctly predicted positive instances to all actual positives. For all classes, most of the precision and recall values are high, demonstrating their substantial superiority over false positives and false negatives.

Class 1-3 are the majority classes, these classes have the highest recall accuracy measures. This is expected due to the availability of more pixels belonging to these classes in the training dataset. The results from previous published work show particularly low recall accuracy for classes 4-6. Class 4 is composed of carbonates, and clay with sandstones from the Chalk and the Rijnland groups, class 5 is claystone that belongs to the Scruff group and class 6 are evaporites that belong to the Zechstein group. Harder classification within these groups could be due to a number of reasons, first all these classes are the minority classes with the least number of pixel samples to train on. Second, the geologic structure of these classes is much more complex in comparison to shallower classes (class 1-3). Third, reflections within these groups have obscure seismic signatures making them harder to classify, this is especially the
case for evaporites. The prediction accuracy demonstrated by the proposed model for these classes is greater than 80%. This significantly outperforms results obtained using other published methods specifically on less frequent classes. The overall accuracy of the model is indicated by the bottom-right cell, which shows a value of 95% and an overall error rate of less than 5%. The total Intersection over Union (IOU) is 84.6%, the Frequency Weighted Intersection over Union (FWIU) is 89%, and the Mean Class Accuracy (MCA) is 92%. These metrics demonstrate the model’s notable accuracy and effectiveness across various evaluation aspects.

Figure 15 showcases the results of a test conducted to substantiate the rationale for incorporating frequency as an additional input in our model. Figure 15a displays a series of seismic inlines, while Figure 15b showcases their corresponding segmentation labels. Figure 15c illustrates the results obtained by solely utilising seismic and depth channels as input to the model, these results highlight the evident inadequacy of seismic data alone in effectively delineating the boundaries between different facies. Despite the inclusion of attention, which leverages information from multiple channels to establish meaningful relationships and enhance segmentation results, seismic data alone remains insufficient. The attention mechanism’s attempt to establish relationships among characteristics within the same class solely based on seismic information falls short in capturing the intricate details and edges between these classes. Consequently, seismic data alone fails to encompass and emphasise the diverse scales and inherent boundaries present in these classes.

On the other hand, Figure 15d showcases the outcomes obtained by employing
our proposed model, which includes the incorporation of supplementary frequency channels. Notably, a significant improvement is observed in the results. The edges of the facies are more accurately captured, and the continuity among the facies is greatly enhanced. The inclusion of frequency information enables a more comprehensive representation of the data, leading to superior performance in highlighting the edges and preserving the coherence of the facies.

To further demonstrate the performance of our proposed method, we perform an ablation study and present segmentation results for three variations of the U-Net architecture: (a) Standard U-Net with standard convolutional blocks and max-pooling operations to down-sample (b) U-Net with residual building blocks in place of the standard blocks. Max-pooling operations are replaced with strided convolutions. (c) U-Net with residual and attention blocks with additional spectral inputs to demonstrate the effectiveness of the proposed features.

Table 3 shows a comparison between the accuracy measures of the three U-Net architectures, the predicted segmentation results of the three models are presented in Figure 16. Standard U-Net has the lowest mean class accuracy of 84.3%. The residual block U-Net demonstrate the advantages of using residual blocks in-place of basic blocks and produces a mean class accuracy score of 89.4%. This model performs better than the standard U-Net due to the additional shortcut connections between the blocks, these connections allow the computed gradients to propagate to lower layers of the network encouraging enhanced learning. Despite that, the accuracy for the underrepresented classes (classes 4-6) are still low due to the model not being
fully equipped to handle complex datasets. The final predicted results are those of
the proposed model, this model combines residual blocks with attention blocks and
spectral inputs. The performance of this model significantly surpasses the other two
models especially for the less frequent classes. The mean class accuracy for this model
is 92.0% and the accuracy of all the classes (including classes 4-6) is more than 80%.
It is clear that the predicted segmentation results for the proposed model shown in
Figure 16e exhibit the highest correlation with the ground truth.

