Pontifícia Universidade Católica
do Rio de Janeiro

Luis Fernando Marin Sepulveda

Generalization of the Deep Learning Model for Natural Gas Indication in 2D Seismic Image Based on the Training Dataset and the Operational Hyper Parameters Recommendation

Tese de Doutorado

Thesis presented to the Programa de Pós–graduação em Informática of PUC-Rio in partial fulfillment of the requirements for the degree of Doutor em Informática.

Advisor : Prof. Marcelo Gattass
Co-advisor: Prof. Aristofanes Corrêa Silva

Rio de Janeiro
January 2024

Pontifícia Universidade Católica
do Rio de Janeiro

Luis Fernando Marin Sepulveda

Generalization of the Deep Learning Model for Natural Gas Indication in 2D Seismic Image Based on the Training Dataset and the Operational Hyper Parameters Recommendation

Thesis presented to the Programa de Pós–graduação em Informática of PUC-Rio in partial fulfillment of the requirements for the degree of Doutor em Informática. Approved by the Examination Committee:

Prof. Marcelo Gattass
Advisor
Departamento de Informática – PUC-Rio

Prof. Aristofanes Corrêa Silva
Co-advisor
Universidade Federal do Maranhão -UFMA

Prof. Raul Queiroz Feitosa
Pontifícia Universidade Católica do Rio de Janeiro – PUC-Rio

Dr. Jan José Hurtado Jauregui
Pontifícia Universidade Católica do Rio de Janeiro, Instituto Tecgraf – PUC-Rio

Prof. António Manuel Trigueiros da Silva Cunha
Universidade de Trás-os-Montes e Alto Douro - UTAD

Prof. Kelson Rômulo Teixeira Aires
Universidade Federal do Piauí - UFPI

Rio de Janeiro, January 24th, 2024

All rights reserved.

Luis Fernando Marin Sepulveda

Graduated in Systems Engineering from the University of Antioquia (Colombia: Medellín). He completed a master's degree in the Computer Science Department of the Federal University of Maranhão-UFMA, with specialization in the area of Computer Science.

Bibliographic Data

Marin Sepulveda, Luis Fernando

Generalization of the Deep Learning Model for Natural Gas Indication in 2D Seismic Image Based on the Training Dataset and the Operational Hyper Parameters Recommendation / Luis Fernando Marin Sepulveda; advisor: Marcelo Gattass; co-advisor: Aristofanes Corrêa Silva. – 2024.

158 f: il. color. ; 30 cm

Tese (doutorado) - Pontifícia Universidade Católica do Rio de Janeiro, Departamento de Informática, 2024.

Inclui bibliografia

1. Aprendizado profundo. 2. Generalização. 3. Recomendação de conjunto de dados de treinamento. 4. Imagem sísmica 2D em terra. 5. Agrupamento. 6. Indicação de gás . I. Gattass, Marcelo. II. Correa Silva, Aristofanes. III. Pontifícia Universidade Católica do Rio de Janeiro. Departamento de Informática. IV. Título.

CDD: 004

To my beloved wife, parents, brother and especially my grandmother.

Acknowledgments

This study was financed in part by the Coordenação de Aperfeiçoamento de Pessoal de Nível Superior - Brasil (CAPES) - Finance Code 001. I appreciate the special support of the Tecgraf Institute (PUC-RIO) and the Applied Computing Group of the Federal University of Maranhão. To CNPq (Conselho Nacional de Desenvolvimento Científico e Tecnológico) and PUC-RIO for financial support and scholarship, which enabled full dedication to the post-graduate program.

I especially thank my advisors whose example, guidance and constant support have made me a better professional and human. Likewise, I thank my colleagues, both from Rio de Janeiro and São Luís, since their constant help has allowed me to advance. Finally, I thank my dear family, who at all times in my life have supported me and encouraged me to overcome any obstacle.

To all of you my most sincere thanks.

Abstract

Marin Sepulveda, Luis Fernando; Gattass, Marcelo (Advisor); Correa Silva, Aristofanes (Co-Advisor). **Generalization of the Deep Learning Model for Natural Gas Indication in 2D Seismic Image Based on the Training Dataset and the Operational Hyper Parameters Recommendation**. Rio de Janeiro, 2024. 158p. Tese de Doutorado – Departamento de Informática, Pontifícia Universidade Católica do Rio de Janeiro.

Interpreting seismic images is an essential task in diverse fields of geosciences, and it's a widely used method in hydrocarbon exploration. However, its interpretation requires a significant investment of resources, and obtaining a satisfactory result is not always possible.

The literature shows an increasing number of Deep Learning, DL, methods to detect horizons, faults, and potential hydrocarbon reservoirs, nevertheless, the models to detect gas reservoirs present generalization performance difficulties, i.e., performance is compromised when used in seismic images from new exploration campaigns. This problem is especially true for 2D land surveys where the acquisition process varies, and the images are very noisy.

This work presents three methods to improve the generalization performance of DL models of natural gas indication in 2D seismic images, for this task, approaches that come from Machine Learning, ML, and DL are used. The research focuses on data analysis to recognize patterns within the seismic images to enable the selection of training sets for the gas inference model based on patterns in the target images. This approach allows a better generalization of performance without altering the architecture of the gas inference DL model or transforming the original seismic traces.

The experiments were carried out using the database of different exploitation fields located in the Parnaíba basin, in northeastern Brazil. The results show an increase of up to 39% in the correct indication of natural gas according to the recall metric. This improvement varies in each field and depends on the proposed method used and the existence of representative patterns within the training set of seismic images.

These results conclude with an improvement in the generalization performance of the DL gas inference model that varies up to 21% according to the F1 score and up to 15% according to the IoU metric. These results demonstrate that it is possible to find patterns within the seismic images using an unsupervised approach, and these can be used to recommend the DL training set

according to the pattern in the target seismic image; Furthermore, it demonstrates that the training set directly affects the generalization performance of the DL model for seismic images.

Keywords

Deep Learning; Generalizability; Training Dataset recommendation; 2D Seismic onshore image; Clustering; Gas indication.

Resumo

Marin Sepulveda, Luis Fernando; Gattass, Marcelo; Correa Silva, Aristofanes. **Generalização do modelo de aprendizado profundo para indicação de gás natural em dados sísmicos 2D com base no conjunto de dados de treinamento e recomendação de hiperparâmetros operacionais**. Rio de Janeiro, 2024. 158p. Tese de Doutorado – Departamento de Informática, Pontifícia Universidade Católica do Rio de Janeiro.

A interpretação de imagens sísmicas é uma tarefa essencial em diversas áreas das geociências, sendo um método amplamente utilizado na exploração de hidrocarbonetos. Porém, sua interpretação exige um investimento significativo de recursos, e nem sempre é possível obter um resultado satisfatório.

A literatura mostra um número crescente de métodos de Deep Learning, DL, para detecção de horizontes, falhas e potenciais reservatórios de hidrocarbonetos, porém, os modelos para detecção de reservatórios de gás apresentam dificuldades de desempenho de generalização, ou seja, o desempenho fica comprometido quando utilizados em imagens sísmicas de novas explorações campanhas. Este problema é especialmente verdadeiro para levantamentos terrestres 2D, onde o processo de aquisição varia e as imagens apresentam muito ruído.

Este trabalho apresenta três métodos para melhorar o desempenho de generalização de modelos DL de indicação de gás natural em imagens sísmicas 2D, para esta tarefa são utilizadas abordagens provenientes de Machine Learning, ML e DL. A pesquisa concentra-se na análise de dados para reconhecer padrões nas imagens sísmicas para permitir a seleção de conjuntos de treinamento para o modelo de inferência de gás com base em padrões nas imagens alvo. Esta abordagem permite uma melhor generalização do desempenho sem alterar a arquitetura do modelo DL de inferência de gás ou transformar os traços sísmicos originais.

Os experimentos foram realizados utilizando o banco de dados de diferentes campos de exploração localizados na bacia do Parnaíba, no Nordeste do Brasil. Os resultados mostram um aumento de até 39% na indicação correta do gás natural de acordo com a métrica de recall. Esta melhoria varia em cada campo e depende do método proposto utilizado e da existência de padrões representativos dentro do conjunto de treinamento de imagens sísmicas.

Estes resultados concluem com uma melhoria no desempenho de generalização do modelo de inferência de gases DL que varia até 21% de acordo com a pontuação F1 e até 15% de acordo com a métrica IoU. Estes resultados demonstram que é possível encontrar padrões dentro das imagens sísmicas

usando uma abordagem não supervisionada, e estas podem ser usadas para recomendar o conjunto de treinamento DL de acordo com o padrão na imagem sísmica alvo; Além disso, demonstra que o conjunto de treinamento afeta diretamente o desempenho de generalização do modelo DL para imagens sísmicas.

Palavras-chave

Aprendizado profundo; Generalização; Recomendação de conjunto de dados de treinamento; Imagem sísmica 2D em terra; Agrupamento; Indicação de gás.

Table of Contents

List of Figures

List of Tables

List of Abbreviations

AvTD – Average Taxonomic Distinction

DA – Domain Adaptation

DFT – Discrete Fourier Transform

DL – Deep Learning

D_S – Source Domain

D_T – Target Domain

E – Extensive Quadratic Entropy

GRU – Gated Recurrent Unit

HDBSCAN – Hierarchical Density-based Spatial Clustering of Applications with Noise

I – Intensive Quadratic Entropy

IoU – Intersection Over Union

ML – Machine Learning

LBP – Local Binary Pattern

LSTM – Long Short Term Memory network

MAE – Mean Absolute Error

MNND – Mean Nearest Neighbor Distance

MPD – Mean Phylogenetic Distance

MSE – Mean Squared Error

PCA – Principal Component Analysis

PDI – Pure Diversity Index

PSO – Particle Swarm Optimization

ROI – Region of Interest,

SEG – Society of Exploration Geophysicists

TL – Transfer Learning

T_S – Source Task

T_T – Target Task

TTD – Total Taxonomic Distinction

1
Introduction

1.1
Background

Seismic image analysis is widely used in hydrocarbon exploration and applied in marine, terrestrial, and transition zone environments. However, its interpretation requires a significant amount of time and the result depends on the experience of the professionals in charge of the analysis (Azzam et al., 2018; Lou et al., 2022; Sarhan and Safa, 2019; Trani et al., 2022).

A technique used for seismic surveying is based on an energy impulse generation that propagates through the Earth's subsoil. The wave generated is reflected by the different layers of rock, being captured by several devices known as receivers that record the wave amplitude with respect to the arrival time. After the acquisition campaign, seismic processing transforms the data captured by the receivers into a vertical trace that represents the internal structures of a vertical section of land. By concatenating several consecutive equally spaced traces, a 2D seismic image is produced (Alsadi, 2017).

Methods based on Machine Learning, ML, have recently been developed for the analysis of seismic data to accomplish specific tasks ranging from data interpretation to detection of specific anomalies or events, these methods represent help for professionals, who otherwise require large amounts of time to perform the analysis, and whose results often present discrepancies when compared to those of other professionals (Bai and Tahmasebi, 2021; Dell et al., 2020; Sarhan and Safa, 2019; Trani et al., 2022; Zhang et al., 2022b; Zhao et al., 2022).

Specifically, methods based on Deep learning models, DL, have been used for the analysis of seismic images in order to obtain interpretations that help in the exploration of hydrocarbon deposits, these methods are applied to different environments that have specifics and particular features. Within these models are those developed for the indication of natural gas reservoirs in 2D seismic images (Alfarhan et al., 2020b,a; Andrade et al., 2021; Fernando Santos et al., 2020; Pan et al., 2021a; Wang et al., 2018).

1.2 Motivation

This work is motivated by the need to use DL-based models created for gas inference in 2D seismic images from different exploration fields.

Normally, when creating a DL-based model for gas inference, a set of seismic images, which come from a single exploitation field, and have labels indicating the position of the gas reservoir within each seismic image, are used as training. The DL model is then tested and adjusted until acceptable performance on the gas inference task is achieved, tested on images that come from the same exploitation field, but that were not used in the training process.

The process of creating the DL model described can be carried out using more than one exploitation field, but there is a need to use the resulting model in seismic images that come from new exploration fields, in which there are no marking labels to carry out a retraining process. However, when using the model trained on new seismic images, there are performance losses that indicate that the new images have features that the DL model is not able to recognize.

In this situation and depending on the number of fields with training sets available, it is possible to create multiple DL models based on the same or even different architectures, but a problem arises in this situation, and that is how to determine which model can recognize the features of the new target seismic images.

Different alternatives arise, such as using the models that were trained using the closest seismic images in terms of terrain distance to the new seismic images. Another option is to carry out new training using only the closest seismic images, regardless of whether they belong to different exploitation fields. These options are based on the hypothesis that seismic from nearby areas have similar features. However, this may not be true in all cases, since there may be changes in the composition of the terrain (Mustafa and AlRegib, 2021; Rollmann et al., 2022; Zhang et al., 2022a).

There is also the possibility of using professional services to select seismic images that are most similar to the new target seismic image. Although this option requires time and makes the method depend on the experience of the professional in charge.

These approaches demonstrate that there is a need to identify the features within the new seismic images that allow the identification of the gas reservoir using DL-based models, in other words, it is necessary to identify a method that allows adapting the creation of DL-based models according to the features of seismic images that come from new explorations fields.

1.3 Problem Statement

When using DL-based models on seismic images, it is observed that when the training data and the target data are collected by the same company and come from the same specific geological region, the results obtain satisfactory performance. However, there is a problem with the performance when these models are used on new data that comes from regions with different geological conditions or is collected by different teams.

This problem is not exclusive to models developed to work with seismic images, in general, it refers to model performance losses caused by differences between the training data and the new data, also known as a model generalization performance (Chang et al., 2021; Wang et al., 2022).

There are several alternatives to face this problem. One option is to use professional expert services to label the new seismic images allowing re-training of the models under new features. However, this option is costly in time and human resources (Huang et al., 2019; Li et al., 2019; Sudharshan et al., 2019; Yu et al., 2017).

A second option is to use Domain Adaptation, DA, a field associated with ML and has different approaches. Some of these modify the models to obtain the ability to extract features common to training and the new data, which allows fulfilling the aiming task (Jin et al., 2021; Li et al., 2021; Sanodiya et al., 2021). A third option is Transfer Learning, TL, which uses models developed to solve a different but related task, i.e., it seeks to transfer the knowledge acquired by a model trained with a different type of data than the target data, but whose task is the same. Transfer learning takes advantage of original model layers by freezing some activation map values and then retraining the model with new data (Duong et al., 2021; Hermessi et al., 2019; Soudani and Barhoumi, 2019). A fourth option considers the situation where there is more than one model available, each created with different data. Here each model will have a different performance depending on the features of the data to be processed. This approach aims to identify the most appropriate model to use with the target data (Rollmann et al., 2021).

In this work, a DA-based approach is used, focused on the analysis of the training data and how the representativeness of different seismic features affects the generalization performance of gas inference models in 2D seismic images. This approach is chosen because there is a large amount of seismic images available to train gas inference models, yet performance with new target data decreases, indicating the existence of multiple patterns that are not correctly recognized by the DL models.

1.4 Research Aim

This work aims to develop three methods to improve the generalization performance of gas reserve inference DL methods on 2D seismic images, compared to the performance of the reference method trained using all available training data.

1.4.1 Research Questions

Main Question:

How to alter DL-based methods for gas inference in 2D seismic images to adapt them to specific patterns of new seismic images allowing better generalization performance?

It is clarified that DL-based methods refer to a complete method that uses a DL model and not just a DL architecture.

Subsequent Questions:

First, compared to the results obtained with the default DL model on seismic images, how can the available data be used to allow better generalization performance of the DL model? This question addresses the use and manipulation of the available data, without modifying the existing DL model, that is, it seeks to improve the model's performance on new data but without altering it.

Second, how to identify patterns within seismic images to allow a comparison that establishes similarities or domains, which improve the indication of natural gas reserves? This question refers to how to extract features that allow different seismic images to be compared, and that these features are also relevant for identifying natural gas.

1.5 Methodology

To carry out this work, three methods are developed, in which techniques based on ML and DL are explored, observing their impact on the generalization performance of the gas reserve inference model.

The research is based on the following work sequence that results in the three proposed methods:

- The components involved in the creation of a DL-based natural gas inference method applied to 2D seismic images are identified and analyzed.
- Experiments are performed to determine which components have the greatest influence on the generalization performance of the models.

- A first proposed method is created that modifies a base gas inference model. The effects on generalization performance are evaluated and the limitations of the proposed method are identified.
- Two new proposed methods are created, each seeks to overcome the limitations found in the previous method and achieve better generalization performance.
- For each of the three proposed methods, experiments are carried out, and their implication within the research objective is analyzed, as well as their advantages and disadvantages.
- Finally, a comparative analysis is carried out between the proposed methods, showing properties and limitations.

1.6 Contributions

There are different contributions to the state of the art, made for the three methods proposed in this work:

- Establishment of a basis for the comparison of seismic image features.

  This work presents evidence that there are patterns within seismic images that can be identified through feature extraction and that can also be used to compare and establish similarities between different seismic images. These patterns can be considered domains. Furthermore, this work presents three feature extraction methods focused on recognizing these patterns.
- Three methods are introduced to compare features extracted from seismic images that allow the creation of clusters containing a set of seismic images with similar patterns.

  Three different ways are established to compare seismic features and create clusters based on their similarity.
- Three methods are presented that allow recommending the training dataset for a gas inference DL model, based on similarity with the target seismic images.

  This work presents evidence that it is possible to select seismic images from the training database that exhibit similar features to the target seismic images and that the recommended set also affects generalization performance.

- An automatic selection of operational hyper parameters is established for the DL gas inference model.

  The conducted experiment demonstrates that a tuning process needs to be performed to identify the appropriate DL operational hyper parameters. This means that it is necessary to modify the hyper parameter of the DL model to allow recognizing the features that represent the different domains.

- Three methods are presented that enable better generalization performance for gas inference DL methods.

  The result showed that for all three proposed methods, there is an improvement in generalization performance compared to the traditional approach that used all available data as training for the DL model. This demonstrates that within the context of seismic imaging, it is important to identify the specific domain of the target images to select the appropriate training data.

1.7 Document Organization

This thesis is structured as follows. Chapter 2 presents the main techniques used for the construction of the proposed method. Chapter 3 shows the analysis carried out in the search for the state of the art. Chapter 4 shows the first proposed generalization method. Chapter 5 shows the second proposed generalization method. Chapter 6 shows the third proposed generalization method. Chapter 7 presents the methods comparison. Chapter 8 presents the conclusion, contribution, future work and scientific productions.

2
Theoretical Foundation

This chapter shows the theoretical foundations used by the three proposed methods, explaining the type of data, presenting several techniques that are used for the extraction, grouping, and classification of features, as well as the relevance analysis methods used on the extracted features.

2.1
Seismic Image Data

Seismic Image is a spatio-temporal sampling of the backscattered seismic wavefield that is collected by seismic surveys or seismic imaging techniques. This data is an ordered collection of traces. It can be considered as a 2/3-D matrix, which is used to help scientists and geophysicists understand the interior structure of the Earth and is commonly used in geophysics and the hydrocarbon industry (Gupta, 2021).

There are different ways to create seismic images, one of them, seen from an elevated perspective, uses waves sent through the earth to create a seismic trace, which is a logarithmic measure of disturbances (particle velocity/acoustic pressure), Vel.log, of waves reflected from the subsurface over time. Traces record in a waveform the intrinsic attributes of amplitude, phase, frequency, polarity, arrival time, and velocity of a reflection signal.

This technique uses a source wave generator and several receivers that collect waves reflected by underground layers. The attributes of the waves are indicators of the rock's composition and the depth at which they are found.

By integrating the data it is possible to obtain the reflection coefficient, Rc, that is, the amplitude of the reflected wave with respect to the incident wave (initial seismic wave), constituting a single vector called seismic trace. By joining different traces, a 2D matrix is created, which is a seismic image called 2D seismic images (Alsadi, 2017; Nanda, 2016).

Figure 2.1 presents a simplified process for obtaining seismic trace data, showing its collection and processing to obtain a seismic trace and Figure 2.2 shows an example of creating a 2D seismic image, starting from a seismic trace and its interpretation as an image, until showing the result of concatenating several consecutive equally spaced traces.

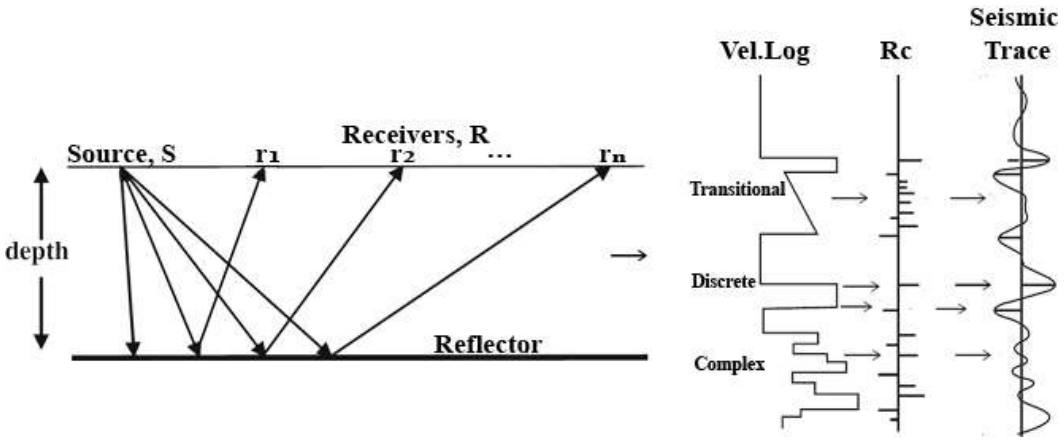


Figure 2.1: Simplified seismic images acquisition process. Adapted from Alsadi (2017) and Nanda (2016).

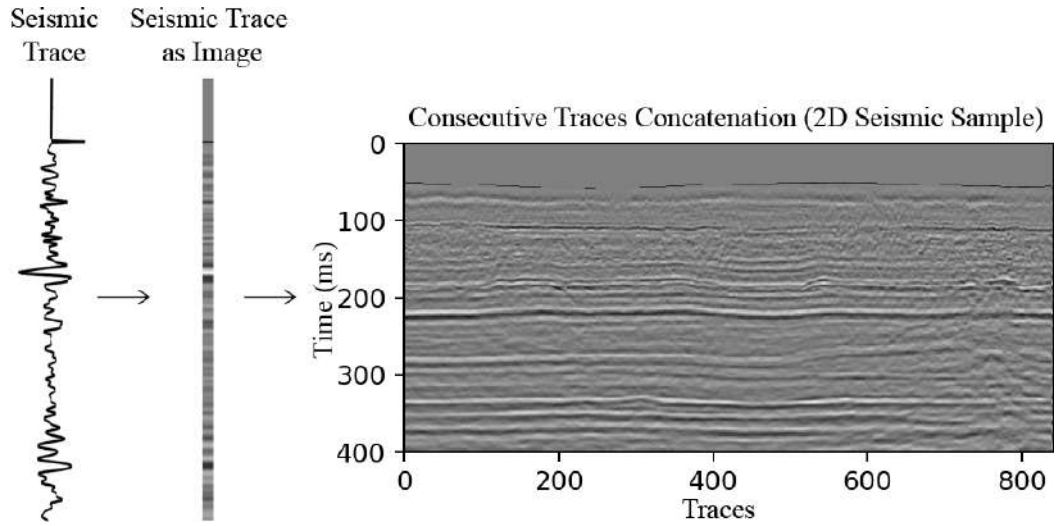


Figure 2.2: Example of trace and 2D seismic image.

2.2 Generalization Performance

This section introduces the concept of Generalization and also presents several learning techniques that are used to alter the generalization of learning models.

Generalization performance is related to the learning model's performance when used with out-of-sample data (Sammut and Webb, 2017a).

Section 1.3 introduces the problem that is studied in the presented work, this is related to the generalization of performance, but this problem is related to the assumption that any pattern recognition technique has, which is that the new data used for inference has the same distribution as the training set (Ghosh et al., 2020).

An example of this situation within the context of gas reservoir identification is that a trained model can only be used on data that has been collected

using the same equipment and parameterization used to collect the training set, it may even be necessary that the terrain have a similar composition.

Within ML there is a subarea focused on emulating the human ability to adapt pre-existing concepts to new environments by transferring the knowledge acquired in previous tasks to a new one or adapting to new data that has similar but not the same features.

This subarea is known as Transfer Learning and has different approaches, some focus on the data, and others on the learning model (Venkateswara and Panchanathan, 2020).

2.2.1 Transfer Learning

Transfer learning is a branch of machine learning and is defined as:

Given a source domain, D_S and a source leaning task T_S, a target domain D_T and a target learning task T_T, transfer learning aims to improve the target predictive function $f_T(.)$ using D_S and T_S, where $D_S \neq D_T$, or $T_S \neq T_T$ (Pan and Yang, 2010).

To put the above definition in a seismic context, we can consider the situation where D_S refers to data that was collected using a specific equipment model, and D_T refers to a new set of data that is collected using a different model. This difference can produce data that vary in resolution, size, or other features, even if both equipment produce the same kind of data.

Another situation may be that the data from both domains were collected by the same equipment but T_S indicates the location of natural gas reserves, and the new T_T was intended to indicate the location of the oil reserve.

In both situations, the transfer learning aim is to improve performance in the target domain but using data available from the source domain.

Several types of transfer learning can be applied depending on the available data or the specific task in which an improvement in model performance is sought.

To understand the different types of transfer learning, the main types and their characteristics are introduced, contextualizing them within the seismic area:

2.2.1.1 Multitask Learning

In Multitask Learning type, multiple target tasks exist, but the domains are considered the same or related. It seeks to create an algorithm that can be applied to K tasks simultaneously, using datasets that come from all domains,

to improve generalization.

A formal definition considers K different tasks, this means that $T = \{T_1, T_2, \cdots, T_K\}$, where the data for each task are sampled from K different domains $D = \{D_1, D_2, \cdots, D_k\}$ respectively. In this case, each domain has labeled data, which means a supervised learning process where the dataset is presented as n tuples, $X_l = \{x_i, y_i\}_{i=1}^{n}$, where $x_i \in X$ and X represent the feature space of the data, and where $y_i \in Y$ and Y represent the set labels. In multitask learning there is also the restriction that it may not be possible to estimate a reliable empirical probability distribution $\hat{P}_k(X, Y)$ for k^{th} domain using only data from the k^{th} domain $D_k = \{x_k^i, y_k^i\}_{i=1}^{n_k}$, $x_k^i \in X_k$ and $y_k^i \in Y_k$ (Caruana, 1997; Venkateswara and Panchanathan, 2020).

Within the seismic context, we can consider the source domain as the seismic images that were collected over time to perform different tasks such as indicating gas, oil, and water deposits, considering that this data already has labels that indicate the deposit location.

In this example, we have several tasks T that are unrelated to each other, but all use the same type of data (seismic) that belong to the same domain D. Multitask Learning aims to take all the datasets with labels from D and create an algorithm that can be used to find gas, oil, and water deposits in new seismic images.

2.2.1.2 Self-Taught Learning

In this type of learning the aim is to use a model that was trained in a source domain D_S different from the target D_T and without labels, which means unsupervised learning. The resulting trained model is then fine-tuned using labeled data from the target domain D_T. The general idea of Self-Taught Learning is to use a model that was trained to extract representational features from a large unlabeled dataset and then apply another training with the target data to tune the feature extraction to the target domain (Venkateswara and Panchanathan, 2020).

From the seismic perspective, Self-Taught Learning can be considered a technique that allows transferring a feature extraction model trained using data that does not even belong to the seismic domain. For example, consider a feature extraction model trained using 2D satellite imagery and then tune that model for use on 2D seismic images.

2.2.1.3 Sample Selection Bias

This is the case where the source data D_S used for model training, and even the target data D_T available for tuning, do not correctly represent the target domain or task. This lack of representation may occur simply because there is insufficient data to recognize the specific task pattern.

For a definition there exists a sample labeled dataset $D = \{x_i, y_i\}_{i=1}^{n}$. Transfer learning in Sample Selection Bias aims to determine the joint distribution $\hat{P}(X, Y)$ which is a true approximation to the joint distribution $P(X, Y)$ of the population using D. Since D is only a tiny subset of the entire population, the approximate $\hat{P}(X, Y)$ is not the same as $P(X, Y)$. This may be because the tiny subset D may lead to an incorrect estimation of the true marginal distribution $P(X)$ with $\hat{P}(X) \neq P(X)$. It could also be due to an incorrect estimate of the class prior with $\hat{P}(Y) \neq P(Y)$ which then leads to an incorrect estimate of the class conditional $\hat{P}(Y|X) \neq P(Y|X)$ (Venkateswara and Panchanathan, 2020).

In a seismic context, consider the example of the case when a model is used to find oil reserves, but the new data presents geological faults that alter the underground layer shape. This specific peculiarity of new seismic images can be considered a bias if there are few or no geologically faulted seismic images in the training dataset. In this case, the model is strongly influenced to learn to identify the pattern that allows finding oil reserves in data without geological faults.

2.2.1.4 Lifelong Machine Learning

In this type of learning the aim is to use a model that was previously trained for several tasks and train it to perform a new one, but without forgetting the previous tasks. This type of transfer learning is different from Multitask learning because new tasks are learned one at a time and not all at the same time.

In Lifelong Machine Learning, a learning model that was trained for the tasks $\{T_1, T_2, \cdots, T_K\}$ is updated to learn the task T_{K+1} with data D_{K+1}. The idea is that learning the $K+1^{th}$ task is easier since the model has already learned the $\{T_1, T_2, \cdots, T_K\}$ tasks, this approach uses knowledge accumulation (Fei et al., 2016; Venkateswara and Panchanathan, 2020).

From a seismic perspective, this type of learning can be illustrated with the example of using a model that was trained for gas indication and then is trained again with new data for the oil indication task, the expected result is

a model that can indicate both oil and gas reserves.

2.2.1.5 Zero-Shot and Few-Shot Learning

This type of transfer learning can be seen as an extreme case because these approaches attempt to learn to recognize new categories of data using a minimal number of samples. The key idea is the ability to transfer knowledge of previously learned categories to learn to recognize the boundaries that allow discriminating a new one. The advantage of Zero-Shot and Few-Shot Learning is that it allows recognition of a new category using only a few (or no) labeled examples. In this case, the model uses a blended learning strategy, which uses a tuple consisting of data and his description for training, then the trained model can use only the description of a new category to learn to recognize it (Fei-Fei et al., 2006; Goodfellow et al., 2016).

Within the seismic context, the following example can be considered: a model for rock type classification was trained using as samples a tuple of 2D images taken at high resolution and a vector containing the compositional features.

To apply Zero-Shot and Few-Shot Learning to the trained model, it must be able to perform a new training using only the description vector, to finally be used to classify a new set of 2D rock images without vector description.

2.2.1.6 Domain Adaptation

In the DA approach, knowledge transfer occurs between two or more domains, source and target. The source domains D_S are different from the target domain D_T, but the aim of DA is to solve a common task $T = \{Y, f(.)\}$. Typically, in DA there are a large number of source data points and no labeled data (or few samples) in the target dataset, so it is difficult to estimate the joint distribution $\hat{P}(X, Y)$. DA approximates $\hat{P}_T(X, Y)$ using the source data distribution estimate $\hat{P}_S(X, Y)$, which is possible since the two domains are correlated (Chattopadhyay et al., 2012; Venkateswara and Panchanathan, 2020).

DA is also used for learning to classify where relevant examples are sampled from the source data to train a classifier that can also classify target samples. Another approach projects data points into feature subspaces common to the source and target datasets, the classifier is trained to obtain the features in the common space of the source data. In these approaches, the source and target feature representations are predetermined and alignment

techniques are applied to reduce the domain distribution difference between domains. They can also be called shallow domain adaptation (Long et al., 2013, 2014; Venkateswara and Panchanathan, 2020).

In the seismic context, DA presents a solution to the situation where a model for indicating oil reserves was trained using data from a specific layer distribution but wants to be used in other regions that have a different layer distribution. In this case, the task is the same (indication of oil reserves), but there are differences between the training source and the target data.

There are variants of DA that introduce restrictions or have different characteristics:

1. Supervised or Semi-Supervised Domain Adaptation

   In this type of DA, the source dataset has labels for each sample, but the target domain only has a few samples with labels, which is insufficient to create a training model without the help of the source domain.

   The source domain has the dataset labeled $D_S = \{x_i^s, y_i^s\}_{i=1}^{n_s}$ and the target domain consists of the dataset $D_T = \{x_i^t, y_i^t\}_{i=1}^{n_t} \cup \{x_i^t\}_{i=n_t+1}^{n_t+n_u}$, where n_t are labeled samples and n_u are unlabeled, and $n_t \ll n_u$. With the small number of labeled samples (n_t) it is not possible to estimate the joint distribution $P_T(X, Y)$ without the risk of overfitting, but the source domain has a large number of samples labeled $n_t \ll n_s$, which can be used to perform training and align the distribution between the source and target domains (Venkateswara and Panchanathan, 2020; Venkateswara et al., 2015).

   From a seismic perspective, this is the case when there is a large amount of seismic images coming from a field that has already been analyzed and has labeled data, but now there is a new field in which there is some seismic images that were verified by exploration wells. In this case, the available seismic images from the new field are not enough to train a model and it is necessary to use both the dataset that has large labeled data, and the available data from the new field.

2. Unsupervised Domain Adaptation

   In this type of DA there are no labels for the target domain, which means that all the training data for the target task comes from the source domain, for this reason, the more similar the domains are, the easier the adaptation will be. However, in cases where the domains are very different, it is necessary to use several techniques to identify a common representation space for both domains.

In unsupervised DA, the source domain has data labeled $D_S = \{x_i^s, y_i^s\}_{i=1}^{n_s}$ and the target domain consists of unlabeled data $D_T = \{x_i^t\}_{i=1}^{n_t}$. The aim is to align the domains to approximate the target joint distribution $P_T = (X, Y)$ that allows the labels for the target domain to be estimated (Venkateswara and Panchanathan, 2020).

Within the seismic context, unsupervised DA can be explained by considering the case where a model for gas inference in 2D seismic images is trained using data coming from an exploitation field that was labeled by a professional expert, but the model needs to be used in another field that only has 2D seismic images without labels. In this case, there are differences between the field data such as the type of terrain, the team, and the equipment used to collect the data, which make it necessary to adapt the knowledge obtained from the exploitation field to the new one.

3. Unconstrained Label Spaces Domain Adaptation

   In the standard form of DA, the source and target domains have identical label spaces, meaning that the task of both domains has a similar label class, for example, identifying a single object or performing a classification with the same number of classes. In an unrestricted adaptation, the source and target label spaces can be different, even the target domain can have an extended label space.

   There are variants of this approach, the first called Partial DA is an approach in which the source label space is a superset of the target label space, this means that the source dataset has all the categories of the target dataset $Y_T \subset Y_S$. Another approach called Openset DA has an intersection between the label space of the source and target domains $Y_T \not\subset Y_S$, $Y_S \not\subset Y_T$ and $Y_S \cap Y_T \neq \varnothing$. There even exists a variant that considers the source space like a subset of the target label space $Y_S \subset Y_T$ (Cao et al., 2018; Geng et al., 2021; Saito et al., 2018).

   From a seismic perspective, Unconstrained label space domain adaptation can be used when there is a model that was trained for rock classification and needs to be adapted for use with a new set, which contains part of the rocks from the source domain, but they also present new types of rocks. In this case, the new model can take advantage of the feature extraction and part of the classification layer included in the original model.

4. Multisource Domain Adaptation

In this type of DA there is K source domain $D_S = \{D_1, D_2, \cdots, D_K\}$ and a target domain T, in this case there are multiple source datasets with labels to train the model and the target domain has a new (and unseen) dataset. The joint distribution $\hat{P}_T(X, Y)$ is different from each element of $\hat{P}_S = \left\{\hat{P}_{S_i}(X_i, Y_i)\right\}_{i=1}^{K}$, the aim of Multisource DA is to align the feature space representation between domains to reduce domain shift and enable knowledge transfer to approximate $\hat{P}_T$ using the source joint distribution estimate $\hat{P}_S$ (Venkateswara and Panchanathan, 2020; Zhao et al., 2020).

Within Multisource DA there are approaches that use the source domains labeled datasets to adapt the new model to be used with the target data, however, there are approaches that consider the fact that the source datasets are not accessible, these approaches focus on performing adaptation using only the models trained in each domain (Ahmed et al., 2021).

Within the seismic context, Multisource DA can be used in situations where there are multiple seismic datasets from different exploited gas fields that were used to train multiple gas inference models. Still, there is no guidance on which model may enable better performance or even whether it exists. In this case, there are several source datasets with different domains within them, from which DA can learn to create a new model for gas inference considering the features of the new seismic images.

The above list presents only some of the transfer learning methods used in the area of ML. Figure 2.3 shows a summary.

2.2.2 Connection of Transfer Learning with the Proposed Work

This section presents several features of the problem under study from a computer science perspective, which allows the problem to be placed within an ML context to determine the subarea on which the proposed solution is focused.

2.2.2.1 Target Task

For each field, the task is to indicate the location of the gas reservoir within the 2D seismic images. This means the model aims to recognize patterns within the training data to indicate gas reserves in new 2D seismic images.

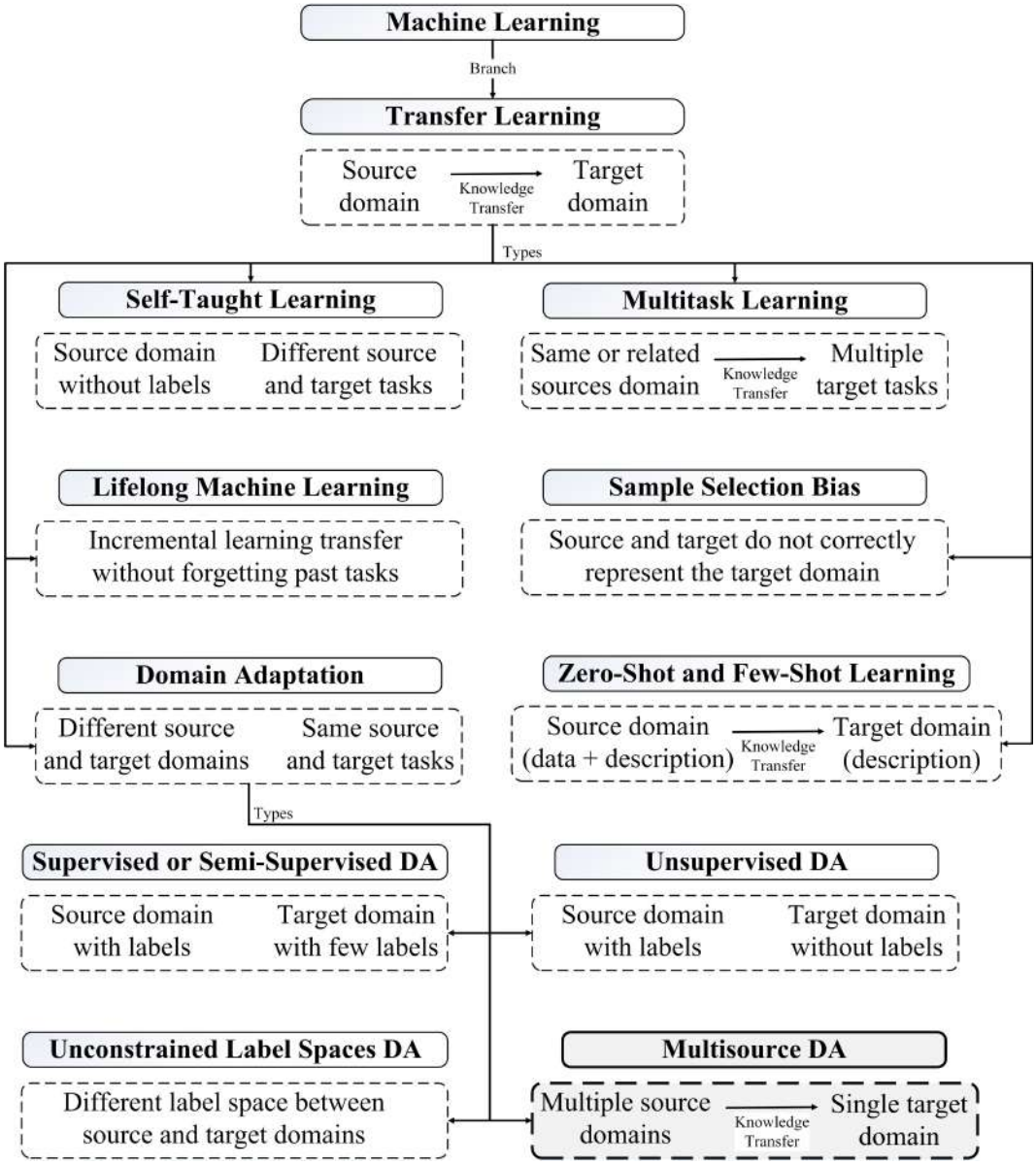


Figure 2.3: Types of Transfer Learning described.

According to the problem described in Section 1.3, this task has the restriction of not having labels for the 2D seismic images of the target field, so it is not possible to train a model using data from the same field, that is, all training data must come from other fields suggesting unsupervised learning process.

From a computer science perspective, this means that each field shares the same task, with an unsupervised learning process, but the study data needs to be analyzed to correctly identify the problem from a computer science perspective.

2.2.2.2 Dataset Characteristics

The problem described in Section 1.3 shows that datasets from different exploitation fields have related features, but there are important differences that do not allow the model trained on a specific field to maintain performance when used in data from another field.

This difference can be explained by the collection process described in Section 4.2.1, which shows that the data was collected on different dates by five collection teams, which also used unknown equipment parameterization, in addition, the fields may have unknown terrain variations.

Considering this dataset's characteristics is possible to affirm that there exist different source domains that may be used to train a new model to be used in a single target field.

2.2.2.3 Problem Characterization

Considering Sections 2.2.2.1 and 2.2.2.2 it can be concluded that there are different domains but they all have the same common task, and the general aim is to use all the available data from the training fields to apply them on the new unlabeled data.

This describes a case with multiple source domains with a single target domain that share the same task, and the aim is to adapt the knowledge of the source domains to generalize the performance of a learning model applied to the target domain.

According to Section 2.2, the type of learning technique that is closest to the previous conclusion is Multisource Domain Adaptation. Figure 2.4 highlights this type of learning within the classification described in the Figure 2.3.

2.3 Classification Techniques

The methods for the indication of natural gas include a classifier based on deep learning, which is trained using seismic traces, in this section the networks used are presented.

2.3.1 Long Short Term Memory Networks, LSTM

LSTM is a special type of recursive network, so data can be persisted by creating loops in the network diagram, allowing it to remember previous

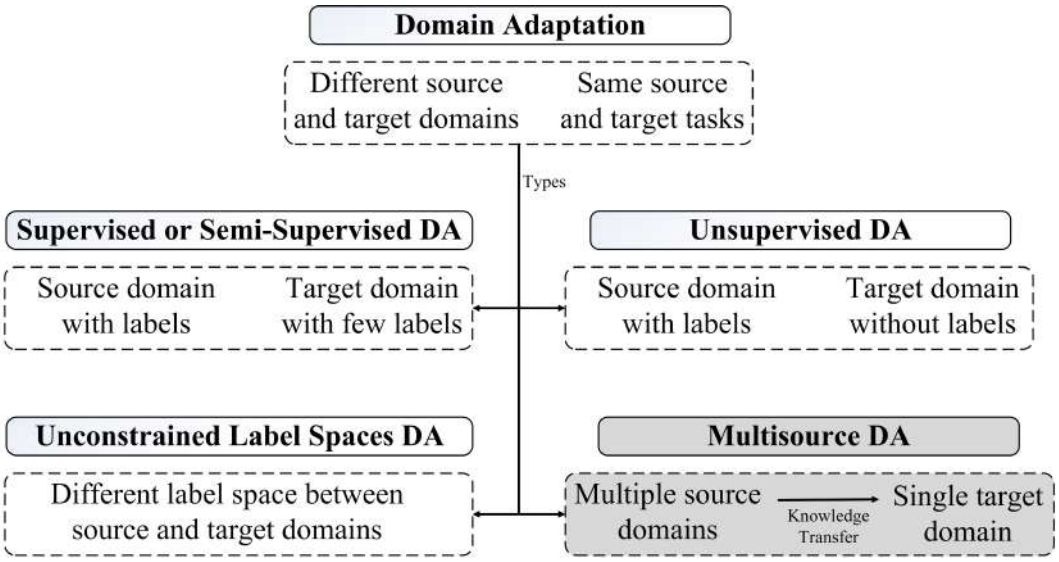


Figure 2.4: Learning technique closest to the problem under study.

states and use this information to decide the next states. LSTM was specifically designed to address the problem of long-term dependency, whereby information is remembered for a long period, for which it can remember, and decide what to forget and how much to learn or ignore from new data (Hochreiter and Schmidhuber, 1997; Ranjbar and Toufigh, 2022). Figure 2.5 present a LSTM cell example.

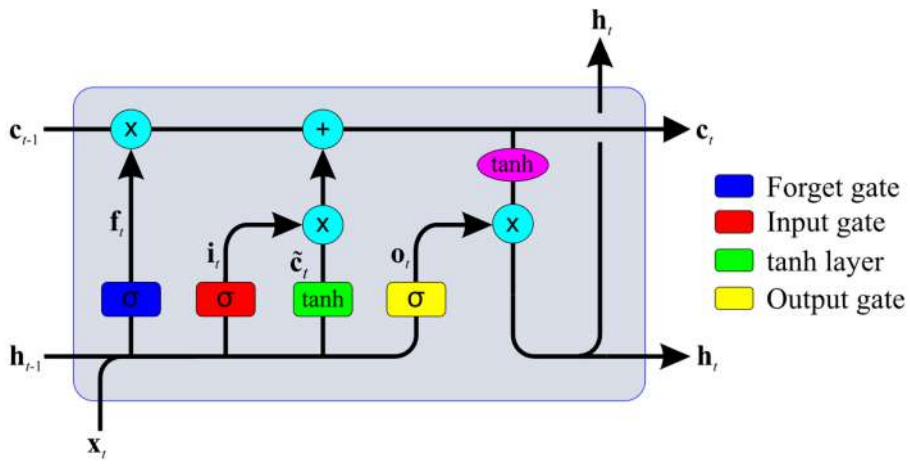


Figure 2.5: LSTM cell example, Adapted from Ranjbar and Toufigh (2022).

2.3.2 Gated Recurrent Unit, GRU

GRU is a variation of LSTM that has two gates, update and reset. The update gate indicates how much to keep from the previous cell, and the reset gate defines how much to incorporate from the new input. The GRU network compared to LSTM has fewer parameters, as it reduces the number of

gates, which allows training to be faster (Cho et al., 2014; Wang et al., 2022). Figure 2.6 present a GRU cell example.

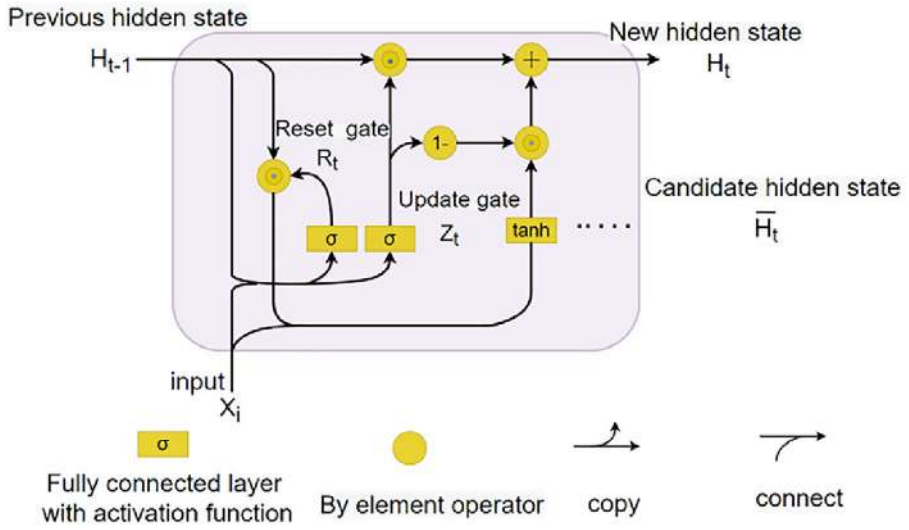


Figure 2.6: GRU cell example, Adapted from Wang et al. (2022).

2.4 Feature Extraction

In this section, features that represent each seismic image are extracted using different techniques, this was selected taking into consideration the ability to characterize attributes and affinity with the type of data under study, in addition to ensuring that each technique captures a different aspect or scale of seismic images.

2.4.1 Phylogenetic Index

Originally used in Biology, phylogenetic indices based on the diversity of species is a technique that allows quantifying the relationships between different individuals and species of a community. Diversity is a term frequently used in the area of ecology, where an index of diversity describes the variety of species present in a community or region, phylogeny is a branch of biology responsible for the study of the evolutionary relationships between the species in order to determine possible common ancestors (Magurran, 2004; Baxevanis and Ouellette, 2004).

Given the similarity of data format between the images and 2D seismic images, it is possible to use this technique as a feature extractor to measure the relationships between different seismic amplitudes (species) and the amount of each amplitude (individuals) in a seismic image (community) (da Cruz et al.,

2020; de Carvalho Filho et al., 2018; de Sousa Costa et al., 2018). To use this approach in images or seismic 2D data, it is necessary to make an analogy between its properties. Table 2.1 shows the correspondence between the terms.

Table 2.1: Matching Terms Proposed Between Biology and Image Processing.

Biology	**Image processing**	**Seismic**
Community	Image	Seismic image
Species	Level of intensity	Intensity
Individual	Pixel or voxel	Wave magnitude

Histogram is the base for the extraction of features of phylogenetic indices, Figure 2.7 shows an example of a histogram to be used to explain the terminology used in phylogenetic indices equations, The total number of different levels of amplitudes for this example is 9, represented by the s, the value of each amplitude is presented by the axis named "Amplitude". An axis called "Abundance" presents the amount of occurrence of each amplitude. This value is represented by the x_i, where i is the specific amplitude. For example for amplitude 1 (value -1.00), its abundance is $x_1 = 5$. w represents the distance that exists between two amplitudes, for example, $w_{1,2} = 0.25$, indicates that the distance magnitude between amplitudes 1 and 2 is 0.25. Finally, n represents the sum of all x_i for $i = (0, ..., s)$.