Comparison to published benchmarks

In this section we compare the segmentation accuracy of the proposed model with
state-of-the-art models presented in literature. In particular two deep learning models,
a facies segmentation benchmark presented by Alaudah et al. (2019), and a second
method presented by Chen et al. (2022), both applied to the F3 dataset. Table 4
shows the accuracy metrics used to measure and compare the performance of the
proposed segmentation model to benchmarks from literature. The pixel accuracy of
the proposed model is 95.2% this value is higher in comparison to model 1’s score of
90.5% and model 2’s score of 93.4%. Based on the accuracy metrics shown in Table 4
the proposed model achieves higher performance over the state-of-the-art semantic
segmentation models. The model proposed by Chen et al. (2022) performs equally
well on class 1-3. However, when we take a closer look at the pixel accuracy of the
three less frequent classes (4, 5 and 6), we observe that the proposed model produces
considerably higher results to those obtained from the two other benchmarks.
Figure 17 compares segmentation results of the three models for inline 200. Results achieved from the benchmark models are promising but there is a clear gap for further improvement. This is especially true for the less frequent classes (class 4-6). Alaudah et al. (2019) show facies mixing and inaccurate facies edge delineation towards the lower left of the image. The model proposed in Chen et al. (2022) improves on the previous results but still suffer from facies mis-classification (lower right corner), additionally we notice segmentation failing to cover the full span of the facie (lower right corner). Results obtained using the proposed framework are highly accurate and outperform state-of-the-art benchmarks.

Penobscot dataset

Figure 18 shows the results for a series of inlines sampled equally to represent the full range of the test-set of the Penobscot dataset. The predicted results highly correlate with the ground truth. The seismic quality below 3 ms is low making it more challenging to predict in the regions corresponding to facie 1-4. Figure 19 shows a comprehensive overview of the confusion matrix. The results clearly demonstrate a higher number of correctly classified pixels compared to incorrect classifications for each of the 8 classes. Even in regions where seismic resolution is low and classification is more challenging (classes 1-4), the model showcases high accuracy values, demonstrating its capability to perform well in difficult areas. The model exhibits a pixel accuracy of 96.7% with an error rate of less than 3%, while achieving a MCA of 92.6%. Additionally, the IOU score is 85.4%, and the FWIU score is 94%. The perfor-
mance of the proposed method in accurately segmenting seismic facies, particularly in regions with low seismic resolution, is further validated by the results obtained from the second dataset.

CONCLUSIONS

We introduced a new deep learning U-Net based model architecture for seismic facies segmentation. The proposed model adapts three main modifications to the standard U-Net model (a) replacing standard convolution blocks with residual blocks that benefit from shortcut connections (b) adding attention blocks for more accurate facies segmentation, and (c) adding depth and spectral channels as input features in addition to the seismic amplitude. This model is general and can also be applied to other segmentation tasks.

We present result for two challenging field examples. Furthermore, we analysed the advantages of the proposed adaptations via a detailed ablation study to demonstrate the contributions of each added component to the standard U-Net model. We also compared the results to published benchmarks. The presented results clearly show that the proposed model outperforms state-of-the-art models when applied to the same dataset. Results are particularly enhanced in areas that contain minority classes with fewer pixels in the training dataset. This shows superior generalisation ability of the proposed model when compared to other models. In the future, we aim to combine this work with semi-supervised learning approaches to handle seismic surveys with limited training data.
ACKNOWLEDGMENTS

This work is funded by the PETRONAS Centre of Excellence in Subsurface Engineering and Energy Transition (PACESET). PACESET is based at Heriot-Watt University in Scotland, United Kingdom.

APPENDIX A

METRICS

Pixels that are correctly classified to a specific class $i$, are calculated by the intersection between the pixels that belong to that class $G_i$, and the pixels classified to that class $F_i$. The set of correctly classified pixels are $G_i \cap F_i$. The metrics used are defined as follows:

1. Pixel Accuracy (PA) is the percentage of pixels over all classes that are correctly classified,

$$PA = \frac{\sum_i |F_i \cap G_i|}{\sum_i |G_i|}. \quad (A-1)$$

2. Class Accuracy for class $i$ ($CA_i$) is the percentage of pixels that are correctly classified in a class $i$.

$$CA_i = \frac{|F_i \cap G_i||G_i|}{|G_i|}. \quad (A-2)$$

We will also define the mean class accuracy (MCA) as the average of $CA$ over all classes,

$$MCA = \frac{1}{n_c} \sum_i CA_i = \frac{1}{n_c} \sum_i \frac{|F_i \cap G_i|}{|G_i|}, \quad (A-3)$$
where $n_c$ is the number of classes.

3. Intersection over Union (IOU) is defined as the intersection of elements in $G_i$ and $F_i$ over the number of elements of their union set,

$$\text{IOU}_i = \frac{|F_i \cap G_i|}{|F_i \cup G_i|}.$$  \hspace{1cm} (A-4)

This metric measures the overlap between the two sets and should equal one if and only if all the pixels were correctly classified. The average IOU over all the classes produces the mean intersection over union (Mean IOU),

$$\text{Mean IOU} = \frac{1}{n_c} \sum_i \text{IOU}_i = \frac{1}{n_c} \sum_i \frac{|F_i \cap G_i|}{|F_i \cup G_i|}.$$  \hspace{1cm} (A-5)

To prevent this metric from being overly sensitive to small classes, it is common to weigh each class by its size. The resulting metric is known as frequency-weighted intersection over union (FWIU),

$$\text{FWIU} = \frac{1}{\sum_i |G_i|} \cdot \sum_i |G_i| \cdot \frac{|F_i \cup G_i|}{|F_i \cap G_i|}.$$  \hspace{1cm} (A-6)

APPENDIX B

ATTENTION WEIGHTS

The attention weights displayed in Figure A-1b-d illustrate the attention weight output of each layer, representing various levels of the encoder in the U-Net architecture. The attention weight corresponds to the output generated immediately after applying the softmax function within the attention block, just prior to the matrix multiplication step involving the value vector. Notably, these weights exhibit higher values at
the boundaries between facies, indicating the network’s heightened emphasis on these regions. This observation offers valuable insights into the network’s decision-making process, showcasing its ability to effectively capture and emphasise the intricate details and transitions among different geological formations. By focusing on these boundaries, the network demonstrates its capacity to discern subtle variations in the seismic data that correspond to distinct facies. This heightened attention to detail greatly contributes to the network’s accuracy in analysing and interpreting seismic profiles, making it particularly valuable for tasks such as facies classification.
REFERENCES


Brown, A. R., 2011, Interpretation of three-dimensional seismic data—7th edi-
tion: AAPG Memoir 42/SEG Investigations in Geophysics, No. 9; doi: 10.1306/m4271346.


Daubechies, I., 1988, Orthonormal bases of compactly supported wavelets: Communications on pure and applied mathematics, 41, 909–996; doi: 10.1002/cpa.3160410705.


Di, H., M. A. Shafiq, Z. Wang, and G. AlRegib, 2019, Improving seismic fault detection by super-attribute-based classification: Interpretation, 7, SE251–SE267; doi:


LeCun, Y., B. Boser, J. S. Denker, D. Henderson, R. E. Howard, W. Hubbard, and


Mitchum Jr, R., P. R. Vail, and S. Thompson, 1977a, Seismic stratigraphy and global changes of sea level: Part 2. the depositional sequence as a basic unit for stratigraphic analysis: Section 2. application of seismic reflection configuration to stratigraphic interpretation, In Payton (1977), 53–62.