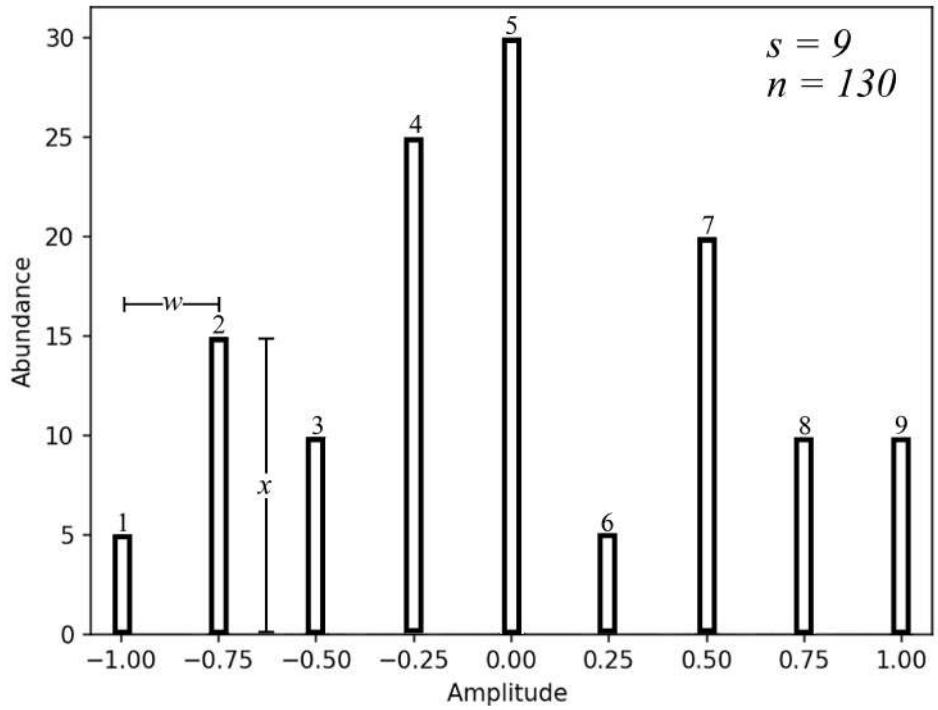


Figure 2.7: Seismic histogram and phylogenetic indices nomenclature example.

Taxonomic Diversity Index, Δ

Quantifies the average taxonomic distance between all amplitude pairs, which means that Δ quantifies how closely related the amplitudes are in a seismic image, Δ is computed by Equation (2-1).

$$\Delta = \frac{\sum \sum_{i<j} w_{ij} x_i x_j}{n\,(n-1)\,/2} \tag{2-1}$$

Mean Phylogenetic Distance, MPD

Represents the distance relationship between all pairs of individuals of different amplitudes and the number of individuals in each pair. This distance quantifies the separation that exists between the amplitudes in a seismic image. MPD is calculated using Equation (2-2).

$$MPD = \frac{\sum \sum_{i<j} w_{ij} x_i x_j}{\sum \sum_{i<j} x_i x_j} \tag{2-2}$$

Intensive Quadratic Entropy, I

Constitute the taxonomic relationship of amplitudes, which measures the relationship between the amplitudes without considering the number of individuals in each one. I is of special relevance when used in normalized images that present the same amplitudes since it is influenced only by their separation. I can be computed by Equation (2-3).

$$I = \frac{\sum_{i,j} w_{i,j}}{s^2} \tag{2-3}$$

Extensive Quadratic Entropy, E

Presents the magnitude of distance between all the amplitudes. This measurement is affected by the number of different amplitudes and their histogram distance, which means that the more amplitudes, or the further the amplitudes are, the greater the magnitude E. Equation (2-4) defines E.

$$E = \sum_{i,j} w_{i,j} \tag{2-4}$$

Average Taxonomic Distinction, AvTD

Considers the average of all distances between a pair of amplitudes to the total number of different amplitudes in the seismic image without considering the number of individuals. When used between the first and last amplitude, AvTD measures how tight the amplitudes are in spectral space. Equation (2-5) defines AvTD.

$$AvTD = \frac{\sum \sum_{i<j} w_{ij}}{s\,(s-1)\,/2} \tag{2-5}$$

Total Taxonomic Distinction, TTD

Represents the sum of the average distances between all amplitudes in the seismic image, TTD is calculated using Equation (2-6).

$$TTD = \sum_{i} \frac{\sum_{i \neq j} w_{ij}}{s-1} \tag{2-6}$$

Pure Diversity Index, PDI

Calculate the sum of the distances between each amplitude and its closest neighbor. PDI is given by Equation (2-7).

$$PDI = \sum w_{imin} \tag{2-7}$$

Mean Nearest Neighbor Distance, MNND

Calculate the mean sum of the distances between each amplitude and its nearest neighbor relative to the number of amplitudes. Equation (2-8) defines MNND.

$$MNND = \frac{\sum w_{imin}}{s} \tag{2-8}$$

2.4.2 Local Binary Pattern, LBP

Allows to extract a binary representation that maps the features of local texture in a image, applying a division mesh to the original image and then performing a component-by-component analysis to the central value of each subdivision (Kar and Banerjee, 2021; Mehmet Bilal, 2021; Pan et al., 2021b; Shu et al., 2021).

LBP performs a seismic image division and, from each subdivision, extracts a binary representation of the texture features created by comparing the value of the central amplitude with the neighboring amplitudes according to a comparison radius. Operationally, LBP uses Equation (2-9) to extract a binary representation of each subdivision:

$$LBP_{P,R} = \sum_{p=0}^{P-1} b(f_p - f_c) \times 2^P, b(x) = \begin{cases} 1, & x \geq 0 \\ 0, & x < 0 \end{cases} \tag{2-9}$$

where P identifies the amplitudes members used within each subdivision, R is the distance radius that identifies the sample amplitude, f_c is the equidistant central amplitude of radius R within the sample P, $f_p(p = 0, ..., P)$ is the neighboring amplitude that is at a distance R of f_c, and $b(x)$ is the comparison operator that assigns the value of 1 to neighboring amplitudes that have a value greater than or equal the central amplitude and 0 otherwise.

However, to refine the extracted features, a uniformity descriptor is applied the Equation (2-10):

$$U(LBP_{P,R}) = \mid b(f_{P-1} - f_c) - b(f_0 - f_c) \mid + \sum_{p=1}^{P-1} \mid b(f_p - f_c) - b(f_{p-1} - f_c) \mid \tag{2-10}$$

Finally, LBP considers the extraction of features that are invariant to the rotation of the image, using Equation (2-11).

$$LBP_{P,R}^{riu2} = \begin{cases} \sum_{p=0}^{P-1} b(f_p - f_c), & if\ U(LBP_{P,R}) \leq 2 \\ P+1 & otherwise \end{cases} \tag{2-11}$$

where the superscript ^{riu2} reflects the use of rotation-invariant "uniform" patterns that have a U-value of at most 2.

2.4.3 Discrete Fourier Transform, DFT

DFT is a discrete transform that is part of the Fourier analysis, which allows transforming a function $f(x)$ belonging to the time domain to obtain its representation $F(x)$ in the frequency domain, using equation (2-12) (Ahmadi et al., 2021; Chui et al., 2021; Osgood, 2019; Qi et al., 2021; Sandwell, 2021; Wu et al., 2021).

$$F(x) = \int_{-\infty}^{\infty} f(x)\, e^{-i2\pi kx} dx \tag{2-12}$$

where i is the imaginary component, x is the distance and k is the wave number, where $k = 1/\lambda$ and λ is the wavelength.

Feature extraction by DFT refers to the extraction of the frequencies that are present after the transformation of the seismic trace within the frequency space, as well as the number of times it occurs, this is equivalent to taking the histogram of the traces after performing the DFT transformation.

2.5 Feature Analysis

When using multiple methods to perform feature extraction, the potential for redundant data arises, so it is viable to use feature analysis techniques to reduce unrepresentative data.

The feature extraction process is carried out to identify attributes in the data that result in quantitative information of interest and allow differentiation of one object class from another, this process is also known as characterization.

The extraction of features in images seeks to obtain insensitivity to capture and lighting noise, in the same way, it must be independent of certain variations such as translation, rotation, scale, and transformations. Image characterization can be used in processes that require segmentation of elements that make up the image or can also be used to extract data that allows a classification of the entire image (Gonzalez and Woods, 2008).

Principal Component Analysis, PCA

The PCA has the purpose of transforming a set of variables, calls of originals, into a new set of variables called principal components. The new variables are linear combinations and are constructed according to the order of importance in terms of the total variability that they collect from the sample (Jolliffe, 2002).

The concept of more information is related to that of greater variability or variance. The greater the variability of the data (variance), is considered that there is more information. That is, the greater its variance, the greater the amount of information that this component has incorporated. For this reason, the one with the highest variance is selected as the first component, while the last component is the one with the lowest variance (Jolliffe, 2002).

After using PCA to perform principal component extraction, a trained model is also created that can be used to extract principal components from new data, in the same way, this model is called the PCA model.

2.6 Performance Metrics

The performance metrics used for natural gas indication are Accuracy, F1 Score, Intersection Over Union, Precision, and Recall. To assess the quality of the clustering, the silhouette coefficient is used.

Accuracy

Accuracy refers to the degree to which the predictions made by a model match the reality that is modeled, it is applied when the test data is labeled, it can be calculated as the number of classified objects correctly divided over the total number of objects. Table 2.2 shows the confusion matrix, on which Equation 2-13 is based to calculate the accuracy (Sammut and Webb, 2017b).

Table 2.2: Confusion Matrix.

		Predicted Class	
		Positive	Negative
True Class	Positive	TP	FN
	Negative	FP	TN

In Seismic the accuracy indicates how well all the pixels in the Gas and No Gas classes were classified, however, given the great imbalance that exists between these classes, it is not possible to use only this metric to evaluate the performance of the DL model, since a correct classification of the pixels of the "No gas" class hides the true performance of the "Gas" interest class.

$$Accuracy = \frac{TP + TN}{TP + FP + TN + FN} \tag{2-13}$$

where TP are the true positive cases, TN are the true negative cases, FP are the false positive cases and FN are the false negative cases.

Intersection Over Union, IoU

IoU is normally used in object detection. It is used to determine true positives and false positives in a set of predictions, based on Equation 2-14 (Rezatofighi et al., 2019).

$$\text{IoU} = \frac{TP}{TP + FP + FN} \tag{2-14}$$

Precision

Precision represents the ability of the predictor to correctly identify positive cases in relation to the total positive cases predicted by the model. Precision is defined in Equation 2-15 (Sammut and Webb, 2017b). In seismic Precision indicates how successful the model is in indicating the location of natural gas, a low precision generally indicates the high presence of false positives.

$$Precision = \frac{TP}{TP + FP} \tag{2-15}$$

Recall

Represents the estimator's ability to correctly identify cases that are considered true positives, represented in Equation 2-16 (Sammut and Webb, 2017b). In seismic, Recall determines how much natural gas the DL model is able to indicate correctly, disregarding false positives, that is, a high recall indicates that the model inferences match the geoscientist's marking labels.

$$Recall = \frac{TP}{TP + FN} \tag{2-16}$$

F1 score

The F1 score metric takes into account FP and FN to calculate the weighted average of Precision and Recall. It is used when classes are not balanced, its interpretation indicates that high values mean greater classification Precision. A high value indicates that the model performs indications that match the gas marking labels and at the same time has a low number of false positives. F1 score is represented by Equation 2-17 (Powers, 2011).

$$\text{F-measure} = 2 * \frac{Precision * Recall}{Precision + Recall} \tag{2-17}$$

Silhouette coefficient, S

The silhouette coefficient is used to measure how compact and well-separated the clusters are, using the Equation 2-18 (Rousseeuw, 1987).

$$S\left(i\right) = \frac{b\left(i\right) - a\left(i\right)}{\max\left\{a\left(i\right), b\left(i\right)\right\}} \tag{2-18}$$

Where $a(i)$ represents the mean distance of the ith element from all others in the same clusters, and $b(i)$ represents the mean distance of the ith element from all elements in the nearest cluster. The resulting value of $S(i)$ will be between $[-1, 1]$, where values close to 1 indicate very good clustering. Values around 0 indicate that there is an intersection between the clusters. Values close to -1 indicate a bad grouping or that the samples are not in any cluster.

3
Related Works

When searching in the state of the art, it was not possible to find works that had the same objective on the same type of data as those proposed in the present work, however, the selected approach to address the generalization problem implies various processes such as the seismic features extraction and clustering, which makes it possible to find works that present solutions to these processes.

3.1
Report of the Analysis of the State of the Art

This section presents the results of applying the literature review.

3.1.1
Principal Findings of Literature Review

The findings after carrying out the review of the state of the art, point to various paths that contribute to the solution of the problem posed, which in turn has various implications that are discussed:

1. The first finding is summarized in the fact that a method wasn't found to solve the problem posed by the specific data, this means that it was not possible to find a method that addresses the generalization problem of DL models designed for 2D seismic images.

2. When considering the various components of a DL model and its training phases, various strategies arise that can be applied to address the research problem, in this way, although a work that solves the investigation problem has not been found in the state of the art., it is possible to find multiple works that focus on each component and that seek to solve similar problems although with different data.

3. To face the problem, two different types of strategies are perceived, the first focuses on the analysis and transformation of the training data and the new data. The second strategy seeks to modify the DL models so that they adapt to data with new features.

4. Within the data analysis and transformation strategy, there is a tendency to use adversarial methods, to make the new data acquire features of the training data without losing relevant information.

5. Within the DL model modification strategy, there is a tendency to interpret features to find a common space that preserves the relevant features of both the training data and the new data. This means that in the first stages of execution of the DL models, data from the common space is extracted from the new data.

6. Some works propose the use of multitasking networks, which allow a grouping of data according to their features, although these works are not focused on improving the generalization of DL models, the use of data grouped by their features to carry out the training of the DL models offers a first approximation to address the research problem.

3.1.2 Generalization Works

In the review of the state of the art was possible to find generalization studies of DL models for geological data, although, with different tasks, that is why in this section a comparison is made focused on the processes, techniques and data management that allow generalization and not on the metrics that each study achieves since these are not comparable when dealing with different tasks:

In Xu et al. (2022) an adversarial Autoencoder-based method is proposed to contrast discrepancies in seismic data that allows feature extraction for the identification of lithological properties, which allows real-time interpretation of seismic data in well drilling, achieving the performance of 70.4% on average. This method seeks to obtain features of the latent space that ignore the particular properties of each domain, understanding by domain a certain and known depositional environment and recording equipment, which allows obtaining a generalization of the DL model for lithological identification.

The work of Xu et al. presents similarities with the methods proposed in the present work, in the first place, it is recognized that the data that are available for training, although they are of the same type, there are differences called domains that originate from the distribution of the field, and by the registry capture equipment and its configuration, secondly, the extraction of features that are representative for all types of domains is proposed, at this point, a difference between the works is presented, since Xu et al. has data that allows the available seismic images to be grouped in several domains,

however, the present work does not have these domain labels, which makes it impossible to use a contrastive approach in the way proposed by Xu et al., although the methods proposed in this work can be used as a feature extractor for lithological identification.

Another difference is the way of using the proposed methods since Xu et al. is designed to be used in real-time in well drilling, while the methods proposed in this work are designed as an interpretation aid tool that allows indicate the location of gas reservoirs, which leads to the drilling of wells. This is an important difference since Xu et al.'s method did not use the seismic data of the new target region to build the model as such data is not available, whereas the present work uses these data as a basis of comparison for defining the training data that allow a generalization of the DL model.

The most significant difference in the way feature extraction is performed between both methods is that Xu et al. takes advantage of available domain data so that Autoencoder's feature extraction model can ignore features associated with the particulars of the domain that do not represent relevant data for lithological identification, in comparison the methods proposed in this work, lacking domain annotation, focuses on the creation of a latent space that represents seismic images regardless of which domain they belong to.

All methods have the same type of limitation in terms of their generalization capacity, given by the representativeness limited only to the samples present in the training data, this means that although both methods are capable of extracting representative features from a latent space, their effectiveness when used for new data will be limited to the existence of representative samples within the training data, in other words, they depend on the existence of samples within the training data that belong to the same domain as the new samples.

Zhu et al. (2020) presents a method to improve the generalization of DL models to detect earthquake signals in seismic data by augmenting the training data, the results show data augmentation can mitigate the bias in training data and improve the performance of a dataset with different statistics. Zhu et al. reaffirm the dependence that exists in DL models with the diversity of training data to achieve good performance, in the same way, it shows the limitation that exists due to the lack of a large set of high quality training data, or the high cost associated with its construction considering good labeling and quality control. This method uses small training data sets to perform transformations that increase the number of samples and their variability, obtaining larger training data sets, however, Zhu et al. highlights the fact that not all transformation processes, commonly used in computer vision can be applied to seismic data,

since some of these processes violate the physical properties of the waveform data of interest, for this reason, it focuses only on those techniques compatible with seismic signals.

Although the form of data collection and the purpose of the model developed in Zhu et al. (2020) are different from the methods proposed in this work, there are similarities that allow a comparison, the most significant being that both methods do not require modification of DL models since they both focus on processing the training data.

Similarly, both methods consider that the quality of the labeling and the variability are factors that influence the generalization of the models, that is, there are different properties that make the training samples have features that allow establishing subsets, despite the fact that all samples belong to the same data type, however, the methods address these properties differently, which can be complementary for both works.

Zhu et al. focuses on the lack of data with enough variability to train a DL model, unlike the present work that focuses on recognizing the features of such variability that allow the creation of clusters. This difference is crucial since in the case of the Zhu et al. method the variability is increased, but the problem of how representative the data used in the training is about the new objective data continues since the DL model is trained to try to recognize all properties within variability, instead of focusing only on those that are most representative for the new data. In contrast, the methods proposed in the present work focus on recognizing the patterns that characterize the variability of the training set and comparing them with those of the new samples to propose a new training set, however, this approach is based on the premise that there is a set of samples that are significantly representative of the new target sample within the training set, and that are sufficient to train the DL model.

When comparing the limitations of all methods, it is found that they can be complementary, making this study a proposal to be developed in future works.

Mustafa and AlRegib (2021) presents a method that improves the generalization of a DL model for seismic phase segmentation, using an active learning approach. In this method, a generalization of the DL model with an Encoder-Decoder architecture is achieved by integrating an Autoencoder branch in the segmentation model, which, starting from the latent space created by the Encoder, adds a parallel Decoder branch for the reconstruction of the input data, in this way the model is trained for the segmentation task and at the same time the capacity of the model to process each data based on the reconstruction error is qualified, which allows the model to request that

the data with a high reconstruction error be labeled, since the representation of this data in the model is assumed to be weak, and more data with markup labels are required. This method achieved the mean Intersection-Over-Union value of 0.773 was tested in the Netherlands F3 block study.

At first, it could be considered that the method proposed by Mustafa and AlRegib does not have much similarity with the present work, however, both focus on the processing of training data and identify the properties of the samples that allow a clustering, in addition to providing the DL model with the most representative samples. Even so, both methods have different ways of working, the method proposed by Mustafa and AlRegib, identifies the samples that have different features, within the training process of the DL model for the fulfillment of the specific task of segmentation of seismic phases, which implies a modification of the original DL model, which differs from the methods proposed in this work, which performs the identification of features for each sample independently of the DL method for gas indication, which makes both tasks independent.

Each approach brings different properties, in the case of Mustafa and AlRegib, the method requests which specific samples it needs to be tagged, which involves active training, but thanks to this it makes the markup tags more representative for the DL model, on the other hand, the methods proposed in the present work uses unsupervised learning to identify the features of the training samples, separating the samples into clusters, in this way the method can be applied independently of the model DL selected but does not offer any consideration about the lack or not of the marking labels.

Another important difference is the way the generalization is achieved, the Mustafa and AlRegib method, obtains a model that was trained to recognize all the representative properties of the different domains found within the training database, on the other hand, the methods proposed in this work focuses on training a different model for each set of new seismic samples, giving the model the ability to recognize the dominant representative properties in the new seismic samples, in this way the resulting model focuses in recognizing the features that are present in the objective data.

Quamer Nasim et al. (2020) presents a method for seismic phase segmentation that allows generalization by domain adaptation using transfer learning. This method proposes a network architecture based on U-net that also uses residual blocks and transposed residual blocks for the phase segmentation task, using the labeled data of the source domain to perform the training and evaluation of the loss function. On the other hand, it integrates a Siamese architecture that is evaluated using a loss function based on the alignment correlation for

domain adaptation, which seeks to minimize the differences between the data extracted from the source domain and the target domain, in this way the DL model learns to extract data that ignores the particularities of each domain, but is still representative for the seismic phase segmentation task, and at the same time achieve a generalization of the model for the target domain. The test was made on the public F3 block 3D dataset from offshore Netherlands and Penobscot 3D survey data from Canada, the maximum class accuracy achieved was 99% for Penobscot class 2 with $> 50\%$ overall accuracy.

Quamer Nasim et al. presents a method that offers the opportunity to compare the methods proposed in this work from a more traditional point of view regarding the problem of generalization of DL models. Both methods require new training for each new domain, in the same way for both methods there is no certainty that the new data really belongs to an unknown domain and that they are not represented within the training database, but it is known that there are no markup labels for the new data. Another similarity is that both methods seek, in addition to a generalization, to fulfill a specific task using seismic images, recognizing that the training samples have particular features that depend on factors such as the distribution of the terrain, which vary depending on the density, type of rock, porosity and permeability of the earth.

However, the methods have significant differences: first, Quamer Nasim et al.'s method requires modifying the original model to include the domain-adaptive Siamese network, which causes the generalization process to be performed at training time, compared to the methods proposed in this paper, which does not require original DL model modification, this difference has implications in the way the training is performed, since for the Quamer Nasim et al. method, it is necessary to generalize the model for each new seismic sample, in comparison with the methods proposed in this work, it performs training data clustering independently of the DL model for gas reserve indication, this means that only the model for gas search is trained based on the training data selected for each new set of seismic samples, which does not require repeating the original training database clustering process. A second difference is that the Quamer Nasim et al. method can perform generalization even when the new samples belong to a domain that is not represented in the training database, compared to the methods proposed in the present work, it can only recommend training samples that are closest to the new seismic samples features.

Another difference is that Quamer Nasim et al. defines a domain about the original data region, in this way it does not consider that multiple domains

can exist within a region, in this way when training the DL model, it is not done the separation of the samples based on their features to then carry out a domain adaptation. Considering this difference, the methods proposed in the present work could be used as a pre processing, identifying the samples that belong to a single domain and then carrying out the adaptation, which would facilitate learning by the DL model, this new application could be considered for future studies.

3.1.3 General Report of Related Works

Table 3.1 - Table 3.3 present the most relevant works found in the review of the state of the art.

Table 3.1: Data collected by reviewing the state of the art.

First Author	Publication source	Date	Paper Name	Main topic area	Search Terms	Research question/issue	Study data	Main techniques used	Main contribution
A. Veillard	80th EAGE Conference and Exhibition 2018: Opportunities Presented by the Energy Transition	2018	Fast 3D Seismic Interpretation with Unsupervised Deep Learning: Application to a Potash Network in the North Sea (Veillard et al., 2018)	Structure seismic interpretation	unsupervised AND deep AND learning AND training AND seismic	How enabling fast and accurate interpretation of 3D geological objects from seismic data?	3D seismic wave amplitude data	Autoencoder and GANS	GAN architecture
Arulmurugan pradhan	Computational Geosciences volume	2020	Seismic Bayesian evidential learning: estimation and uncertainty quantification of sub-resolution reservoir properties (Pradhan and Mukerji, 2020)	Estimation of low-dimensional sub-resolution reservoir properties	autoencoder AND seismic AND data	Estimation of low-dimensional sub-resolution reservoir properties directly from seismic data, without requiring the solution of a high-dimensional seismic inverse problem	3D pre-stack seismic data	Bayesian evidential learning approach	Incorporation of non-linear statistical models for seismic estimation problems
Fanbi Meng	IEEE Geoscience And Remote Sensing Letters	2022	SelfSupervised Learning for Seismic Data Reconstruction and Denoising (Meng et al., 2022)	Seismic denoising	no labeled AND data AND seismic.	in the seismic denoising field, collecting large numbers of labeled samples is impossible; thus, the main challenge to using deep learning methods is a lack of labeled data.	Synthetic and field stacked seismic data	J-invariant funtion, U2 net, feature in different scale extraction	Feature extraction at different scales
Feng Qiao	GEOPHYSICS (Society of Exploration Geophysicists)	2018	Unsupervised seismic facies analysis via deep convolutional autoencoders (Qiao et al., 2018)	Seismic Attributes	unsupervised AND deep AND learning AND training AND seismic	How to highlight stratigraphic and depositional information from prestack seismic data?	Prestack seismic data	Deep convolutional autoencoder	Seismic facies creation
Filippo Gatti	Computer Methods in Applied Mechanics and Engineering	2020	Towards blending Physics-Based numerical simulations and seismic databases using Generative Adversarial Network (Gatti and Clouteau, 2020)	Earthquake ground motion prediction	unsupervised AND deep AND learning AND training AND seismic	How to blend the outcome of physics-based numerical simulations with massive but poorly-labeled experimental databases such as in-situ data routinely recorded for monitoring purposes?	broadband seismic signals	adversarial learning techniques	Adversarial learning techniques
Guoyin Zhang	Journal of Natural Gas Science and Engineering	2022	Seismic characterization of deeply buried paleocaves based on Bayesian deep learning (Zhang et al., 2022b)	paleocaves shape identification	autoencoder AND seismic AND data	Deeply buried paleocaves exist widely in the world and act as an important type of hydrocarbon reservoirs. Three-dimensional (3-D) seismic data are commonly used to detect deeply buried paleocaves, but regular interpretation methods can hardly characterize their shape and uncertainty.	Syntetic Three-dimensional (3-D) seismic data	Bayesian encoder-decoder model	Bayesian encoder-decoder and estimates uncertainty
Jingcao Li	Expert Systems With Applications	2021	Feature concatenation for adversarial domain adaptation (Li et al., 2021)	Adversarial domain adaptation	feature AND domain AND adaptation.	The domain-invariant feature representation guarantees the transferability. However, to obtain domain-invariant features, certain domain-specific information is suppressed, which may cause the loss of discriminability. To this end, we aim to enhance the discriminability by enriching the information contained in the domain-invariant features	Objects image.	Adversarial domain adaptation	Feature Concatenation for adversarial Domain Adaptation (FCDA)
Jinsheng Jiang	IEEE Geoscience And Remote Sensing Letters	2022	A Convolutional Autoencoder Method for Simultaneous Seismic Data Reconstruction and Denoising (Jiang et al., 2022)	Reconstruction and denoising of seismic data	autoencoder AND seismic AND data	How to be representative features of seismic data by removing random noise?	Synthetic seismic wave amplitude data	Autoencoder	Feature extraction and analysis
Juan Manuel Davila Delgado	Applied Soft Computing	2021	Deep learning with small datasets: using autoencoders to address limited datasets in construction management (Delgado and Oyedele, 2021)	data augmentation and generation of synthetic data.	autoencoder AND seismic AND data	Poor data management practices and the low level of digitisation of the construction industry represent a big hurdle to compiling big datasets; which in many cases can be prohibitively expensive.	Synthetic and fiel seismic data	undercomplete, sparse, deep and variational autoencoders	Variational autoencoders
Kai Zhang	Journal of Petroleum Science and Engineering	2022	Unsupervised-learning based self-organizing neural network using multi-component seismic data: Application to Xujiahe tight-sand gas reservoir in China (Zhang et al., 2022b)	Unsupervised learning gas reservoir prediction	unsupervised AND deep AND learning AND training AND seismic	The tight sandstone reservoirs of the Xujiahe Formation of the Western Sichuan Basin Depression typically exhibit characteristics such as low porosity, low permeability, and strong heterogeneity, resulting in weak seismic response differences between gas and water.	Seismic wave amplitude data	self-organizing neural network (SOM)	self-organizing neural network (SOM)
Kejiu Hu	Structures	2022	Mode shape prediction based on convolutional neural network and autoencoder (Hu and Wu, 2022)	predict the mode shape	autoencoder AND seismic AND data	Mode shape is a dynamic characteristic that plays an important role in civil engineering. In this paper, an approach to predict the mode shape of a bridge is proposed using a convolutional neural network (CNN) and an autoencoder.	bridge mode shape	convolutional neural network (CNN) and an autoencoder.	Finite element method feature extractor
Kim, Minkyu	Journal of Engineering Mechanics	2022	Near-Real-Time Identification of Seismic Damage Using Unsupervised Deep Neural Network (Kim and Song, 2022)	Seismic damage identification	autoencoder AND seismic AND data	Prompt identification of structural damage is essential for effective postdisaster responses. To this end, this paper proposes a deep neural network (DNN)-based framework to identify seismic damage based on structural response data recorded during an earthquake event.	Operational modal analysis (OMA) matrix	self-supervised DNN	Variational Autoencoder
Kislov K. V.	Seismic Instruments	2018	Deep Artificial Neural Networks as a Tool for the Analysis of Seismic Data (Kislov and Gravirov, 2018)	Multiple DL aplication on seismic data	autoencoder AND seismic AND data	Possibility of applying deep networks in seismology	Single wave amplitude data	NN, autoencoder, multitask	Multiple application review
Kislov, K.V	Solid Earth	2020	Possibilities of Seismic Data Preprocessing for Deep Neural Network Analysis Kislov et al. (2020)	Preprocessing for reduce the noise level, to remove the anthropogenic noise, and to reduce the dimensionality of the data	autoencoder AND seismic AND data	Adequate preprocessing can increase the efficiency of the further analysis by order and more. However, specialized preprocessing cannot be used for solving other tasks or with other postprocessing algorithms	Single wave amplitude data	Wavelet transform, autoencoder	Preprocessing study
Kunhong Li	IEEE Geoscience And Remote Sensing Letters	2022	Simultaneous Seismic Deep Attribute Extraction and Attribute Fusion (Li et al., 2022)	Feature and attributes extraction and Fusion	multi AND task AND autoencoder.	Data-driven deep attributes bring new challenges to interpretation as they lack the support of intrinsic physical mechanisms.	Seismic wave amplitude data	Autoencoder	Extract deep seismic attributes and merge traditional seismic attributes simultaneously
Kunhong Li	Artificial Intelligence in Geosciences	2020	Seismic labeled data expansion using variational autoencoders (Li et al., 2020)	Automatic seismic tagging	unsupervised AND deep AND learning AND training AND seismic	Supervised machine learning algorithms have been widely used in seismic exploration processing, but the lack of labeled examples complicates its application.	Synthetic seismic wave amplitude data	Variational autoencoders	Variational features extraction

Table 3.2: Data collected by reviewing the state of the art.

First Author	Publication source	Date	Paper Name	Main topic area	Search Terms	Research question/issue	Study data	Main techniques used	Main contribution
Kyungbook Lee	Journal of Petroleum Science and Engineering	2018	Feature extraction using a deep learning algorithm for uncertainty quantification of channelized reservoirs (Lee et al., 2018)	efficient uncertainty assessment	autoencoder AND seismic AND data	Reservoir models are generated by geostatistics using available static data. However, there is inherent uncertainty in the reservoir models due to limited information. A number of reservoir models with equivalent probabilities are created to quantitatively assess model uncertainty.	reservoir image	stacked autoencoder (SAE)	Stacked autoencoder (SAE)
Liu Xingge	Geophysical Prospecting	2021	Semi-supervised deep autoencoder for seismic facies classification (Liu et al., 2021)	Clustering and classification	autoencoder AND seismic AND data	Facies identification based on supervised machine learning methods usually requires a large amount of labelled data, which are sometimes difficult to obtain	3D seismic data	DAE, SSDAE	Structure of semi-supervised deep autoencoder (SSDAE)
Liu Xingge	IEEE Transactions on Geoscience and Remote Sensing	2022	Deep Classified Autoencoder for Lithofacies Identification (Liu et al., 2022)	Lithofacies classification	autoencoder AND seismic AND data	Lithofacies classification is an indispensable procedure in well logging and seismic data interpretation	Lithofacies and trace seismic data	deep classified autoencoder learning	Sparse autoencoder
Maximilian Trapp	Mechanical Systems and Signal Processing	2019	Intelligent optimization and machine learning algorithms for structural anomaly detection using seismic signals (Trapp et al., 2019)	Anomaly detection	unsupervised AND deep AND learning AND training AND seismic	he lack of anomaly detection methods during mechanized tunnelling can cause financial loss and deficits in drilling time	synthetic SpecFEM3D wave.	Multi-output supervised machine learning	Multi-output supervised machine learning
Mostafa Mousavi	IEEE Geoscience and Remote Sensing Letters	2019	Unsupervised Clustering of Seismic Signals Using Deep Convolutional Autoencoders (Mousavi et al., 2019)	Feature extraction and clustering	autoencoder AND seismic AND data	supervised learning and relies on labeled data where the quantity and quality of labeled training sets play an important role in determining the effectiveness of an algorithm, however, what can be done when large labeled data sets are not available?	Earthquake waveforms	Autoencoder with Kullback-Leibler	Clustering Autoencoder
Mustafa, Ahmad	2021 IEEE International Conference on Image Processing (ICIP)	2021	Man-Recon: Manifold Learning For Reconstruction With Deep Autoencoder For Smart Seismic Interpretation (Mustafa and AlRegib, 2021)	Domain Adaptation	unsupervised AND deep AND learning AND training AND seismic	Deep learning can extract rich data representations if provided sufficient quantities of labeled training data. For many tasks however, annotating data has significant costs in terms of time and money.	Seismic wave amplitude data	Autoencoder with active learning	Active learning methodology based on learning reconstruction manifolds with deep autoencoders
Na Yang	Natural Resources Research	2022	Mineral Prospectivity Prediction by Integration of Convolutional Autoencoder Network and Random Forest (Yang et al., 2022)	Features extractor	autoencoder AND seismic AND data	The convolutional neural networks used widely in mineral prospectivity prediction usually perform mixed feature extraction for multichannel inputs. This results in redundant features and impacts further improvement of predictive performance.	Stream sediment data	Parallel Convolutional autoencoder networks	Parallel feature data extraction
Pengcheng Xu	Journal of Geophysics and Engineering	2019	A semi-supervised learning framework for gas chimney detection based on sparse autoencoder and TSVM (Xu et al., 2019)	Unsupervised and semisupervised learning	autoencoder	How to solve the problem of the lack of labeled data that often limits the applications of supervised classifiers?	Seismic wave amplitude data	Sparse autoencoder (SAE) and the transductive support vector machine (TSVM)	Learning process and combination of techniques.
Qian, Feng	IEEE Transactions on Geoscience and Remote Sensing	2022	DTAE: Deep Tensor Autoencoder for 3-D Seismic Data Interpolation (Qian et al., 2022)	Seismic trace interpolation	autoencoder AND seismic AND data	The core challenge of seismic data interpolation is how to capture latent spatial-temporal relationships between unknown and known traces in 3-D space.	3D seismic data	deep tensor autoencoder (DTAE), tensor back-propagation (TBP),	DTAE
Qingkai Kong	Artificial Intelligence in Geosciences	2021	Deep convolutional autoencoders as generic feature extractors in seismological applications (Kong et al., 2021)	Feature extraction	feature AND extraction AND seismic	The idea of using a deep autoencoder to encode seismic waveform features and then use them in different seismological applications is appealing	Earthquake waveforms	Overcomplete autoencoder	Feature extraction
Qiuner Nasim	IEEE Transactions on Geoscience and Remote Sensing	2022	Seismic Facies Analysis: A Deep Domain Adaptation Approach (Qiuner Nasim et al., 2020)	Domain Adaptation	seismic AND domain AND adaptation.	Deep neural networks (DNNs) can learn accurately from large quantities of labeled input data, but often fail to do so when labeled data are scarce.	3D seismic wave amplitude data	EarthAdapNet (EAN)	Unsupervised deep domain adaptation network (EAN-DDA)
Seyvandi Kandanaarachchi	Information Sciences	2022	Unsupervised anomaly detection ensembles using item response theory (Kandanaarachchi, 2022)	unsupervised anomaly detection	autoencoder AND seismic AND data	Constructing an ensemble from a heterogeneous set of unsupervised anomaly detection methods presents challenges because the class labels or the ground truth is unknown.	Outlier Detection DataSets (ODDS)	Item Response Theory (IRT)	Item Response Theory (IRT)
Shulin Pan	Computers & Geosciences	2020	A partial convolution-based deep-learning network for seismic data regularization (Pan et al., 2020)	reconstruct the missing data	autoencoder AND seismic AND data	Spatial undersampling is a common problem in actual seismic data due to limitations in seismic survey environments, which can be satisfactorily solved by data regularization. The convolution-based deep-learning reconstruction methods require fewer assumptions than the conventional reconstruction methods (e.g., Curvelet-domain and F-X domain data regularization methods). However, the traditional convolution methods are not suitable for the large percentages of missing data.	pre-stack and post-stack seismic data	partial convolution-based (PConv-based) deep-learning network	Hierarchical regional learning

Table 3.3: Data collected by reviewing the state of the art.

First Author	Publication source	Date	Paper Name	Main topic area	Search Terms	Research question/issue	Study data	Main techniques used	Main contribution
SiBo Zhang	Petroleum Science	2022	A comparison of deep learning methods for seismic impedance inversion (Zhang et al., 2022c)	seismic inversion problem.	autoencoder AND seismic AND data	Deep learning is widely used for seismic impedance inversion, but few work provides in-depth research and analysis on designing the architectures of deep neural networks and choosing the network hyperparameters.	3D seismic wave amplitude data	multi-scale architecture	Network arquitectura comparison
Smith W.A. Caudinumni	Computers & Geosciences	2019	Towards a robust parameterization for conditioning facies models using deep variational autoencoders and ensemble smoother (Canchumuni et al., 2019)	construction of a continuous parameterization of facies	autoencoder AND seismic AND data	Ensemble-based methods have been applied with remarkable success for data assimilation in geosciences. However, they sometimes fail to preserve the geological realism of the model, which is particularly evident in reservoirs with complex facies distributions.	Syntetic channelized facies model	convolutional variational autoencoder and the ensemble smoother with multiple data assimilation	Multiple data assimilation
Stefanos Nikolopoulos	Engineering Applications of Artificial Intelligence	2022	Non-intrusive surrogate modeling for parametrized time-dependent partial differential equations using convolutional autoencoders (Nikolopoulos et al., 2022)	Modeling differential equation	autoencoder AND seismic AND data	Recent advances in the field of computational mechanics have allowed researchers to develop high-fidelity models of complex physical systems that emulate their behavior. With this approach, the response of a system under investigation can be efficiently predicted via computer simulations in lieu of computationally costly and time-consuming experiments. However, certain applications of practical interest such as optimization, uncertainty quantification and parameter identification require a large number of model runs.	differential equation f.unitions	Autoencoder + MLP	Feature extraction
Ting Xu	Journal of Natural Gas Science and Engineering	2022	Domain generalization using contrastive domain discrepancy optimization for interpretation-while-drilling (Xu et al., 2022)	Generalizability	domain AND generalization	The collected logging data has the same probability distribution even if it comes from different wells. In this way, the model trained on multiple drilled wells could be directly used in a new well. However, this assumption is invalid in practice since there is always a large difference in depositional environment and logging equipment.	Seismic well data	Contrastive domain discrepancy based adversarial autoencoder (CDD-AAE)	Adversarial Autoencoder
Wang, Yingying	Geophysics	2020	Seismic trace interpolation for irregularly spatial sampled data using convolutional autoencoder (Wang et al., 2020a)	Seismic trace interpolation	autoencoder AND seismic AND data	Seismic trace interpolation is an important technique as irregular or insufficient sampling data along the spatial direction may lead to inevitable errors in multiple suppression, imaging and inversion.	Synthetic and field seismic trace	Autoencoder, transfer learning	Seismic corrupt data reconstruction
Weiqiang Zhu	Advances in Geophysics	2020	Seismic signal augmentation to improve generalization of deep neural networks (Zhu et al., 2020)	Generalizability	model AND generalization	A sufficiently large and complete training data set is a requirement that can be difficult to meet due to the significant effort and time involved in data collection and labeling.	Earthquake waveforms	Data Augmentation	Filters for data augmentation specific for seismic
Youngbin Ahn	Journal of Petroleum Science and Engineering	2022	Reliable channel reservoir characterization and uncertainty quantification using variational autoencoder and ensemble smoother with multiple data assimilation (Ahn and Choe, 2022)	channel reservoir characterization and uncertainty quantification	autoencoder AND seismic AND data	Reservoir characterization is essential for reliable performance prediction and decision making.	2D channel reservoir models	variational autoencoder(VAE) and ensemble smoother with multiple data assimilation(ES-MDA)	EAV Resulting Vector Enhancement Using ES-MDA
Zengmao Wang	IEEE Transactions on Neural Networks and Learning Systems	2020	Domain Adaptation With Neural Embedding Matching (Wang et al., 2020b)	Domain adaptation	domain AND adaptation	How to transfer information from the source domain to the target domain where labeled data is scarce?	Images of numbers and objects.	Neural embedding matching (NEM)	Progressive Learning
Zhenjiao Jiang	Geothermics	2021	Combining autoencoder neural network and Bayesian inversion to estimate heterogeneous permeability distributions in enhanced geothermal reservoir: model development and verification (Jiang et al., 2021)	Permeability distribution inference	autoencoder AND seismic AND data	Determining permeability distributions in reservoirs is critical for the management of limited earth resources. While hydraulic fracturing is widely used to enhance the permeability of deep geothermal, gas and oil reservoirs, it remains challenging to infer heterogeneous distributions of permeability.	Syntetic 3D image of fracture probability	neural network + Bayesian inversion algorithm based on Markov Chain Monte Carlo (MCMC)	Neural network + Bayesian inversion algorithm based on Markov Chain Monte Carlo (MCMC)

4
Generalization of Natural Gas Reserve Indication Deep Learning Model Based on Four Feature Extraction Techniques and Training Dataset Recommendation - Method 1

This method analyzes training data based on features extracted by four techniques, creating multiple training clusters based on similarity. A recommendation model is then used to select clusters that have the most similar features to the target seismic images. This method is proposed to provide a solution to the problem presented in Section 1.3, which is addressed as a generalization performance problem according to Section 2.2.

The specific contributions of method 1 to the state of the art are: First, the creation of a method to cluster seismic images based on the similarity of their features in an unsupervised manner. This clustering improves the performance of DL models on target images that is different from that used in the training process but shares some features. This approach does not require modifying the original network architecture for gas indication. Secondly, a method is developed to extract features in seismic images, introducing the technique of phylogenetic indices and Short-Time Fourier Transforms for each seismic trace. Third, a basis is established for the seismic features comparison that allows measuring similarity in multiple domains. This is possible because the four extraction techniques allow different features of the seismic images to be measured. Fourth, from the point of view of DL training models, method 1 presents evidence that supports the importance of training sets analysis, showing that not all the data are suitable, which affects generalization performance. Finally, the proposed method 1 improved the generalization performance of the DL model.

4.1
Proposed Method

This section describes the method, techniques, and models used to detect gas in 2D seismic images, based on the use of resources that belong to both ML and DL.

Figure 4.1 provides an overview of the proposed method 1, which consists of three sub-processes: clusterization, recommendation, and classification. The

first two processes are considered ML algorithms to select the training data used in the last DL-based natural gas indication process. Note that clusterization and recommendation processes are independent of the DL model. This independence allows the DL model to be changed without having to rebuild the training data set.

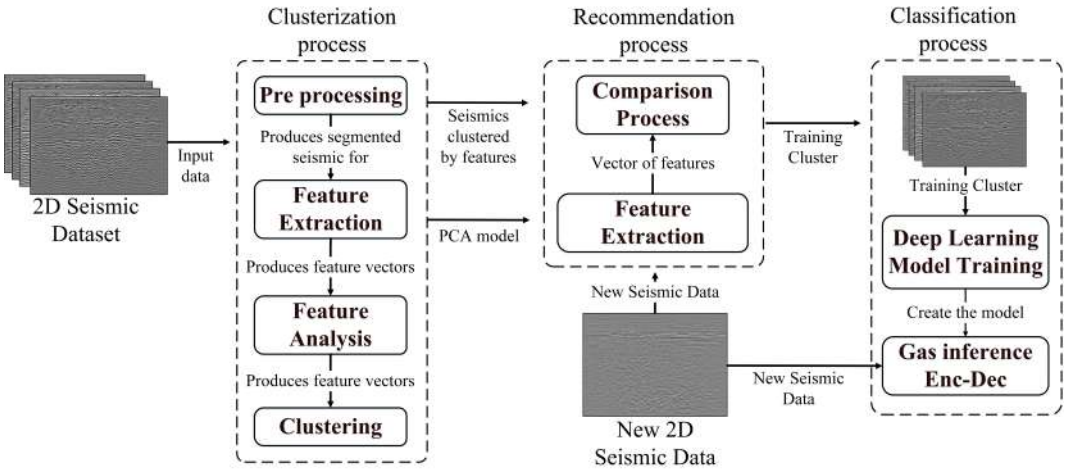


Figure 4.1: Proposed method 1 for generalization based on dataset recommendation.

4.1.1 Clusterization Process

The creation of clusters is the basis of the method 1. These clusters seek to group seismic images with similar features. The challenge here is that clusters are unknown, i.e., there is no ground truth to train a DL model to find them. In short, this section proposes a method that takes all available seismic images as input and produces a training dataset as output, Figure 4.2 present the pipeline.

4.1.1.1 Pre processing

This step aims to prepare the original seismic images for the feature extraction process, through the selection of study subregions and normalizing the amplitude.

This step performs three operations on the original seismic images. The first step is to divide the seismic image in the horizontal direction according to the region of interest. This region is obtained through an analysis of the geology, which reveals the region where gas reserves are likely to exist. The professional responsible for labeling delimits an area between two horizons, one close to the surface and another deeper. The region between these horizons is

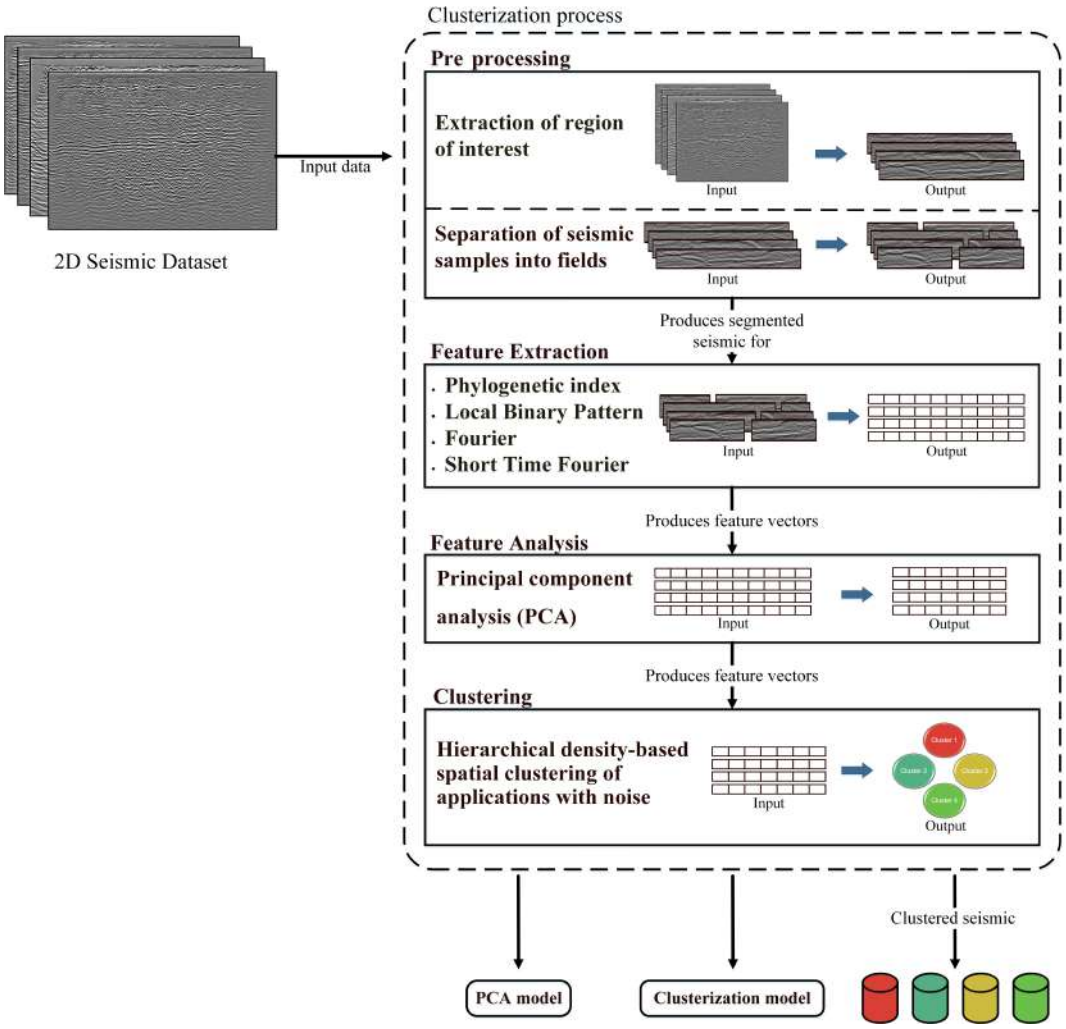


Figure 4.2: Clusterization process pipeline.

also called the region of interest, ROI, it is in this region where the studies will be carried out.

The second split each seismic image according to the exploration field, i.e., we divide the 2D seismic image vertically so that each sub-image belongs to only one exploration field, Figure 4.3 presents an example showing how a seismic image contains traces that belong to more than one field. By separating the image into subregions, sets of traces that belong to a single field are obtained.