Scheck-Wenderoth, M., and J. Lamarche, 2005, Crustal memory and basin evolution


Tschannen, V., M. Delescluse, M. Rodriguez, and J. Keuper, 2017, Facies classi-


LIST OF TABLES

1. Encoder building blocks for the proposed architecture.
2. Percentage of class labels in the F3 dataset.
3. Accuracy comparison of three U-Net architectures for the F3 dataset.
4. Comparison of accuracy metrics between the proposed model and the state-of-the-art benchmarks for the F3 dataset.
5. Percentage of class labels in the Penobscot dataset.
LIST OF FIGURES

1 Standard U-Net architecture. The contracting path on the left is the encoder and the expanding path on the right is the decoder. Arrows show direction of information flow. Coloured blocks represent particular layers as per the legend. The numbers represent the output channels. “I” represents the spatial size of the image, downscaled by the pooling operators (red) and upscaled by transposed convolution operators (blue).

2 Comparison of building blocks with an input \( x \) passing through convolution layers \( w_i \) and mapped to \( F(x) \) in (a) a feed forward convolutional block and (b) a residual block with an identity shortcut connection.

3 The attention block illustrates the computation of the Query, Key, and Value matrices. The input matrix \( X \) is multiplied by the weight matrices, followed by scaling and applying the softmax function to produce the attention. The attention is then multiplied with the value matrix to generate the output of the self-attention layer.

4 Proposed U-Net architecture with residual and attention blocks. Arrows show direction of information flow. Coloured blocks represent particular layers as per the legend. The numbers represent the output channels. “I” represents the spatial size of the image, downscaled by strided convolutions (pink, orange) and one max-pooling operation (red) and upscaled by transposed convolution operators (blue). The black dotted line are skip connections and the blue dotted arrow are shortcut connections for identity mapping.

5 (a) F3 seismic cube and the (b) annotated 3D cube showing the 6 class la-
bels.

6 F3 cube divided into the training set (red shaded area), the validation set (yellow shaded area) and the test set (green shaded area).

7 (a) Penobscot seismic cube and the (b) annotated 3D cube showing the 8 class labels.

8 Penobscot seismic cube divided into the training set (red shaded area), the validation set (yellow shaded area) and the test set (green shaded area).

9 Example of augmentations applied to (a) seismic patches and (b) segmentation label patches.

10 Power spectrum for the F3 seismic cube.

11 Power spectrum for the Penobscot seismic cube.

12 Summary of the workflow showing the input patch of size 128 × 128 × 5 fed to the U-Net. The patch consists of five layers composed of seismic, depth and three frequency profiles.

13 (a) Seismic profiles, (b) the ground truth labels, and (c) the predicted results using the proposed U-Net.

14 Confusion matrix illustrating the performance of the proposed model on the F3 dataset. Each cell represents the percentage of observed pixels relative to the total number of pixels in the dataset. The diagonal represents correctly classified pixels, while the off-diagonal represents incorrectly classified pixels. The rightmost column presents precision values in green, indicating true positives, and false positives in red. The bottom row shows recall values in green, representing true positives, and false negatives in red.
15 (a) Seismic profiles and (b) the ground truth labels. A comparison of the predicted labels using (c) no additional frequency channels to the input and (d) the proposed U-Net with frequency channels added to the input.

16 (a) Seismic profiles and (b) the ground truth labels. A comparison of the predicted labels using (c) a standard U-Net, (d) a U-Net with a residual blocks, and (e) the proposed U-Net.

17 (a) A single seismic profile, (b) ground truth labels for the single seismic profile, (c) segmentation result reproduced after Alaudah et al. (2019), (d) segmentation result reproduced after Chen et al. (2022), and (e) segmentation result achieved by our proposed method.

18 The first column shows an array of seismic inlines extending throughout the entire Penobscot dataset. The second column shows the ground truth labels and the third column shows the predicted labels.