The third operation is on the amplitude scale. The seismic amplitude varies according to unknown parameters in the acquisition of seismic images (see Section 4.2.1). For this reason, it is also necessary to apply normalization to put each image on the same scale. Here all amplitudes are scaled to fit in the interval $[-1, +1]$ using the Equation (4-1).

$$N(x) = 2\frac{x - min\ x}{max\ x - min\ x} - 1 \tag{4-1}$$

where x is the 2D image and $N(x)$ is the 2D image normalized between [-1,1].

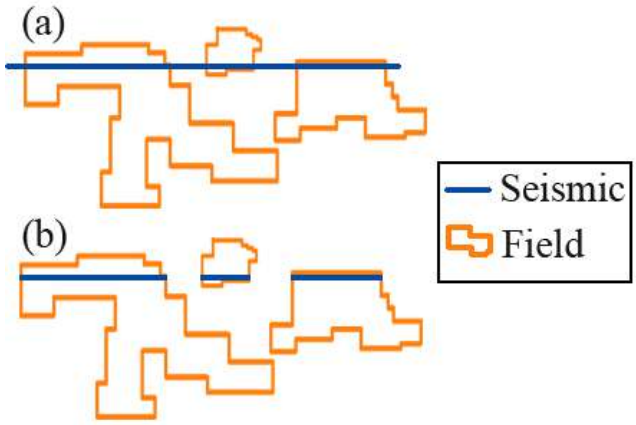


Figure 4.3: Seismic image: (a) original seismic crossing multiple fields, (b) segmented seismic.

4.1.1.2 Feature Extraction

In this section, features that represent each seismic image are extracted using four techniques: Phylogenetic index that extracts 8 features, Local Binary Pattern that extracts 10 features, Fourier that extracts 2 features, and Short Time Fourier that extract 3 features, presented in Section 2.4. Techniques were selected, taking into consideration the ability to characterize attributes and affinity with the type of data under study, in addition to ensuring that each technique captures a different aspect or scale of seismic images. In total, 23 features are obtained from each seismic image, stored in a single vector.

4.1.1.3 Feature Analysis

The extracted feature vectors for each seismic image are analyzed to concentrate the features and reduce the size of the vectors, this process is carried out to identify more representative features, using the PCA technique described in Section 2.5.

As a preprocessing step, the feature vector must be centralized in zero and normalized. The PCA algorithm then reduces the size of the feature vector and produces a model that is saved for the analysis of new images.

4.1.1.4 Clustering

This section aims to create several clusters of seismic images that are similar based on their extracted features. Likewise, a clustering model is

created that allows new seismic images to identify which cluster they belong to.

The reduced feature vectors created in section 4.1.1.2 are then separated into clusters of similar features. The idea of using a cluster as a training set is based on the hypothesis that a DL model trained in one cluster yields better results for images in the same cluster. In this way, when using the DL gas inference model on images coming from a new exploration field, First, the cluster with similar features is determined and then the training is carried out with the seismic images contained in the cluster.

Hierarchical density-based spatial clustering of applications with noise, HDBSCAN, (Lentzakis et al., 2020; Lin et al., 2019) is used as a clustering technique taking into account several properties of this technique that satisfy the limitations of the problem under study. The first property is that does not require specifying the number of clusters to create, which is necessary since there is no ground truth that indicates how many clusters it is possible to separate the seismic images based on their similarity.

The second property is that it does not require specifying the maximum distance between group samples. This property is essential to deal with the restriction of not having a ground truth that allows determining this distance in the feature space.

The third property is that HDBSCAN also identifies samples that do not fit any cluster and stores them in a cluster as outliers. This property prevents clusters containing samples that may corrupt the DL training process.

The last property specifies the minimum number of samples that each cluster must contain, preventing clusters from being too small to be useful in the DL model training process.

In summary, the HDBSCAN performs two tasks. The first is the separation of the seismic images into clusters. The second is the indication of a cluster that is more similar to a new seismic image.

4.1.2 Recommendation Process

This step aims to extract features from the target seismic images to compare with each training cluster to identify the most similar seismic images set, thereby creating a recommended set of training seismic images for the DL gas inference model. This step describes how the new seismic images are compared to the clusters to determine the training set for the DL models.

As input data, this step uses clusters and the clusterization model from Section 4.1.1.4, the PCA model from Section 4.1.1.2, and the target seismic

images to determine the training set. As a result, the cluster containing the seismic images recommended as a training set for DL models for natural gas reserves detection is identified, Figure 4.4.

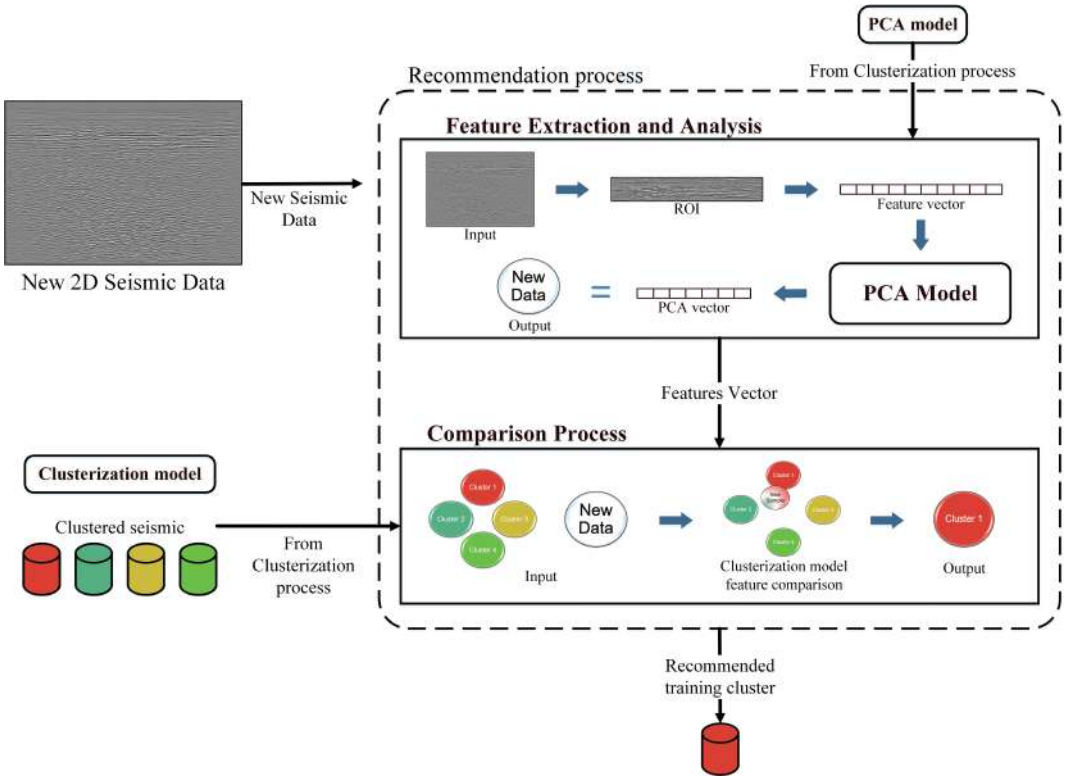


Figure 4.4: Recommendation process pipeline.

For each seismic image, the recommendation process executes the steps separately:

1. Feature Extraction: Techniques described in Section 4.1.1.2 are applied to the target seismic image.

2. PCA Model: The resulting feature vector is subjected to the PCA model, which produces a new feature vector in the same sample space as that used in the clustering process, i.e., This step places the target features in the same representation space as the training features.

3. HDBSCAN Model: Applies the cluster model to the most representative features vector. As a response, the recommendation of the training cluster most similar to the target seismic image is obtained.

   In this step explained at a high level of abstraction, the HDBSCAN model determines the Euclidean distance that exists between the new seismic image and each of the images that make up the clusters (within the same representation space created by the extracted features), then, the cluster with the shortest distance to the new seismic is selected.

In summary, the recommendation process compares features between a target seismic image and each cluster's features. It is important to note that the recommendation is made for each target seismic separately, which implies that different training clusters can be recommended for different target seismic images. The individual recommendation implies that it is necessary to perform new training of the DL models for each recommended cluster.

Finally, up to this section, all the analyses were directed to the seismic features to determine similarities that would allow the grouping of the seismic training images and creating clusters. In this process, there is no ground truth to evaluate the strategy. We can only evaluate the effectiveness of the recommended clusters through the performance of the DL gas inference model.

4.1.3 Classification Process

This section uses the cluster containing the recommended seismic images for training a DL model, with the objective of indicating the location of natural gas reserves in 2D seismic images. Figure 4.5 shows the process carried out for the classification.

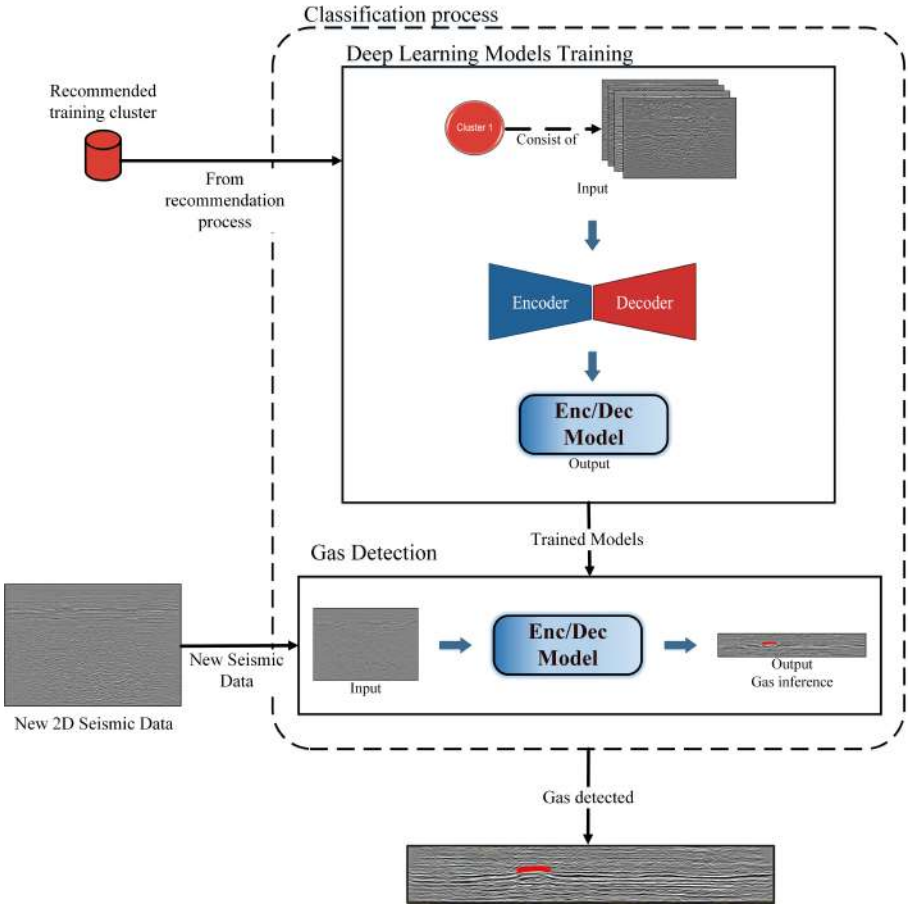


Figure 4.5: Classification process pipeline.

4.1.3.1 Encoder-Decoder

In this step we use the method proposed by Andrade et al. (2021) for gas reservoir indication. This method consists of two steps. In the first one, preprocessing is applied that transforms original seismic images into a readable dataset for the DL model, performing ROI extraction, class balancing, and data normalization. The second step uses a DL model based on a recurring neural network known as LSTM (Section 2.3.1), to carry out a segmentation that allows obtaining a binary result with gas and non-gas classes.

This method receives as input data a set of seismic images in segy format, separated into three different folders, representing the training, validation, and test sets. Additionally, it is necessary to indicate the path where the files that delimit the ROI and the gas labels for each seismic file are stored, in the same way, the training parameters of the network must be included. As a result, the method delivers an image that identifies sections with gas according to specialist marking, sections where the method makes a correct indication, and areas where the method indicates gas but does not match the labels. The results are also delivered about the metrics for each of the seismic images and the complete set.

The LSTM network used is trained considering a time-based early stopping technique, an Adam optimizer with a learning rate of 0.0001, and a weighted categorical cross-entropy-based loss function, the training monitor uses data loss. of validation, measured concerning the convergence metric F1 score (Sammut and Webb, 2017b).

Although the DL model presents an Encoder-Decoder architecture, the convolution cells are replaced by LSTM cells, which better handle temporal sequences by using a memory mechanism. Additionally, a connection layer called Time Distributed is used that resizes the final encoder data to have the output layer size. For more details see Andrade et al. (2021).

4.2 Results

This section presents two different experiments in which the same gas indication method is trained in two ways, firstly with all available seismic images and secondly using a dataset recommended by method 1 proposed in this chapter.

The first experiment uses the same DL model (Section 4.1.3.1) to evaluate two tests that differ from each other only by the training dataset used. These tests are performed separately for each of the nine available fields, which means

that the training datasets change for each new target field.

In the first test, the training dataset consists of all seismic images that do not belong to the target field. This training dataset is labeled as "All except the target". The second test uses the proposed training cluster recommendation to determine the training dataset. This test is named "The Cluster".

The second type of experiment uses only the Belo exploration field. The Belo field is geographically separate from most available fields. Three tests are carried out, using the same DL model and changing the training set. The first test uses all the seismic images available in the eight remaining fields as a training set. The second test uses a training dataset named here as "Expert Cluster", created specifically for the Belo field by an expert. The expert selected the training set without the help of selection tools. Finally, we test the cluster recommendation method proposed here. Note that each new seismic image uses its own recommended cluster.

The performance metrics used are described in Section 2.6, additionally, the size of the training set (Train size) is presented, indicating the number of seismic images that were used in each test.

4.2.1 Seismic Image Database

This chapter presents the seismic image database used as the study object, establishes its origins, general properties, and annexes, and describes limitations that introduce challenges from the perspective of pattern recognition.

4.2.1.1 Background

The database used to test was provided by Eneva, a private energy company in Brazil that operates onshore natural gas. The fields under study come from the Paleozoic Basin of Parnaíba, located in the northeast of Brazil, with more than 600,000 km^2, Figure 4.6.

The basement of this geological formation is a metamorphic and igneous rock, with ages that vary from the Archaic to the Neoproterozoic. Most of these rocks were formed between the Paleoproterozoic and at least the Lower Cambrian, corresponding to the end of the consolidation of the South American Platform (De Miranda et al., 2018).

In the last decades, the investments made by the National Petroleum Agency of Brazil and the concessionaires have elevated the Paraíba Basin to the category of an important onshore basin producing natural gas (Abelha et al., 2018).

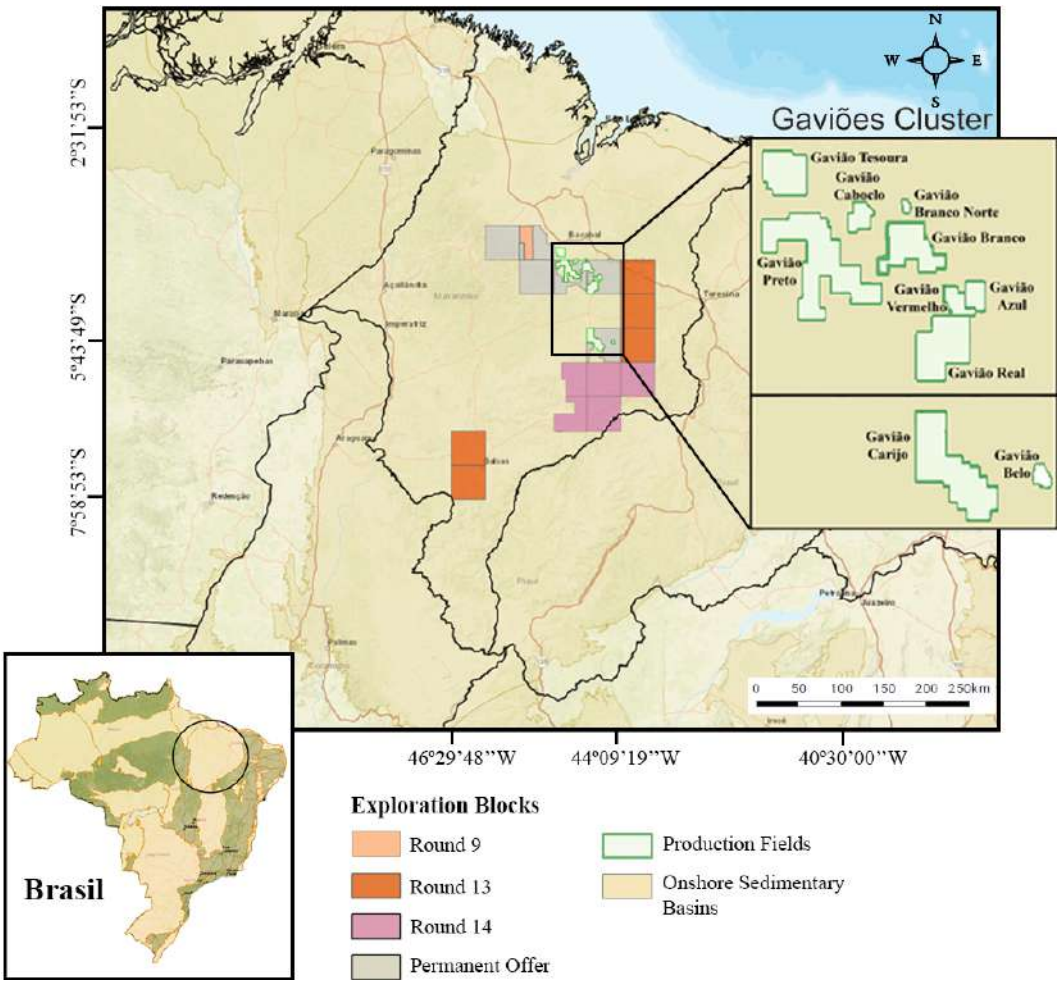


Figure 4.6: Production fields in the Parnaiba Basin (ANP, 2012).

Available seismic images come from nine exploration fields, since these fields are geographically located in the same region, a single seismic image can cross more than one field, which means that seismic files are duplicated in each field that the seismic image crosses.

Seismic images are stored in the segy format, developed by the Society of Exploration Geophysicists, SEG, and contains a 2D matrix that forms seismic amplitude traces as a function of arrival time. Each file has additional information that spatially locates the seismic image within the Parnaíba region and the exploration regions.

In addition, Eneva provides two types of marking label files. The first identifies a region of interest, ROI, within each original seismic image. This mark indicates in which area it is most likely to find natural gas deposits, and therefore it is the region of focus for the studies.

The second marking label refers to the location within the 2D seismic of the gas reservoirs, which is the basis for training the DL gas indication models. However, this marking presents a challenge. Although the quality of

the marking of the location of the natural gas reservoir is supported by studies of wells and interpretations of geoscientists, the regions with no gas marking are regions where it is not certain that analysis for natural gas has been carried out and therefore its content is uncertain. This ambiguity in markup introduces a challenge from the ML perspective. Figure 4.7 presents an example of the labels.

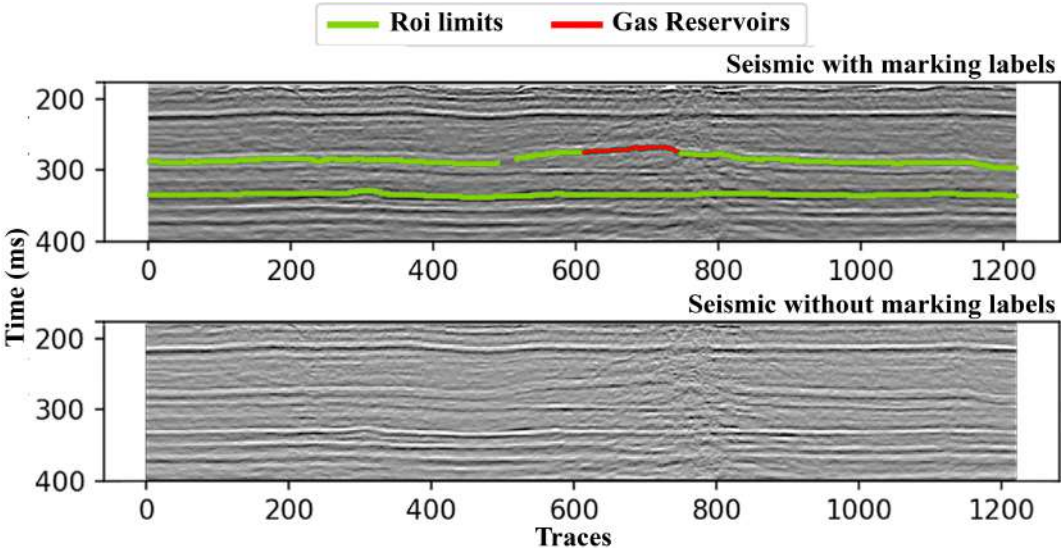


Figure 4.7: Example of ROI and Gas reservoirs marking labels.

Gas label brands present a particular challenge. The gas class has a high level of confidence since they have different analyses and are supported by well drilling that proves the correct marking, however, confidence in the "No Gas" class label decreases as the region becomes away from gas marks. This means that, for the labels of no gas regions, there is no certainty whether there are indeed no gas reservoirs or whether that specific region has not yet been analyzed.

4.2.1.2 General Feature Data

The seismic database comprises 313 seismic images with ROI, but only 168 have gas labels. The seismic images are distributed in the various exploration fields according to Table 4.1. The variation in the total number is because there are seismic images whose extension of land crosses more than one field, creating replicated data. This work will not consider the Gavião Branco Norte field due to the low number of seismic images. Therefore, the study will focus on the remaining nine fields.

The data comes from several acquisitions performed by five teams between 2011 and 2020. This variation in time and the teams introduces

variations in the data features that can be considered domains. In addition, each team did not report the technical equipment model, its configuration, and acquisition parameters, so there was no labeling that identified the different domains within the training data, also, there are particularities caused by different types of terrain. However, all seismic images are encoded following the SEG revision 1 standard.

Seismic images have an average spatial resolution of 15 meters. Nevertheless, given the variability in the acquisition, this resolution may vary within the limits of the SEG standard.

The distribution of the seismic images according to date and team acquisition appears in Table 4.1.

Table 4.1: Seismic images by the team, collected data and field (Not Defined, N/D).

		Gavião									
		Azul	Belo	Branco	Branco Norte	Caboclo	Carijo	Preto	Real	Tesoura	Vermelho
Team	1	7	-	5	-	-	-	6	3	-	6
	2	10	-	63	5	21	4	47	46	21	26
	3	11	10	-	-	6	24	15	-	9	9
	4	-	-	1	1	-	-	-	-	-	-
	5	-	-	-	-	-	10	-	-	-	-
Acquisition Date	2011	9	-	8	-	-	-	1	29	-	7
	2014	7	-	55	6	21	-	50	18	21	24
	2016	-	-	5	-	-	-	-	-	-	7
	2017	9	-	-	-	6	1	16	-	9	-
	2019	1	-	1	-	-	23	1	2	-	1
	2020	-	10	-	-	-	14	-	-	-	-
	N/D	2	-	-	-	-	-	-	-	-	2
Total images		28	10	69	6	27	38	68	49	30	41

4.2.2 First Experiment

This experiment consists of two tests, in the first the comparison baseline is obtained, and in the second the proposed method is applied.

The first test establishes the comparison baseline and is made up of nine different tests, one for each field of exploration. All of them are performed following the same pattern and changing only the target field. Firstly, the seismic images belonging to the field selected as the target are separated, secondly, all the other seismic images from the rest of the fields are taken (paying special attention that none cross the target field), and they are divided into two sets, one with 70% of the images identified as training, and 30% validation. Finally, the gas inference method described in Section 4.1.3.1 is used.

For example, if the target is the Belo exploration field, a total of 10 seismic images are used for the DL model test, and in total, we would have

298 images (after removing the repeats) to separate in training and validation sets.

The second test uses the proposed training dataset recommendation method, in this case, one test is also performed for each field. Firstly, the seismic images belonging to the target field are separated and known as the test set, secondly, the seismic images of the other fields (that do not intersect the target field) are taken and lead to the method proposed in this work, which returns a set of recommended seismic images. Third, the recommended images are separated into two sets, 70% for the training set and 30% for the validation set. Finally, it is used in the gas inference method described in Section 4.1.3.1.

For example, for the Belo field, the method proposed recommends using only 35 seismic images for training the gas inference model, Figure 4.8 shows the distribution of these seismic images in the fields.

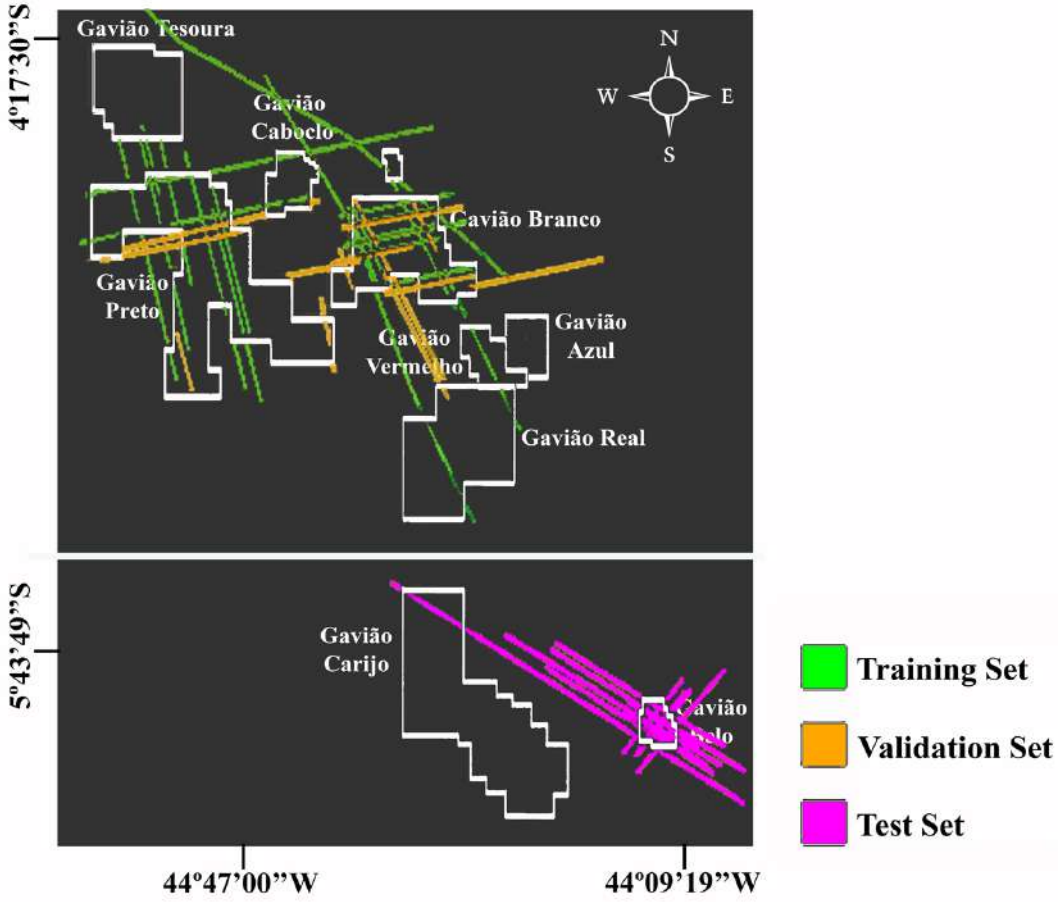


Figure 4.8: Recommended training cluster for Belo exploration field.

Table 4.2 presents the first experiment results, when analyzing the metrics a similar behavior is observed for Azul, Belo, Branco, Real and Tesoura fields. For these, the Accuracy metric does not show the change between the tests, however, a better performance is obtained in the rest of metrics

Table 4.2: First experiment results.

Field Target	Training source	Train size	Accuracy	Precision	Recall	F1_Score	IoU
Gavião Azul	All except target	281	0.99	0.45	0.49	0.43	0.30
	Cluster	116	0.99	**0.46**	**0.52**	**0.45**	**0.32**
Gavião Belo	All except target	298	0.99	0.54	0.45	0.47	0.35
	Cluster	35	0.99	**0.60**	**0.51**	**0.53**	**0.41**
Gavião Branco	All except target	240	0.99	0.45	0.39	0.39	0.26
	Cluster	69	0.99	**0.47**	**0.42**	**0.40**	**0.27**
Gavião Caboclo	All except target	282	0.99	**0.30**	0.38	0.29	0.21
	Cluster	95	0.99	0.28	**0.43**	0.29	0.21
Gavião Carijo	All except target	271	**0.99**	0.29	0.17	0.21	0.13
	Cluster	76	0.98	**0.31**	**0.22**	**0.24**	**0.16**
Gavião Preto	All except target	241	0.99	0.28	0.22	0.22	0.15
	Cluster	52	0.99	**0.29**	**0.23**	0.22	0.15
Gavião Real	All except target	263	0.99	0.28	0.24	0.23	0.16
	Cluster	19	0.99	**0.36**	**0.26**	**0.26**	**0.17**
Gavião Tesoura	All except target	279	0.99	0.23	0.28	0.24	0.15
	Cluster	37	0.99	**0.28**	**0.39**	**0.31**	**0.21**
Gavião Vermelho	All except target	268	0.99	0.45	0.50	0.42	0.29
	Cluster	59	0.99	**0.49**	0.50	**0.43**	**0.30**

when using the cluster recommendation method, this translates into a better inference of regions with gas reservoirs with fewer false positives and false negatives.

To facilitate the identification of the improvement in the performance of the metrics when using the proposed method, Table 4.3 is presented. The Train Size column shows the percentage of the training set that was used for each field; for example, for Gavião Azul only 41% of all available seismic images were used for training. The other metrics show the difference between the performance obtained when using the proposed method and the baseline, those highlighted in blue represent an improvement, while those highlighted in pink present an equal or worse result.

Table 4.3: Improvement of metrics in relation to results using all available data for training.

Field Target	Train size	Precision	Recall	F1_Score	IoU
Gavião Azul	0.41	0.01	0.03	0.02	0.02
Gavião Belo	0.12	0.05	0.04	0.04	0.05
Gavião Branco	0.29	0.02	0.03	0	0.01
Gavião Caboclo	0.34	-0.02	0.05	0	0
Gavião Carijo	0.28	0.01	0.05	0.03	0.02
Gavião Preto	0.22	0.01	0.01	0	0
Gavião Real	0.07	0.08	0.02	0.02	0.01
Gavião Tesoura	0.13	0.05	0.11	0.07	0.06
Gavião Vermelho	0.22	0.05	0	0.01	0.01

The changes in statistical terms reflect an improvement in the overall performance of the DL model when using the cluster recommendation method, but the largest change is observed in the size of the training dataset, for

example, for the Real field, the first test uses 263 seismic training images, instead, the recommended cluster is made up of 19 images, which means that only 7.22% of the original training images are representative of the Real target field. Likewise, only 41.28% for Azul, 11.74% for Belo, 28.75% for Branco and 13.26% for Tesoura are necessary.

Caboclo field results indicate that by using the cluster recommendation method, a higher percentage of gas reservoir identification is achieved, however, a 2% lower precision is also obtained, which translates into more false positives. Regarding the training dataset size, only 33.69% of images present representative data for DL model learning when the recommended cluster is used.

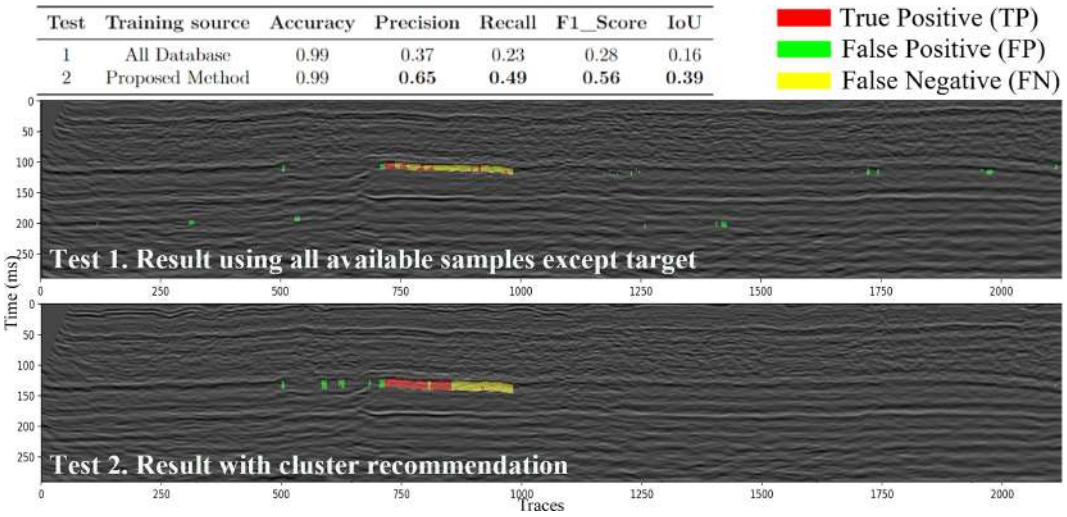
Carijo field shows a 1% loss of Accuracy when using the cluster recommendation method, while at the same time showing better performance on all other metrics. Given the imbalance of the gas and non-gas classes, and that Accuracy is calculated based on all classes, this behavior means that a better identification of the class of interest is obtained, in this case, the gas class, and a lower performance to identify no gas class. In general, the results show a better performance of the DL model when the recommended cluster is used, also only 28.04% of seismic images present relevant information for training.

Tests in the Preto field show an improvement in Recall and Precision when using the recommended training cluster, which means a better performance of the DL model to identify the regions with gas reservoirs and with fewer false positives, in the same way, only 21.6% of training images present relevant data.

Finally, for the Vermelho field, tests show that when using the cluster recommendation method, only 22.01% of images contain relevant information for DL model training, in the same way, the results show fewer false positives, which means a better performance of the model.

Figure 4.9 shows the first experiment result for a seismic image belonging to the Azul field, which shows an improvement in the identification of regions with a gas reservoir and a reduction in false negatives.

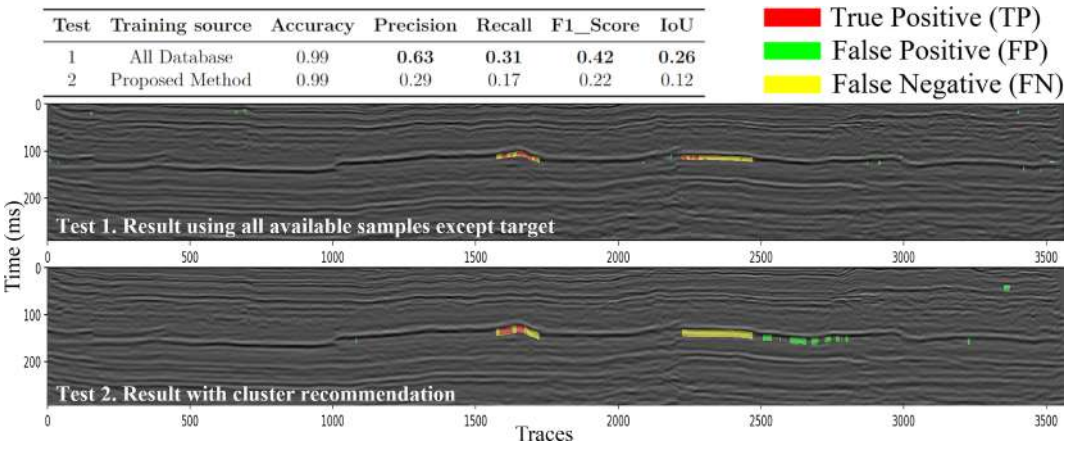
Figure 4.10 shows the result for a different seismic image belonging to the same field, this shows a deterioration in the identification of regions with gas reservoirs and an increase in false positives and false negatives, this result is explained because it not all the seismic images within the Azul field share the same cluster recommended by the proposed method. The first experiment uses a single recommended cluster for a set of target seismic images, which has two implications, the first is that the recommended cluster is selected based on the number of images in the target set for which the cluster is most representative.



Test	Training source	Accuracy	Precision	Recall	F1_Score	IoU
1	All Database	0.99	0.37	0.23	0.28	0.16
2	Proposed Method	0.99	**0.65**	**0.49**	**0.56**	**0.39**

Figure 4.9: First Experiment example of the improvement in gas reservoir indication.

For example, if in a set of 10 target seismic images, 8 are represented by cluster 1, it will be selected to evaluate the 10 target seismic images. The second implication is that there are images to be evaluated with a cluster that is not the most suitable for them, as is the case of the 2 seismic images of the previous example.



Test	Training source	Accuracy	Precision	Recall	F1_Score	IoU
1	All Database	0.99	**0.63**	**0.31**	**0.42**	**0.26**
2	Proposed Method	0.99	0.29	0.17	0.22	0.12

Figure 4.10: First Experiment example of deterioration in gas reservoir detection.

When analyzing the first experiment results in general, there is an improvement in the gas reservoir identification with a marked reduction of the training set size. This implies a better generalization of the gas inference model used.

4.2.3 Second Experiment

In this experiment, three tests are conducted at Belo field, the first taking the seismic images from the remaining eight exploratory fields as a training set. In the second test, a training set built by an expert (without the help of a recommendation tool and identified as Expert Cluster) is used, taking as a selection criterion the similarity with Belo field. The last test uses the proposed method in this work. This test differs from the one performed in the first experiment in that each recommended cluster is used and not the most recommended, this means that for each seismic image from the Belo field, a DL model is trained using the recommended cluster for each image, unlike the first experiment where a unique DL model was trained with a single selected cluster to evaluate all the seismic images from the Belo field.

Results presented in Table 4.4 show that by using a specific cluster for each seismic image, an appreciable improvement is obtained in the metrics, even higher than those achieved in the first experiment concerning gas reservoir identification. In the same way, the results using the Expert cluster demonstrate the high level of difficulty involved in creating an appropriate training set, since the results obtained do not equate with the metrics achieved in the remaining tests.

Note that the "Train Size" column is not included in Table 4.4, this is because the training set recommendation made by the proposed "Cluster" method is made for each of the target seismic images separately, which implies that there is no single number of training images that can be compared with the other tests.

Table 4.4: Second experiment results.

Field Target	Training source	Accuracy	Precision	Recall	F1_Score	IoU
Gavião Belo	All except target	0.99	0.54	0.45	0.47	0.35
	Expert Cluster	0.99	0.45	0.29	0.34	0.22
	Cluster	0.99	**0.58**	**0.55**	**0.54**	**0.43**

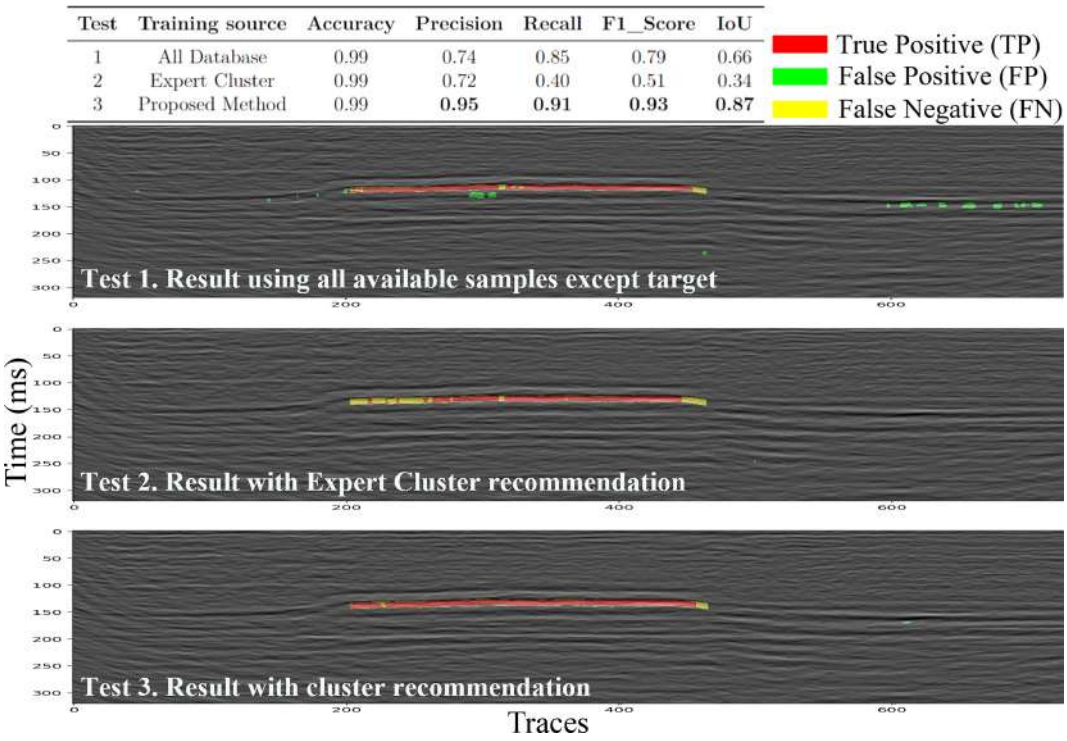
To facilitate the identification of the improvement in the performance of the metrics, Table 4.5 is presented. The metrics show the difference between the performance obtained when using either the dataset proposed by the specialist or the proposed method and the baseline, those highlighted in blue represent an improvement, while those highlighted in pink present an equal or worse result.

Figure 4.11 shows the result of the second experiment for a single seismic image, which shows a significant improvement in the region with gas reservoir

Table 4.5: Improvement of Belo field metrics in relation to results using all available data for training.

Field Target	Training source	Precision	Recall	F1_Score	IoU
Gavião Belo	Expert Cluster	-0.08	-0.16	-0.13	-0.12
	Cluster	0.02	0.09	0.06	0.06

identification and a reduction in false positives and false negatives when using the training cluster recommended by the proposed method.



Test	Training source	Accuracy	Precision	Recall	F1_Score	IoU
1	All Database	0.99	0.74	0.85	0.79	0.66
2	Expert Cluster	0.99	0.72	0.40	0.51	0.34
3	Proposed Method	0.99	**0.95**	**0.91**	**0.93**	**0.87**

True Positive (TP)
False Positive (FP)
False Negative (FN)

Figure 4.11: Second experiment example of the improvement in gas reservoir indication.

Figure 4.12 shows the result for a second seismic image, which shows a significant improvement in gas reserves identification, but with an increase in false positives, when using the proposed cluster recommendation method, this result presents the least favorable case obtained, given the deterioration in precision, this means that the increase in true positives constitutes an important improvement in the DL model, however, the increase in false positives constitutes a problem since when observing the Figure 4.12 it is interpreted that there are two regions with gas reservoirs.

Finally, Figure 4.13 shows the result for a third seismic image, this is of special interest since it demonstrates the limitations of the proposed method, where none of the tests performed achieves a favorable result. Even when the metrics show an improvement when using the proposed cluster recommendation method, the improvement is not significant, and in the same

Test	Training source	Accuracy	Precision	Recall	F1_Score	IoU
1	All Database	0.99	0.36	0.62	0.45	0.29
2	Expert Cluster	0.99	**0.49**	0.48	0.48	0.32
3	Proposed Method	0.99	0.35	**0.85**	**0.50**	**0.33**

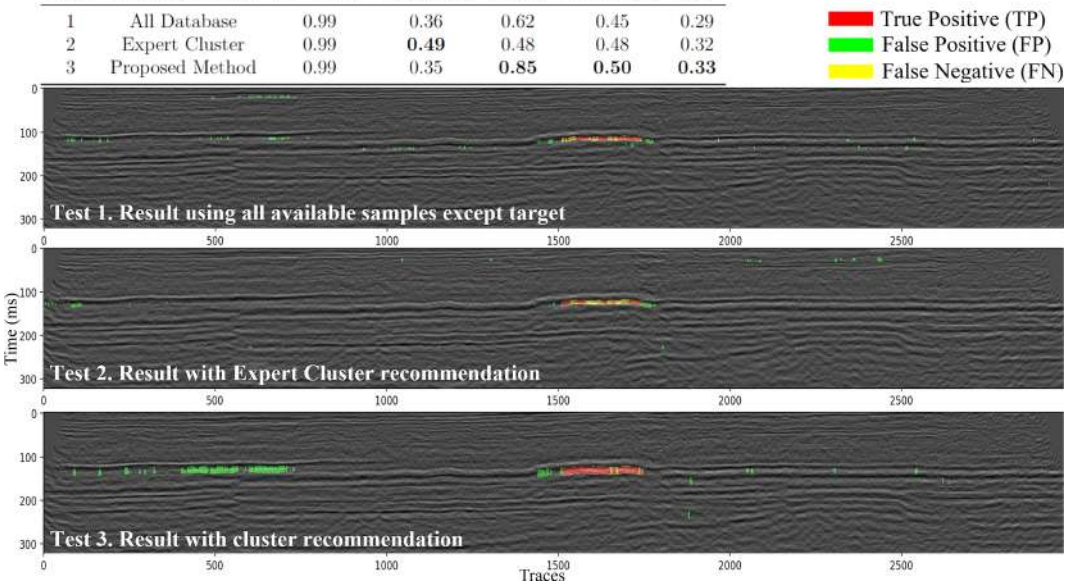


Figure 4.12: Second experiment example of the improvement in gas reservoir indication and an increase in false positives.

way, it does not identify gas reserves.

Test	Training source	Accuracy	Precision	Recall	F1_Score	IoU
1	All Database	0.99	0.03	0.01	0.01	0.01
2	Expert Cluster	0.99	0.00	0.00	0.00	0.00
3	Proposed Method	0.99	0.07	0.02	0.03	0.02

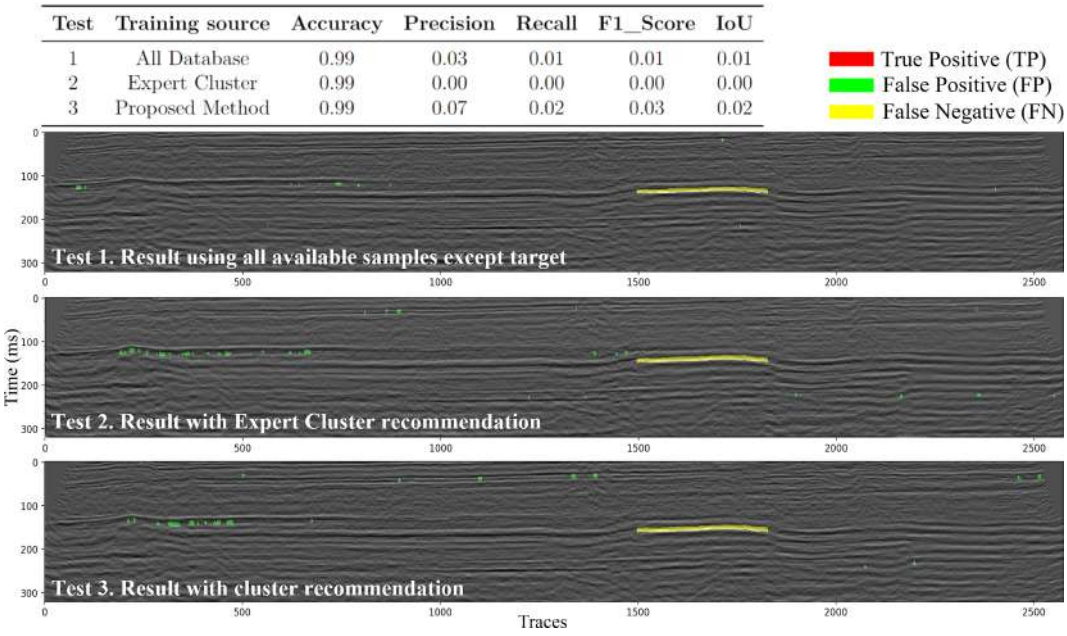


Figure 4.13: Second experiment example without significant changes.

Overall, the results of the second experiment show the utility of using ML-based techniques in training data analysis and gas reservoir inference. Similarly, the method proposed in this paper presents a better generalization of the DL model, compared to the use of all available data for training.

4.3 Discussion

This section presents discussions about the results obtained, analyzing their implications at a high level of abstraction. Additionally, important aspects of the proposed method are presented.

4.3.1 First Experiment Discussion

By analyzing all the results, it is found that the use of the cluster recommendation method shows an improvement of 3% in precision, 4% in Recall and 2% in F1 and IoU, requiring only 23% of the original training data, compared to results using all available training data. These results show from the seismic point of view, that there are images that are more representative and that contribute more to learning, when they are accompanied by others with similar features, in the same way, it is understood that it is possible to create clusters of images based on its seismic features. From the ML point of view, it is clear that seismic representatives and their affinity with the target seismic images are more important for learning than quantity, which also means that when using seismic images with different features in the same training set, they make learning difficult.

The use of a single recommended cluster to train a single DL model, which is used on all seismic images belonging to the same exploration field, presents an overall improvement in performance metrics for the entire exploration field, however, it does not ensure that each seismic image presents a better performance, since each one can have seismic features that coincide with a different cluster.

Cluster conformation indicates that seismic images that are geographically located in the same exploration field do not necessarily share features, which implies that geographic location is not an adequate criterion for training set selection, and at the same time, that two continuous seismic images may or may not share features. This analysis can be clearly evidenced in Figure 4.8, where the target field is Belo, and the seismic images selected to carry out the model training come from multiple geographically separated fields, and where it is observed that no image from the Carijo field was selected to be part of training cluster, despite these closer physically.

4.3.2 Second Experiment Discussion

When analyzing the results, it is found that the use of the cluster recommendation method for each seismic image achieves an improvement of 4% in precision, 10% in Recall, 7% in F1, and 8% in IoU, however, this implies the creation of multiple DL models, each one trained with a different and specific cluster for each seismic target image.

Although using a specific cluster for each target image involves multiple training sessions, the results obtained exceed those obtained when using the cluster recommendation by the exploration field. Almost all seismic images processed with this approach show improvements in gas and non-gas class identification, reducing both false negatives and false positives.

The seismic images with which it was not possible to obtain improvements in the performance of the DL model are seismic images that obtain poor performance in each experiment including the baseline, which may indicate that for the type of features that these seismic images have there are not enough representative data.