19 Confusion matrix illustrating the performance of the proposed model on the Penobscot dataset. Each cell represents the percentage of observed pixels relative to the total number of pixels in the dataset. The diagonal represents correctly classified pixels, while the off-diagonal represents incorrectly classified pixels. The rightmost column presents precision values in green, indicating true positives, and false positives in red. The bottom row shows recall values in green, representing true positives, and false negatives in red.

A-1 (a) A single seismic profile, followed by the attention weight outputs of each layer, corresponding to different levels of the encoder in the U-Net architecture, including (b) Layer 1, (c) Layer 2, (d) Layer 3, and (e) Layer 4.
Figure 1. Standard U-Net architecture. The contracting path on the left is the encoder and the expanding path on the right is the decoder. Arrows show direction of information flow. Coloured blocks represent particular layers as per the legend. The numbers represent the output channels. "I" represents the spatial size of the image, downscaled by the pooling operators (red) and upscaled by transposed convolution operators (blue).

264x121mm (300 x 300 DPI)
Figure 2. Comparison of building blocks with an input $x$ passing through convolution layers $\omega_i$ and mapped to $F(x)$ in (a) a feed forward convolutional block and (b) a residual block with an identity shortcut connection.

322x158mm (197 x 197 DPI)
Figure 3. The attention block illustrates the computation of the Query, Key, and Value matrices. The input matrix $X$ is multiplied by the weight matrices, followed by scaling and applying the softmax function to produce the attention. The attention is then multiplied with the value matrix to generate the output of the self-attention layer.

$$[Z] = \text{Softmax}\left(\frac{QK^T}{\text{Scale}}\right)V$$
Figure 4. Proposed U-Net architecture with residual and attention blocks. Arrows show direction of information flow. Coloured blocks represent particular layers as per the legend. The numbers represent the output channels. "I" represents the spatial size of the image, downscaled by strided convolutions (pink, orange) and one max-pooling operation (red) and upscaled by transposed convolution operators (blue). The black dotted line are skip connections and the blue dotted arrow are shortcut connections for identity mapping.

399x180mm (300 x 300 DPI)
Figure 5. (a) F3 seismic cube and the (b) annotated 3D cube showing the 6 class labels.

99x35mm (300 x 300 DPI)
Figure 6. F3 cube divided into the training set (red shaded area), the validation set (yellow shaded area) and the testset (green shaded area).

50x39mm (300 x 300 DPI)
Figure 7. (a) Penobscot seismic cube and the (b) annotated 3D cube showing the 8 class labels.

80x61mm (300 x 300 DPI)
Figure 8. Penobscot seismic cube divided into the training set (red shaded area), the validation set (yellow shaded area) and the test set (green shaded area).

52x54mm (300 x 300 DPI)
Figure 9. Example of augmentations applied to (a) seismic patches and (b) segmentation label patches.
Figure 10. Power spectrum for the F3 seismic cube.

308x225mm (300 x 300 DPI)
Figure 11. Power spectrum for the Penobscot seismic cube.

308x225mm (300 x 300 DPI)
Figure 12. Summary of the workflow showing the input patch of size $128 \times 128 \times 5$ fed to the U-Net. The patch consists of five layers composed of seismic, depth and three frequency profiles.
Figure 13. (a) Seismic profiles, (b) the ground truth labels, and (c) the predicted results using the proposed U-Net.

66x47mm (300 x 300 DPI)
Figure 14. Confusion matrix illustrating the performance of the proposed model on the F3 dataset. Each cell represents the percentage of observed pixels relative to the total number of pixels in the dataset. The diagonal represents correctly classified pixels, while the off-diagonal represents incorrectly classified pixels. The rightmost column presents precision values in green, indicating true positives, and false positives in red. The bottom row shows recall values in green, representing true positives, and false negatives in red.

<table>
<thead>
<tr>
<th>Predicted class</th>
<th>True class</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>100%</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1</td>
<td>18.97%</td>
<td>0.17%</td>
<td>0.01%</td>
<td>0.00%</td>
<td>0.00%</td>
<td>99.05%</td>
<td>0.95%</td>
</tr>
<tr>
<td>2</td>
<td>2</td>
<td>0.55%</td>
<td>9.09%</td>
<td>0.77%</td>
<td>0.04%</td>
<td>0.01%</td>
<td>86.95%</td>
<td>13.05%</td>
</tr>
<tr>
<td>3</td>
<td>3</td>
<td>0.00%</td>
<td>0.43%</td>
<td>43.91%</td>
<td>0.16%</td>
<td>0.40%</td>
<td>0.01%</td>
<td>97.77%</td>
</tr>
<tr>
<td>4</td>
<td>4</td>
<td>0.00%</td>
<td>0.00%</td>
<td>0.27%</td>
<td>6.15%</td>
<td>0.70%</td>
<td>0.02%</td>
<td>86.04%</td>
</tr>
<tr>
<td>5</td>
<td>5</td>
<td>0.00%</td>
<td>0.00%</td>
<td>0.01%</td>
<td>0.51%</td>
<td>15.60%</td>
<td>0.27%</td>
<td>95.16%</td>
</tr>
<tr>
<td>6</td>
<td>6</td>
<td>0.00%</td>
<td>0.00%</td>
<td>0.04%</td>
<td>0.08%</td>
<td>0.30%</td>
<td>1.54%</td>
<td>78.78%</td>
</tr>
</tbody>
</table>

97.20% | 93.75% | 97.55% | 88.76% | 91.67% | 83.66% | 95.25% | 2.80% | 6.25% | 2.45% | 11.24% | 8.33% | 16.34% | 4.75%

594x595mm (197 x 197 DPI)
Figure 15. (a) Seismic profiles and (b) the ground truth labels. A comparison of the predicted labels using (c) no additional frequency channels to the input and (d) the proposed U-Net with frequency channels added to the input.

149x102mm (300 x 300 DPI)
Figure 16. (a) Seismic profiles and (b) the ground truth labels. A comparison of the predicted labels using (c) a standard U-Net, (d) a U-Net with a residual blocks, and (e) the proposed U-Net.

99x47mm (300 x 300 DPI)
Figure 17. (a) A single seismic profile, (b) ground truth labels for the single seismic profile, (c) segmentation results reproduced after Alaudah et al. (2019), (d) segmentation results reproduced after Chen et al. (2022), and (e) segmentation results achieved by our proposed method.

99x161mm (300 x 300 DPI)
Figure 18. The first column shows an array of seismic inlines extending throughout the entire Penobscot dataset. The second column shows the ground truth labels and the third column shows the predicted labels.

59x177mm (300 x 300 DPI)
Figure 19. Confusion matrix illustrating the performance of the proposed model on the Penobscot dataset. Each cell represents the percentage of observed pixels relative to the total number of pixels in the dataset. The diagonal represents correctly classified pixels, while the off-diagonal represents incorrectly classified pixels. The rightmost column presents precision values in green, indicating true positives, and false positives in red. The bottom row shows recall values in green, representing true positives, and false negatives in red.