4.3.3 General Analysis of the Results

Seismic image comparison allows the clustering of those that present similar features, for this process no labeled data is required since the extraction of features can be carried out using unsupervised ML techniques. Clusters, made up of seismic images with similar features, represent a training base for DL models that obtain superior performance in almost all metrics when it is used in images that share or are located close in the features space, with those of training clusters.

Concentration of the original training database in multiple clusters has several consequences, the first is the need to create as many models as clusters, the second is a marked reduction in the training times necessary for the DL models, and consequently a reduction in the necessary training data. Finally, given the developed method, it is possible to identify those seismic images that do not fit into any cluster which means that the method provides the ability to identify when there is training data available to find gas reservoirs in new images, and when seismic images that share features with the new images are not available, in other words, the proposed method identifies if appropriate training data is available to be used in new data.

Experiments demonstrate that results in gas reservoir indication improve when training image features are similar to new explorations data, although

the amount of this data is markedly less than the available data, which implies that the unused data introduces into the DL model the ability to recognize gas reservoirs in seismic images that have different features but make it difficult to recognize in the target images.

Finally, the generalization of the DL model is improved by introducing the cluster recommendation method, which although it does not affect DL model architecture, provides a way to improve its performance by creating model replicas, but trained with different clusters, highlighting the fact that recommendation basis is feature similarity.

4.3.4 Important Aspects of the Proposed Method 1

The proposed method 1 includes various ML and DL techniques that allow pre processing, feature extraction, feature analysis, data clustering, and inference of gas reservoirs for 2D seismic images.

From a high level of abstraction, in this work two different tasks are carried out, the first performs the clustering of the seismic images based on their features, this task focuses on creating sets made up of images that are considered similar and therefore are more representative for any new seismic sample that fit into the set. The second task refers to the classification of each point contained in a 2D seismic image, to determine whether or not it belongs to the gas class, this task is what makes the gas inference.

The following advantages are highlighted:

1. This work is an approach that allows a better generalization of the DL models for the detection of natural gas in 2D seismic images based on the recommendation of the training cluster, and that also does not require altering the architecture of the models or the original data.

2. When carrying out the first experiment, the advantages offered by the use of the training data recommendation method are demonstrated, which establishes a basis for the seismic images comparison, which is essential for the DL model generalization and the natural gas improvement detection.

3. A combination of unsupervised ML techniques used for seismic image representation is an advantage offered by the cluster recommendation process since it allows obtaining a representation from the same seismic image even when it is not labeled. In other words, the proposed method does not require labeled seismic image for the creation of recommended

training clusters, since as the results of the second experiment show, grouping seismic images by their features is a challenging task.

4. The proposed method makes it possible to improve the generalizability performance of DL models for the detection of natural gas without modifying the model architecture and using the available seismic images. This is made possible by a subspecialization of the model by using a training data set specifically selected for the target seismic images, which helps maintain or improve model performance, which is a great challenge.

5. By using the cluster-per-field recommendation method, it is possible to gain the advantage of significantly reducing the size of the training set required for the DL model, while maintaining and even improving the performance of metrics for the entire target field.

6. By using the cluster recommendation method for each seismic image, the learning of the DL model focuses on those features that are representative of the target seismic image, this for almost all seismic images leads to better gas detection and a reduction in false positives and false negatives.

7. Finally, the combination of all the techniques for the creation of the proposed method allows better detection of gas reserves, as well as a better generalization of the DL model. According to the results of the analysis of the state of the art, the present work is the first to use this combination of ML and DL techniques, applied in the processing of 2D seismic images.

The proposed method has limitations, the following are highlighted:

1. The DL model generalization performance using the proposed method completely depends on the available training seismic images, therefore, generalization performance depends on the existence of images with features comparable to those of the new target seismic images. To overcome this limitation, it is possible to consider Domain Adaptation techniques that seek to modify the input data to fit different domains.

2. The use of different techniques for feature extraction, and clustering also implies hyperparameter tuning for each of them, this limitation can be overcome by using hyperparameter optimization methods or by exploring alternatives for feature extraction in seismic images, however, it will be explored in future research.

3. Each seismic image contains within it data representing large tracts of terrain and depth, however, feature extraction considers the image as a single specimen, when in fact a subdivision of it could provide a more detailed comparison of features.

4. The proposed method focuses only on the training dataset recommendation and does not directly interfere with the DL gas inference model, which must be manually fitted to the training data. To address this limitation, a hyperparameter optimization approach can be used, although it is outside the scope of this work.

4.3.5 Research implications

This subsection provides information on how the proposed method 1 will have an impact on current research trends in this area.

This work demonstrates that the analysis of the DL model training data has an impact on the generalizability. In the same way, it shows that the search for training data with similar features to that of the new objective seismic images allows obtaining better performance for natural gas inference, without modifying the network architecture or the original data.

From the point of view of ML, this work shows that using all the available training databases in problems with multi-domain unknowns does not guarantee that the model learns all the features of each domain. From this point of view, using a set of images grouped according to their features offers better performance, however, it is necessary to make a comparison with the objective data to determine the appropriate training set.

4.4 Conclusion

Method 1 is proposed for gas reservoir identification in 2D seismic images using DL models, which includes a recommendation for a training cluster. To validate the proposed method, the database provided by O&G Eneva from the Paleozoic Basin of Parnaíba located in the northeast of Brazil is used.

Experiments results show that the training data directly influence the generalizability of the DL models, in the same way, it is found that the seismic images are prone to presenting great variability caused by various factors during collection campaigns. However, it is possible to establish relationships between the seismic image features, which creates a basis for comparison.

Experiments show an improvement of 3% in precision, 4% in Recall, and 2% in F1 and IoU, requiring only 23% of the original training data, when the

proposed method is used to recommend the cluster of training per field and a 4% improvement in accuracy, 10% in Recall, 7% in F1 and 8% in IoU when using the seismic image training cluster recommendation method, compared to the performance of the DL method when using all available data for training. The results obtained show that the proposed method represents a useful tool for gas detection in 2D seismic images.

5
Generalization of the deep learning model of natural gas reserve indication based on feature extraction with Autoencoder and recommendation of training dataset and operational hyper parameters - Method 2

This chapter presents a method that seeks to overcome the main limitations found in the Chapter 4, focusing on improving the generalization performance of DL models for the indication of natural gas reservoirs in 2D seismic images. This method is shown independently of Chapter 4, presenting a new DL-based feature extraction approach. It also treats the problem described in Section 1.3 in a more specialized way as a domain adaptation problem according to Section 2.2.2.

Method 2 separates the task of analyzing the properties of the seismic images to make a comparison between the training data and the target data, from the task of indicating natural gas, that is, a method is proposed that first performs an analysis of the features of the seismic images to recommend a training set and then performs the training of the DL model for the specific task of natural gas indication.

This approach focuses on identifying the seismic images that are most representative of the target images, considering the feature extraction step as an unsupervised training problem, it is possible to tackle it using the unsupervised artificial neural network known as Autoencoder (Jiang et al., 2022; Kong et al., 2021; Xu et al., 2019).

The specific contributions of method 2 to state of the art are: First, it introduces an unsupervised DL feature extraction method that allows to establishment of a basis for comparing seismic training data and creating clusters with concentrated representativeness based on pattern similarity. Second, a new training data recommendation method for natural gas indication DL models is presented. Third, an automated method is presented to recommend operational hyper parameters of natural gas indication DL models. Finally, the proposed method 2 improved the generalization performance of the DL model.

5.1 Proposed Method

This section describes the proposed method, as well as the techniques and models used to improve the DL model generalization performance for gas indication in 2D seismic images, based on the recommendation of both the operational hyper parameter and the dataset used for DL model training.

The recommendation process allows feature extraction to compare the training seismic images (with ground truth) and the new target seismic images (without ground truth) to choose several representative training sets. In addition, the recommendation process determines the operating hyper parameters for the DL model based on the performance of these training sets. This process is crucial as it provides an automatic way to determine training data and hyper parameters that may otherwise need to be selected by the user, depending on user experience, or require multiple tests, leading to a high time cost.

Figure 5.1 provides a high-level description of the proposed method composed of three main processes, explained in Section 5.1.1 to Section 5.1.3.

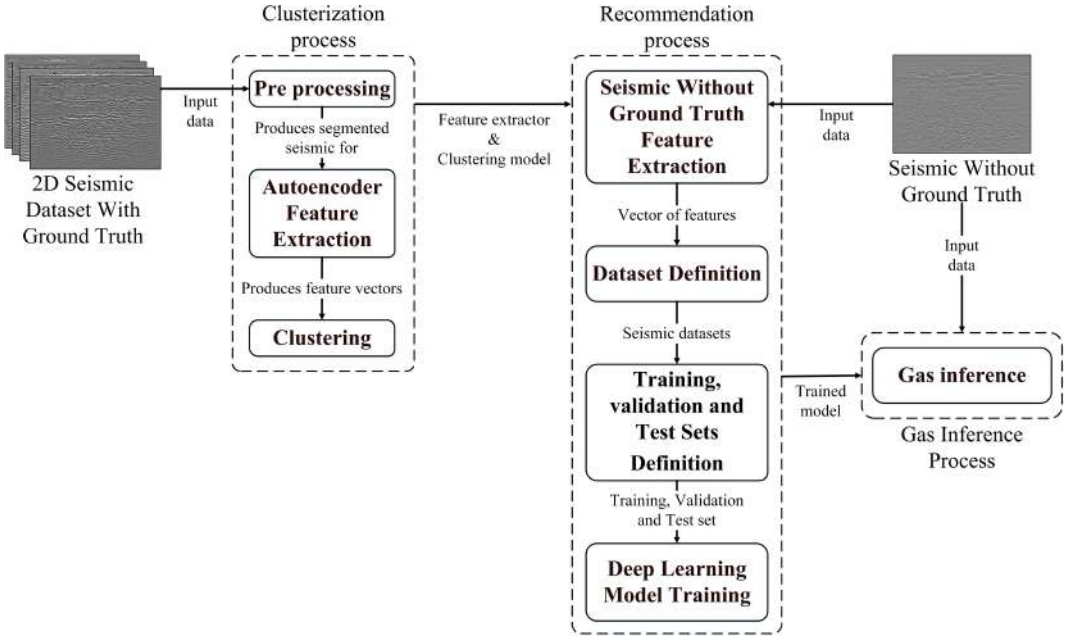


Figure 5.1: Proposed method 2.

5.1.1 Clusterization Process

This process establishes a seismic images comparison base that allows the creation of several clusters with similar features. There are no ground truth

labels indicating which seismic images are similar for this task, so the entire process is executed from an unsupervised approach.

It uses the training set described in Section 4.2.1 as input data and then creates a seismic images clustering model and a encoder feature extraction model.

The clustering process consists of three sequential stages, as shown in Figure 5.2, explained in Section 5.1.1.1 to Section 5.1.1.3.

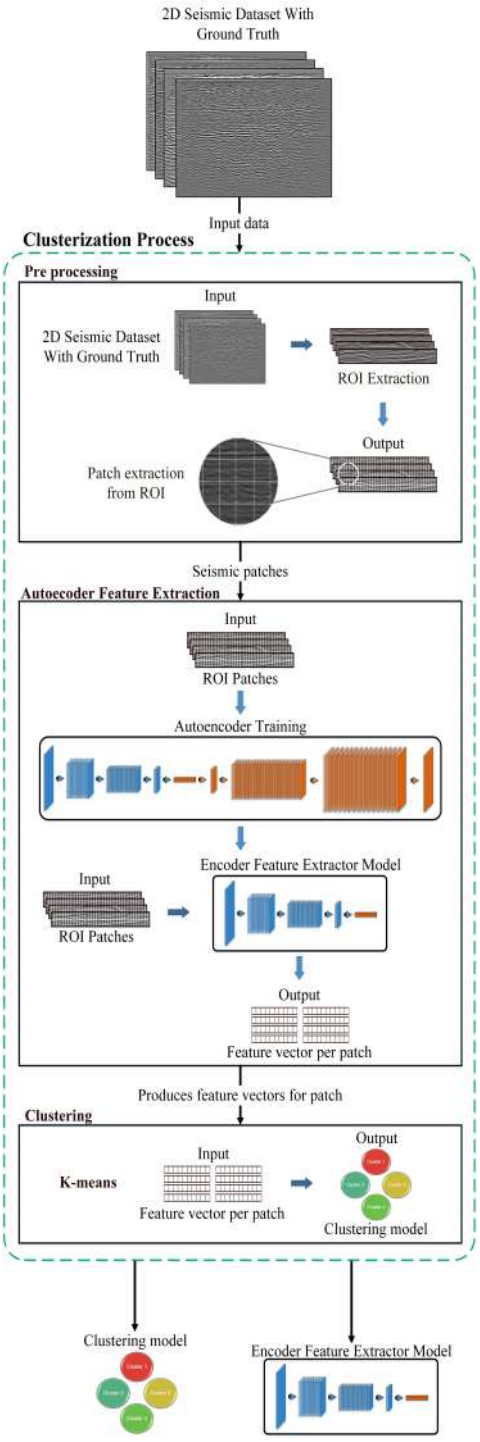


Figure 5.2: Clusterization process pipeline.

5.1.1.1 Pre processing

This step aims to separate the original seismic images to create standard-sized patches, allowing features to be extracted using a DL approach.

The training database consists of 2D seismic images which, when seen on a geographic map, represent lines of seismic information on the ground. The data does not have a standard for the land extension represented. Thus, it is necessary to perform a pre processing to standardize the size of all the seismic images.

Method 2 uses the ROI indication described in Section 4.2.1, to reduce the depth size of each image and focus analyzes on a specific region, causing pattern recognition efforts to be centralized on those representative regions.

Each seismic image has a specific ROI to perform the first pre processing extraction. Only the information of the upper limit is considered, determining the depth component q closest to the surface, according to Equation 5-1.

$$q = \min\left(P(x)\right) \tag{5-1}$$

where, $P(x)$ is the set of the depth component of the points of the upper limit of the ROI.

Then the work area T, which represents the region to be extracted, is formed by the points that are extracted from q plus depth of interest d, this is $(q + d)$, for each trace that forms the seismic image, as shown in Equation 5-2. Depth of interest refers to the size of the ROI, to determine its value, the maximum depth value existing in the entire ROI database is taken as a reference. Thus, at the end of the first pre processing, a sub region of each original seismic image is obtained, with a constant depth for the entire database as shows in Figure 5.3.

$$T = \left[(x_q, y_1), (x_{q+1}, y_2), (x_{q+2}, y_3), \cdots, (x_{q+d}, y_n)\right] \tag{5-2}$$

where n is the number of traces in the seismic image.

The second pre processing performs a tessellation of T without overlapping and with a size adjustment with zero padding, producing a set of patches of defined standard size for all seismic images. The padding of zero value is necessary since the number of traces in each image is nonstandard. This fact implies the existence of spaces at the end of the seismic image filled with traces of zero value.

For tessellation, patches of size, (a, b) are created, this size was defined through experimentation, considering the operational restrictions imposed by the processing load. For a and b, values of 360 and 6, respectively, are

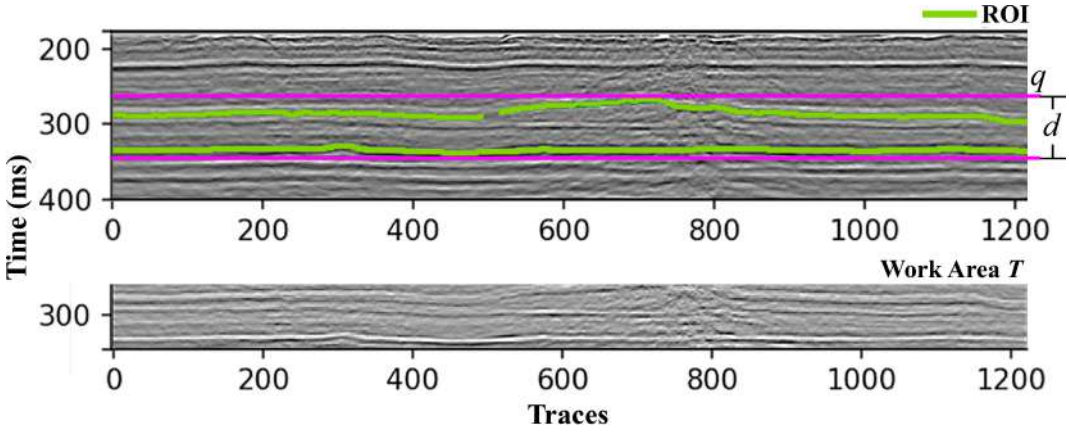


Figure 5.3: Example of extracted work area.

recommended.

5.1.1.2 Autoencoder Feature Extraction

This step aims to create a feature extraction method based on Autoencoder. This method can be used both to perform feature extraction from training image patches and from new target seismic image patches.

This stage builds a feature extraction model based on Autoencoder (Jiang et al., 2022; Kong et al., 2021; Li et al., 2022), which contains two internal models with an Encoder-Decoder structure. Figure 5.4 shows the proposed architecture. The Autoencoder is used because it does not require markup labels to train a model and obtain a lower-dimensional representation of the input data that can be used as features. Not requiring markup labels is a necessary quality for the feature extraction process since the training database described in Section 4.2.1 does not have markup labels that identify each trace domain. In other words, there is no information about what kind of properties each trace of each seismic image possesses.

Autoencoder training uses the seismic patches created in Section 5.1.1.1 as input and ground truth for the loss function. In this context, the Decoder model aims to reconstruct the original input patches from a representation built by the Encoder model. This representation acts as the features, at the end of the training, the Encoder Feature Extraction model is obtained and used for feature extraction, producing a single vector for each patch, which is stored in an array containing all seismic training patch features.

The Autoencoder model training uses an approach with an early stop, with an Adam optimizer and a Mean Square Error, MSE, loss function (Theodoridis, 2020) using one thousand five hundred epochs.

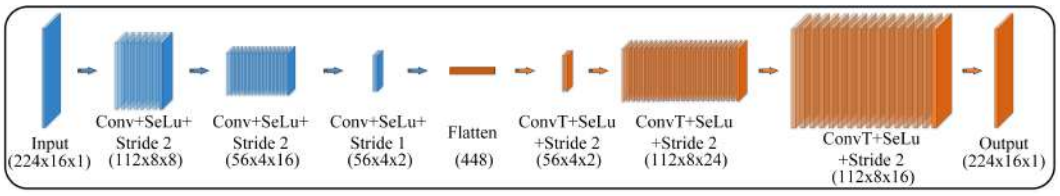


Figure 5.4: Autoencoder (Encoder-Decoder) model architecture summary.

5.1.1.3 Clustering

This step aims to separate seismic image patches into groups based on their similarity. Additionally, it creates a clustering model that can be used to compare target seismic image patches.

This stage uses the features extracted from the patches to train the model based on the K-means technique (Han et al., 2012; Lloyd, 1982), which separates the original training set into patch subsets based on similarity and creates a clustering model to be used on target seismics.

The K-means technique is used given its ability to work with many samples, which is necessary since dividing the original seismic images into patches significantly increases the amount of data that it must group. Although the technique requires reporting the number of desired clusters, it allows a separation that considers small differences between the patches, which otherwise would be classified as a single set.

Selecting the number of groups in which the seismic patches will be separated requires an evaluation of the quality of the clusters created because it is not known how many domains are in the original seismic database. The silhouette coefficient measures the quality of a defined number of clusters c. We selected this technique since its response is within a fixed range which facilitates the comparison and interpretation of the results (Bhandari and Pahwa, 2023; Jin et al., 2022; Leng et al., 2022).

The clustering model tests various numbers of clusters, ranging from three to fifty, to define the appropriate value. Fewer clusters than stated do not offer enough variability for the seismic images. For example, although the patches within a seismic may contain different properties, Using only two clusters to separate the seismic patches would result in an original seismic image composed of patches belonging to both clusters, which makes clustering useless. We tested many clusters empirically to define the upper limit, the results show that fifty is an appropriate limit.

For each value within these limits, K-means separates the entire set of seismic training patches created in Section 5.1.1.2 into c clusters. Then, the

silhouette coefficient of all clusters with a different value of c is compared, selecting the one with the highest silhouette coefficient according to Equation 5-3.

$$c = \max\left(f_{Silhoutte}\left(C_{i=3}\right), f_{Silhoutte}\left(C_{i=4}\right), \cdots, f_{Silhoutte}\left(C_{i=50}\right)\right) \tag{5-3}$$

where C_i is the assignment of seismic patches in i cluster performed by K-means.

At the end of the clustering process, it obtains the grouping of seismic patches in the c cluster. In addition, it produces a trained K-means model, which can be used in target seismic patches to determine which of the c cluster is the nearest.

Although the clustering model uses patches of defined size to perform the grouping, the final clusters comprise the original seismic images. The assignment of each seismic image to a cluster is carried out by analyzing the patches that make up the image and their assignment to each cluster, in this way the cluster to which the most patches have been assigned is selected. This approach preserves the initial seismic images without altering or cropping them.

5.1.2 Recommendation Process

This process aims to recommend both the training set and the operating hyper parameter for a gas inference DL model, based on the similarity of the target seismic image features to the clusters.

This process takes the target seismic images to indicate the region containing the gas reserves (data without ground truth), compares the seismic features to both define the training datasets, and identifies a set of operational hyper parameters for a specific DL model. The recommendation process can be applied individually or to a set of seismic images, and includes four stages, as shown in Figure 5.5, which produce a trained DL model as a response.

This process improves the generalization performance of the DL model without altering its net architecture, working only on the training data and performing a comparative analysis between the features of the target seismic images (without ground truth) and the training seismic database.

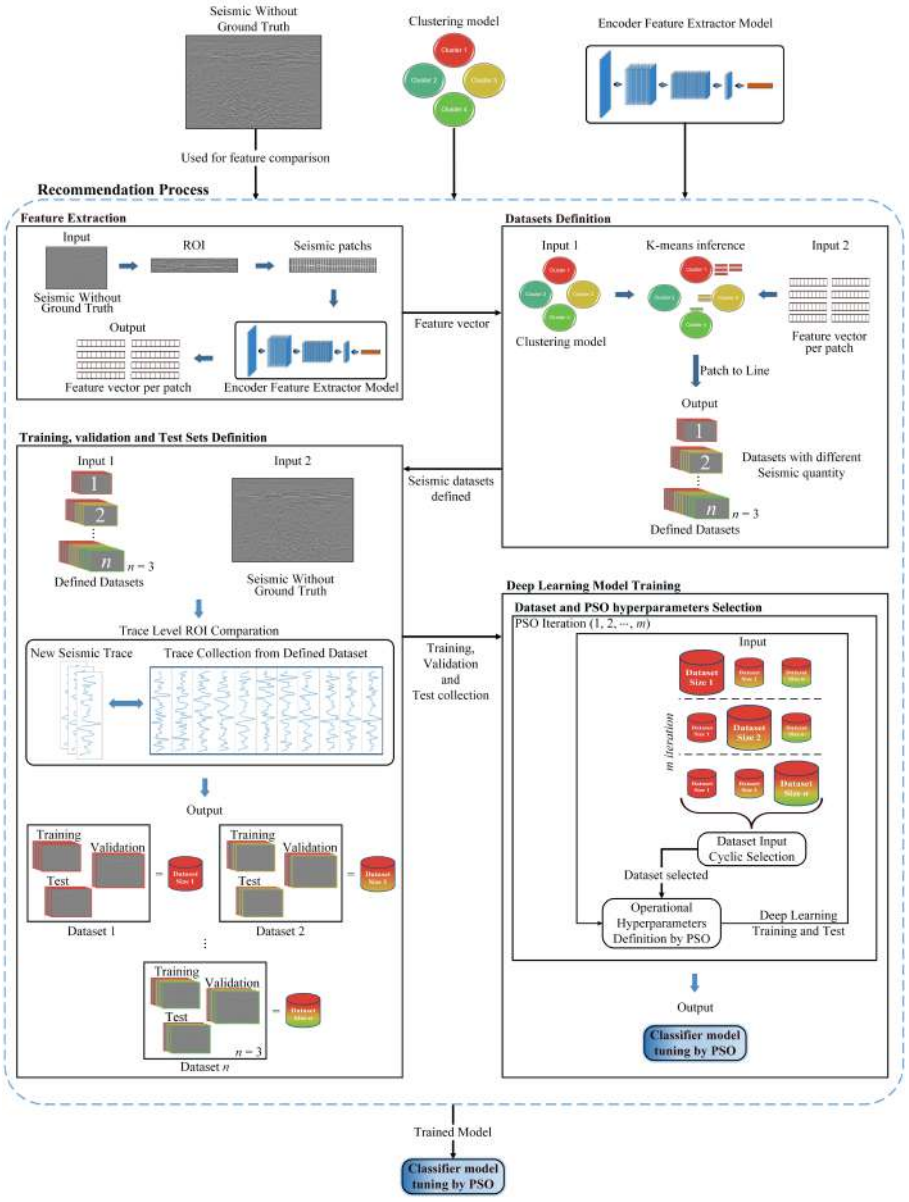


Figure 5.5: Method 2 recommendation process pipeline.

5.1.2.1 Seismic Without Ground Truth Feature extraction

This step aims to represent the target seismic image in the same feature space of the training set to allow a similarity comparison.