<table>
<thead>
<tr>
<th>Predicted class</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
<th>8</th>
</tr>
</thead>
<tbody>
<tr>
<td>True class</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1</td>
<td>54.77%</td>
<td>0.05%</td>
<td>0.00%</td>
<td>0.00%</td>
<td>0.00%</td>
<td>0.00%</td>
<td>0.00%</td>
<td>0.00%</td>
</tr>
<tr>
<td>2</td>
<td>0.76%</td>
<td>8.44%</td>
<td>0.08%</td>
<td>0.00%</td>
<td>0.00%</td>
<td>0.00%</td>
<td>0.00%</td>
<td>0.00%</td>
</tr>
<tr>
<td>3</td>
<td>0.01%</td>
<td>0.38%</td>
<td>15.62%</td>
<td>0.05%</td>
<td>0.02%</td>
<td>0.03%</td>
<td>0.01%</td>
<td>0.00%</td>
</tr>
<tr>
<td>4</td>
<td>0.00%</td>
<td>0.00%</td>
<td>0.33%</td>
<td>1.56%</td>
<td>0.19%</td>
<td>0.07%</td>
<td>0.01%</td>
<td>0.00%</td>
</tr>
<tr>
<td>5</td>
<td>0.00%</td>
<td>0.00%</td>
<td>0.00%</td>
<td>0.08%</td>
<td>1.13%</td>
<td>0.10%</td>
<td>0.00%</td>
<td>0.00%</td>
</tr>
<tr>
<td>6</td>
<td>0.00%</td>
<td>0.00%</td>
<td>0.00%</td>
<td>0.00%</td>
<td>0.13%</td>
<td>6.58%</td>
<td>0.23%</td>
<td>0.00%</td>
</tr>
<tr>
<td>7</td>
<td>0.00%</td>
<td>0.00%</td>
<td>0.00%</td>
<td>0.00%</td>
<td>0.00%</td>
<td>0.08%</td>
<td>5.36%</td>
<td>0.13%</td>
</tr>
<tr>
<td>8</td>
<td>0.00%</td>
<td>0.00%</td>
<td>0.00%</td>
<td>0.00%</td>
<td>0.00%</td>
<td>0.00%</td>
<td>0.11%</td>
<td>3.73%</td>
</tr>
</tbody>
</table>

|         | 98.61%  | 95.22%  | 97.42%  | 92.31%  | 77.32%  | 95.91%  | 93.77%  | 96.61%  |
|         | 1.39%   | 4.78%   | 2.58%   | 7.69%   | 22.68%  | 4.09%   | 6.23%   | 3.39%   |

750x755mm (197 x 197 DPI)
Figure A-1. (a) A single seismic profile, followed by the attention weight outputs of each layer, corresponding to different levels of the encoder in the U-Net architecture, including (b) Layer 1, (c) Layer 2, (d) Layer 3, and (e) Layer 4.
<table>
<thead>
<tr>
<th>Layer name</th>
<th>Output size</th>
<th>Layer</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>conv1</td>
<td>112 × 112</td>
<td>7 × 7, 64, stride 2</td>
<td></td>
</tr>
<tr>
<td>conv2, x</td>
<td>56 × 56</td>
<td>3 × 3 maxpool, stride 2</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>[3 × 3, 64] ×2</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>[3 × 3, 64]</td>
<td></td>
</tr>
<tr>
<td>conv3, x</td>
<td>28 × 28</td>
<td>3 × 3, 128 ×2</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>[3 × 3, 128]</td>
<td></td>
</tr>
<tr>
<td>conv4, x</td>
<td>14 × 14</td>
<td>3 × 3, 256 ×2</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>[3 × 3, 256]</td>
<td></td>
</tr>
<tr>
<td>conv5, x</td>
<td>7 × 7</td>
<td>3 × 3, 512 ×2</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>[3 × 3, 512]</td>
<td></td>
</tr>
</tbody>
</table>

Table 1. Encoder building blocks for the proposed architecture.
<table>
<thead>
<tr>
<th>Class 1</th>
<th>Class 2</th>
<th>Class 3</th>
<th>Class 4</th>
<th>Class 5</th>
<th>Class 6</th>
</tr>
</thead>
<tbody>
<tr>
<td>(Upper N. S.)</td>
<td>(Middle N. S.)</td>
<td>(Lower N. S.)</td>
<td>(Rijnland/Chalk)</td>
<td>(Scruff)</td>
<td>(Zechstein)</td>
</tr>
<tr>
<td>28.09%</td>
<td>11.89%</td>
<td>48.59%</td>
<td>6.64%</td>
<td>3.28%</td>
<td>1.51%</td>
</tr>
</tbody>
</table>

Table 2. Percentage of class labels in the F3 dataset.
<table>
<thead>
<tr>
<th>Model</th>
<th>Class 1</th>
<th>Class 2</th>
<th>Class 3</th>
<th>Class 4</th>
<th>Class 5</th>
<th>Class 6</th>
<th>MCA</th>
</tr>
</thead>
<tbody>
<tr>
<td>Standard U-Net</td>
<td>94.7%</td>
<td>93.2%</td>
<td>97.3%</td>
<td>77.3%</td>
<td>66.0%</td>
<td>76.97%</td>
<td>84.3%</td>
</tr>
<tr>
<td>Residual block U-Net</td>
<td>97.2%</td>
<td>93.2%</td>
<td><strong>98.0%</strong></td>
<td>79.0%</td>
<td>87.8%</td>
<td>81.1%</td>
<td>89.4%</td>
</tr>
<tr>
<td>Proposed model</td>
<td><strong>97.2%</strong></td>
<td><strong>93.7%</strong></td>
<td>97.6%</td>
<td><strong>88.7%</strong></td>
<td><strong>91.7%</strong></td>
<td><strong>83.9%</strong></td>
<td><strong>92.0%</strong></td>
</tr>
</tbody>
</table>

Table 3. Accuracy comparison of three U-Net architectures for the F3 dataset.
<table>
<thead>
<tr>
<th>Model</th>
<th>Pixel accuracy</th>
<th>IOU</th>
<th>Class 1</th>
<th>Class 2</th>
<th>Class 3</th>
<th>Class 4</th>
<th>Class 5</th>
<th>Class 6</th>
<th>MCA</th>
<th>FWIU</th>
</tr>
</thead>
<tbody>
<tr>
<td>(Alaudah et al., 2019)</td>
<td>90.5%</td>
<td>-</td>
<td>97.4%</td>
<td>93.8%</td>
<td>94.1%</td>
<td>77.2%</td>
<td>67.4%</td>
<td>60.2%</td>
<td>81.7%</td>
<td>83.2%</td>
</tr>
<tr>
<td>section</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>(Alaudah et al., 2019)</td>
<td>86.2%</td>
<td>-</td>
<td>92.6%</td>
<td>91.2%</td>
<td>97.4%</td>
<td>67.3%</td>
<td>28.6%</td>
<td>45.8%</td>
<td>70.5%</td>
<td>75.7%</td>
</tr>
<tr>
<td>patch</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>(Chen et al., 2022)</td>
<td>93.4%</td>
<td>77.5%</td>
<td>97.9%</td>
<td>93.9%</td>
<td>97.1%</td>
<td>84.5%</td>
<td>73.4%</td>
<td>72.6%</td>
<td>86.6%</td>
<td>88.2%</td>
</tr>
<tr>
<td>Proposed model</td>
<td>95.2%</td>
<td>84.6%</td>
<td>97.2%</td>
<td>93.7%</td>
<td>97.6%</td>
<td>88.7%</td>
<td>91.7%</td>
<td>83.9%</td>
<td>92.0%</td>
<td>89.0%</td>
</tr>
</tbody>
</table>

Table 4. Comparison of accuracy metrics between the proposed model and the state-of-the-art benchmarks for the F3 dataset.
Table 5. Percentage of class labels in the Penobscot dataset.

<table>
<thead>
<tr>
<th>Class 1</th>
<th>Class 2</th>
<th>Class 3</th>
<th>Class 4</th>
<th>Class 5</th>
<th>Class 6</th>
<th>Class 7</th>
<th>Class 8</th>
</tr>
</thead>
<tbody>
<tr>
<td>57.37%</td>
<td>8.48%</td>
<td>15.46%</td>
<td>1.63%</td>
<td>1.30%</td>
<td>6.67%</td>
<td>5.40%</td>
<td>3.71%</td>
</tr>
</tbody>
</table>
DATA AND MATERIALS AVAILABILITY

Data associated with this research are available and can be obtained by contacting the corresponding author.