The seismic images described in Section 4.2.1 have general properties, such as coming from an exploration based on wave propagation reflection, being 2D data, and using coding based on the Geophysical Exploration Society standard. However, they contain features that depend on the properties of the terrain and the technology and the parameterization used to collect them, so each seismic image has particular features, which in turn allow the comparison and creation of representative clusters for each feature.

In the same way, the target seismic images (without ground truth) contain features that allow them to be analyzed and compared with the clusters. However, these features must be represented in the same space as that used for the training seismic database, so we perform a treatment similar to that described in Section 5.1.1.1. This treatment extracts the region of interest and divides the new seismic image into patches. Then, the Encoder Feature Extractor model created in Section 5.1.1.2 is applied, resulting in a collection of features in the form of a vector that represents each patch of each target seismic in the same representation space as the training seismic database.

5.1.2.2 Datasets Definition

This stage aims to recommend several training sets but with different numbers of seismic images. These images allow obtaining training sets with different concentrations of features that are present in the target seismic image.

This stage creates several training datasets for the target seismic images based on the patch assignment of the clustering model created in Section 5.1.1.3.

The reason for creating multiple training datasets is that the DL model processes the target seismic image individually to indicate the gas reservoir location. Thus, the same DL model processes all traces of a single seismic image. However, it is not common for all traces in the same image to have the same features and have the same recommended cluster. For example, sixty percent of the traces that belong to a target seismic image have features that identify them as part of a C^1 cluster, twenty percent belong to the C^2 and the rest belong to the C^3. However, the region containing the gas reservoir occurs in the traces with features similar to C^2 for the previous case. If we select only the cluster with a greater representation as the training dataset, we will not provide correct representative data for the DL model to learn. On the other hand, if we take the three clusters and form a single training dataset, we will provide representative features that the model does not need and will only hinder the learning process. For this reason, several recommended training datasets are created, each adding new clusters to provide seismic images with more features present in the target seismic image. They allow a more nearly complete representation that may be necessary to indicate the natural gas location.

In this way, each defined dataset contains seismic lines in which a small number of features predominate. From the point of view of machine learning, the model will learn to recognize these features contained in the training

seismic. Each defined dataset includes new features to learn, which may or may not be necessary for the gas reservoir search on the target seismic images.

In addition, determining which or how many of the recommended clusters for new seismic images are relevant would imply having information on where the gas reserves exist, that is, the unavailable marking labels. So, it must create multiple training sets containing one or more recommended clusters.

For a single seismic image, the cluster to which the clustering model has recommended the largest number of patches is taken as the first recommended dataset. The second recommended dataset contains, in addition to the images belonging to the first dataset, those belonging to the following recommended clusters, prioritizing those with the highest number of recommended patches. The number of defined datasets n to be used can extend to the total number of clusters.

In the case of a set of target seismic images, the model chooses the most recommended cluster in common for all target seismic as the first defined dataset, which implies the creation of an assignment ranking for all recommended clusters for each seismic. For the second dataset, the model uses the most recommended clusters and those that follow in the rankings.

5.1.2.3 Training, Validation, and Test Sets Definition

This stage aims to separate each recommended training set into training, validation and test subsets that will be used for the DL model. Paying special attention to making the test subset have features as similar as possible to the target seismic image.

This stage separates the seismic images of each dataset defined in Section 5.1.2.2 into three sets without substitution. This process takes special care to select the test set based on the similarity of the original training data with the target seismic images. This step is necessary to define the sets that will train the gas indication DL model and evaluate its performance to select the operational hyper parameters.

The seismic from each defined dataset is similar to the target seismic images since the clustering model selected them. However, some images have a trace distribution more like the new data. This stage compares them at the trace level so that the test data is as similar as possible to those of the target seismic images. It evaluates the similarity through the Mean Absolute Error, MAE, metric (Sammut and Webb, 2017a), and It selects a number l of seismic images from the defined dataset without replacement to represent each target seismic image, where l is the twenty percent of the dataset. Finally, there are

randomly selects thirty percent of the remaining images in the defined dataset for the validation set and seventy percent for the training set. It performs this process n times independently for each defined dataset.

5.1.2.4 Deep Learning Model Training

This stage aims to determine both the operational hyper parameters and the recommended training set, in addition to creating the gas inference DL model that will be used in the target seismic image.

This stage introduces the operational hyper parameters of a specific DL model to indicate gas reservoirs in 2D seismic images. It also identifies which of the defined datasets contains the most representative features based on the test sets performance metrics. The result of this stage is a trained DL model that can be used on the target seismic images.

The developed proposal operates in a way that does not affect the original architecture of the DL model. For this reason, it is possible to use the proposed method 2 regardless of the selected classifier model. Two different DL networks will demonstrate this property. The first model used was proposed by Andrade et al. (2021) based on an LSTM network. A GRU (Section 2.3.2) replaces the neural network in the second model.

The choice of operational hyper parameters refers to configuration parameters that affect how the DL model works without affecting the architecture. It performs this search because these parameters directly affect the ability of the DL model to learn the specific features of the training data (Amirabadi et al., 2020; Kannammal et al., 2022; Nematzadeh et al., 2022). Method 2 identified five hyper parameters to tune, as shown in Figure 5.6:

1. Gas Pixel Spacing, Gas_{pixel}: Determines the number of pixels added above and below the marking label to consider it a gas region. This enlargement of the original label allows the model to recognize the patterns that occur before and after the gas reserve areas.

2. Region of Interest Pixel Spacing, ROI_{pixel}: Indicates the number of pixels to add before the ROI upper bound. The geoscientist usually marks the upper and lower limit of the region of interest. This mark follows the natural limits given by the rock structures. This work takes the ROI point closest to the surface as the upper limit of ROI, and then ROI_{pixel} is extended towards the surface, allowing the capture of seismic threshold properties delineated by the geoscientist.

3. Size of the region of interest, ROI_{size}: Determines the size of the ROI extracted from all training data. Although each seismic image contains an indication of ROI, it is necessary to define a standard size so that the gas indication DL model can process them.

4. Batch Size: Indicates the number of traces to process in training the gas indication DL model before applying the loss and feedback function to adjust the internal weights of the network.

5. Balance: Determines whether to perform a balancing between the seismic images of each class (Gas, No gas) within the training data set, It is a Boolean type variable.

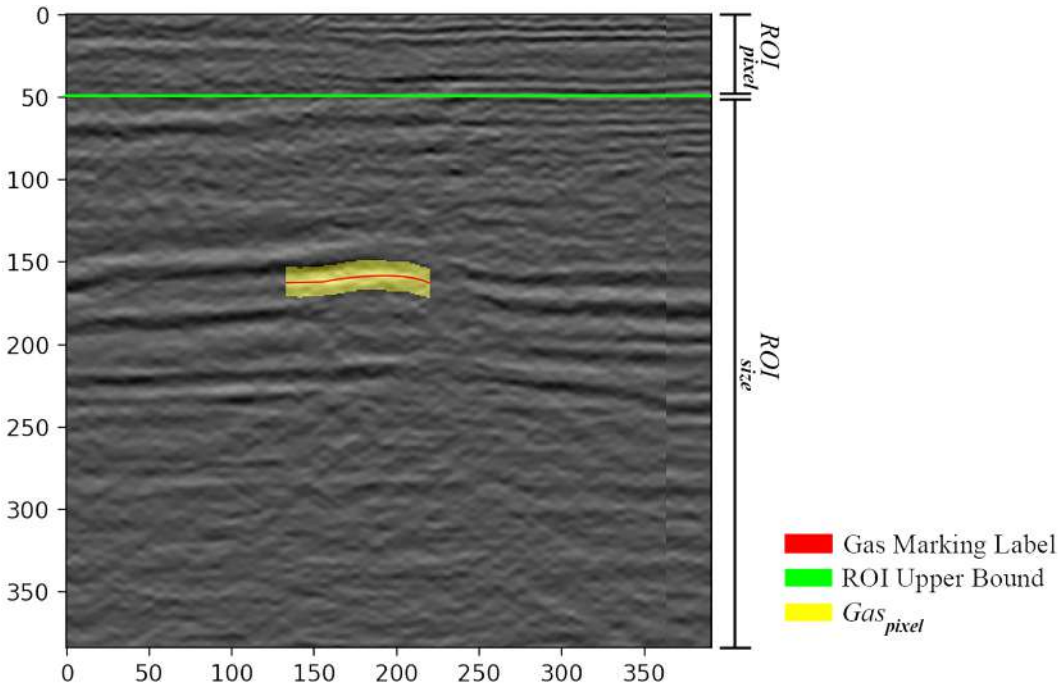


Figure 5.6: Hyper parameters example.

If the search for hyper parameters is manual, the result and the time required will depend on the user's experience, and the complexity of the task increases when considering the multiple possible combinations. For this reason, it uses an optimization technique that allows the search to occur within a multidimensional space with an iterative adjustment, an adjustable learning rate, a defined number of iterations, and a focus on reducing the cost function that allows the production of a reasonable solution. Considering all these characteristics, this work uses Particle Swarm Optimization, PSO, (George et al., 2020; Ma et al., 2022; Muisyo et al., 2022; Shi et al., 2022).

PSO performs a defined number of iterations to test various hyper parameters on the DL model, which implies training and testing the model

for the indication of gas reservoirs. This process uses all the defined datasets in Section 5.1.2.3. The training and validation sets teach to the DL model, and the test set measures performance and provides feedback to the PSO, which updates the hyper parameters before starting a new iteration.

Since there is more than one defined dataset, it is necessary to identify which contains the most representative features for the target seismic based on the test set. Thus, the model cyclically changes the defined dataset used in each iteration of the PSO. When it has tested all defined datasets, it will use the first dataset again in the next iteration of the PSO. This process tests each defined dataset multiple times with different hyper parameters.

When working with DL, three sets (training, validation, and tests) usually perform the study and validate of results (Lei et al., 2023; Lu et al., 2022; Maharjan et al., 2022; Waqas and Ahmed, 2022). The DL model is usable when achieving adequate performance for the objective task. This work defines these three sets based on the similarity with the new target seismic images to improve the generalization. Thus, training a new DL model does not require new marking labels and focuses on identifying the seismic images with the most representative features for the target seismic images in the original training database.

At the end of the iterative PSO process, the method identifies the hyper parameters and the dataset that produces a better indication of the gas reservoir for the test 2D seismic images within the scope of the PSO heuristic technique, which implies that the solution found may not be the best within the entire search space and may be a local minimum. As a result of this stage, it also obtains a specific trained DL model for the target seismic images.

5.1.3
Gas Inference Process

This process uses the DL model trained in Section 5.1.2.4 for each new seismic image and indicates the location of the gas reservoirs. Figure 5.7 presents the pipeline.

5.2
Experiments and Results

This section presents the experiments carried out and their results, following the performance comparison when using a DL model for the indication of natural gas reservoirs in 2D seismic images trained with all the available data versus the results when using the same DL model architecture trained by the proposed recommendation method.

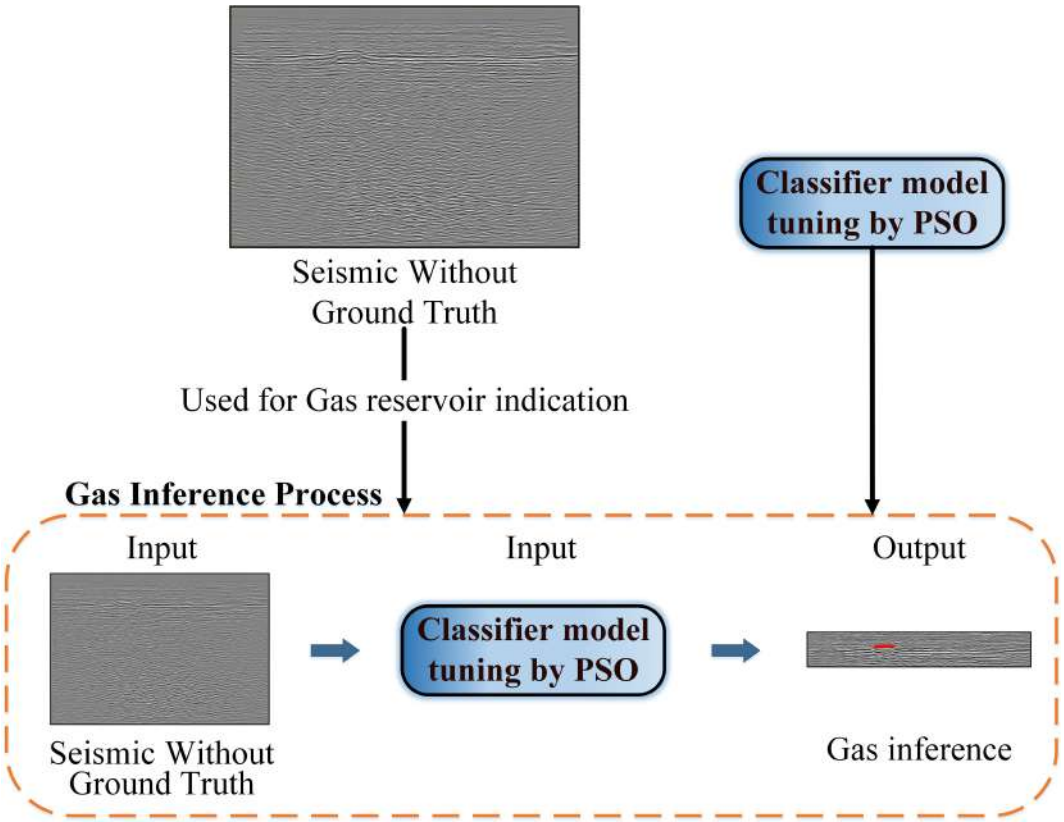


Figure 5.7: Gas indication process pipeline.

The performance metrics used are described in Section 2.6.

5.2.1 First Experiment

This experiment performs tests in two stages using the method proposed by Andrade et al. (2021) based on an LSTM network architecture to indicate natural gas reservoirs in 2D data.

The first stage obtains the comparison baseline for the nine study fields described in Section 4.2.1. This experiment uses all available data to train the gas indication DL model and then uses it for a specific field. For example, if Gavião Belo is the target, we will use all the seismic images that belong to the other fields and that do not intersect with the Gavião Belo field as the training base for the DL model. We will reserve thirty percent for validation and seventy percent for training.

In the second stage, the recommendation method proposed in this work defines the training seismic images and assigns the training, validation, and test sets. Then, it defines the hyper parameters for the gas indication DL model. Figure 5.8 presents an example of a set recommended for the Gavião Belo field.

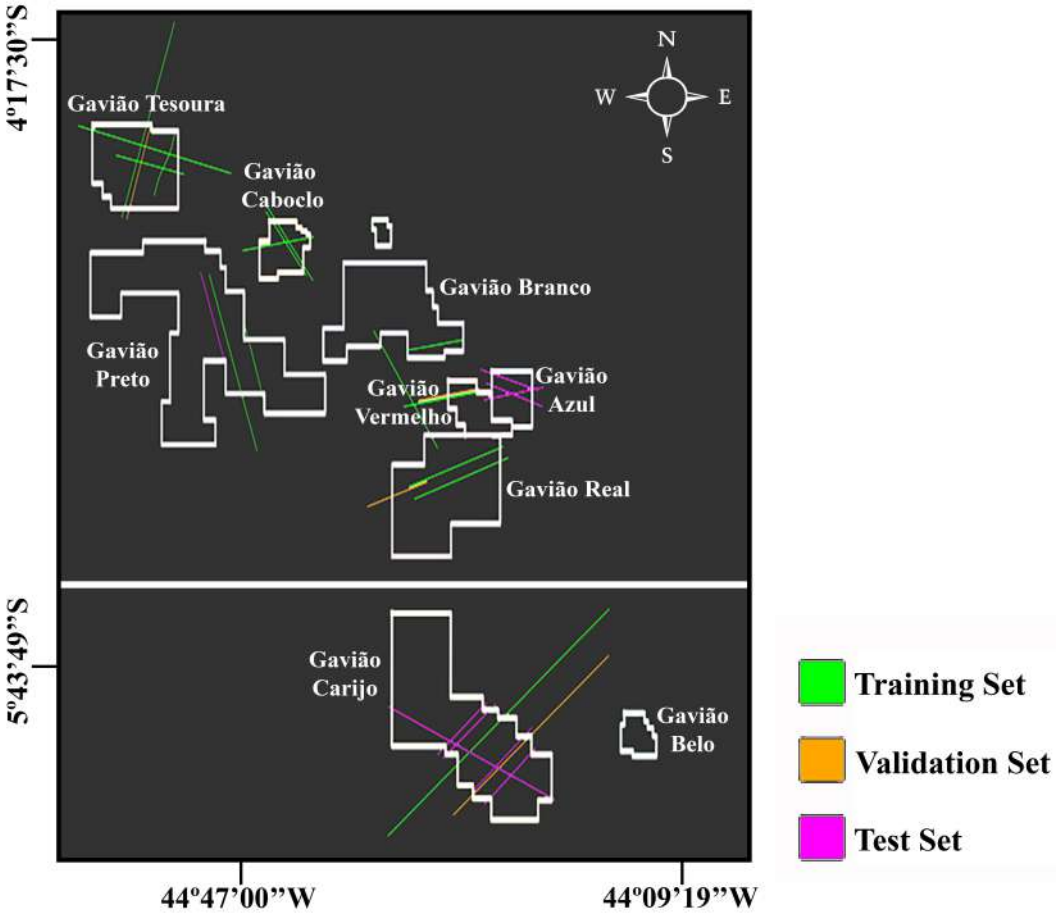


Figure 5.8: Training seismic images recommendation for Gavião Belo field.

Table 5.1: First experiment table results.

Field Target	Training source	Database size	Accuracy	Precision	Recall	F1_Score	IoU
Gavião Azul	All Database	281	0.99	**0.45**	0.49	**0.43**	**0.30**
	Proposed Method	95	0.98	0.32	**0.76**	**0.43**	0.29
Gavião Belo	All Database	298	0.99	0.54	0.45	0.47	0.35
	Proposed Method	29	0.99	**0.56**	**0.63**	**0.56**	**0.44**
Gavião Branco	All Database	240	0.99	**0.45**	0.39	0.39	0.26
	Proposed Method	89	0.99	0.40	**0.72**	**0.49**	**0.35**
Gavião Caboclo	All Database	282	0.99	**0.30**	0.38	**0.29**	**0.21**
	Proposed Method	72	0.99	0.23	**0.50**	**0.29**	**0.21**
Gavião Carijo	All Database	271	0.99	**0.29**	0.17	0.21	0.13
	Proposed Method	63	0.99	0.22	**0.33**	**0.25**	**0.16**
Gavião Preto	All Database	241	0.99	**0.28**	0.22	0.22	0.15
	Proposed Method	105	0.99	0.26	**0.41**	**0.28**	**0.19**
Gavião Real	All Database	263	0.99	**0.28**	0.24	0.23	0.16
	Proposed Method	49	0.99	0.27	**0.49**	**0.30**	**0.20**
Gavião Tesoura	All Database	279	0.99	**0.23**	0.28	**0.24**	**0.15**
	Proposed Method	83	0.99	0.16	**0.57**	0.23	0.14
Gavião Vermelho	All Database	268	0.99	**0.45**	0.50	0.42	0.29
	Proposed Method	79	0.98	0.35	**0.70**	**0.43**	**0.30**

Table 5.1 presents the performance metrics achieved for the first experiment. Also included is the "Database Size" column, which shows the number of seismic images used to train the DL model.

To facilitate the identification of the improvement in the performance

of the metrics when using the proposed method, Table 5.2 is presented. The Train Size column shows the percentage of the training set that was used for each field; for example, for Gavião Azul only 34% of all available seismic images were used for training. The other metrics show the difference between the performance obtained when using the proposed method and the baseline, those highlighted in blue represent an improvement, while those highlighted in pink present an equal or worse result.

Table 5.2: Method 2 first experiment, improvement of metrics in relation to results using all available data for training.

Field Target	Train size	Precision	Recall	F1_Score	IoU
Gavião Azul	0.34	-0.13	0.27	0	-0.01
Gavião Belo	0.1	0.02	0.19	0.09	0.1
Gavião Branco	0.37	-0.05	0.33	0.1	0.09
Gavião Caboclo	0.26	-0.07	0.12	0	0
Gavião Carijo	0.23	-0.07	0.16	0.04	0.03
Gavião Preto	0.44	-0.02	0.19	0.06	0.04
Gavião Real	0.19	-0.01	0.25	0.07	0.04
Gavião Tesoura	0.3	-0.07	0.29	-0.01	-0.01
Gavião Vermelho	0.29	-0.1	0.2	0.01	0.01

The results of the first experiment present two types of behavior when using the proposed method: the first is predominant, obtained by the Gavião Azul, Gavião Belo, Gavião Branco, Gavião Cabloco, Gavião Carijo, Gavião Preto, Gavião Real, and Gavião Vermelho fields, which present an increase in the correct indication of natural gas of 27%, 19%, 33%, 12%, 16%, 19%, 25%, and 20%, respectively. The metric Precision presents a deterioration in almost all cases. Compared with the baseline, a variation of -13%, 2%, -5%, -7%, -7%, -2%, -1%, and -10% appears for each field. Despite the loss of precision, it obtains a general improvement of the generalization according to the F1 score metric of 0%, 9%, 10%, 0%, 4%, 6%, 7%, and 1%. The model achieved these results using only 34%, 10%, 37%, 23%, 26%, 19%, and 29% of the training data set.

Only the Gavião Tesoura field presents the second behavior. The results of the proposed method show an increase in the correct indication of natural gas of 29%. However, there is a loss of precision of 7%, which results in a loss in the F1 Score Metric of 1% with a reduction of the training set of 70%.

Generally, the results using the proposed method show an increase in the correct indication of natural gas, leading to a better model DL generalization. The results based on the performance metrics represent the general trend of each field. Three images illustrate the effects of the proposed method on the

seismic images. Figure 5.9 shows an example of the ideal case in which an improvement occurs in all metrics.

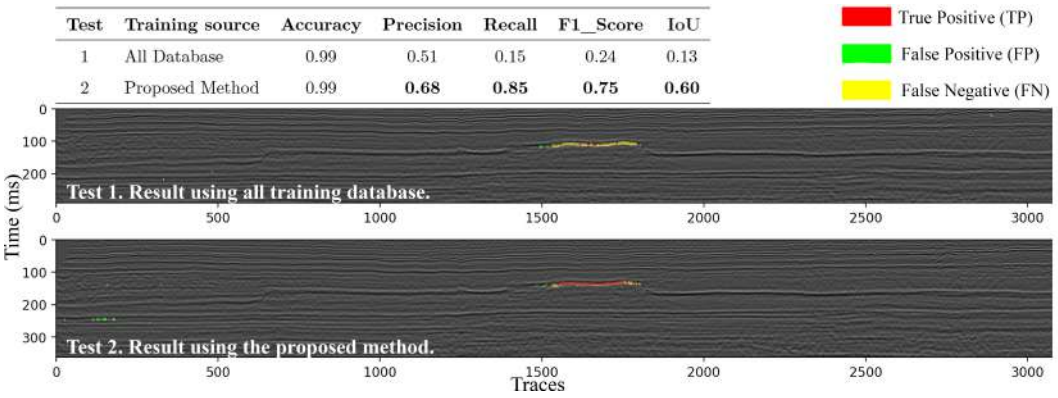


Figure 5.9: First experiment example of improvement in the indication of natural gas.

Figure 5.10 shows the case that obtains an increase in the correct marking of natural gas but with a loss of precision.

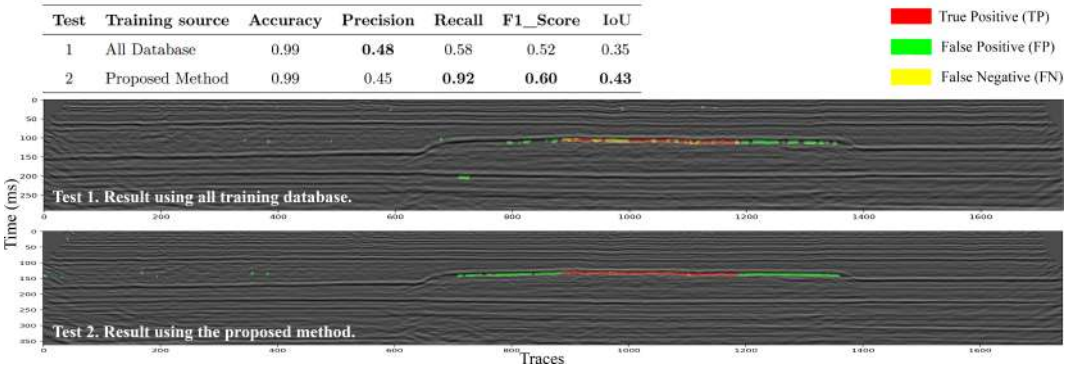


Figure 5.10: First experiment example of improvement in the indication of natural gas and precision deterioration.

Finally, Figure 5.11 shows an important case where representative training seismic imagess do not exist for the target seismic images, so the proposed method can not lead to a better generalization.

Table 5.3 presents the first experiment's recommended operational hyper parameters of the method for each field.

Considering all the fields, when using the proposed method, a percentage variation of the metrics of $-13 \leq Precision \leq 2$, $12 \leq Recall \leq 33$, $-1 \leq F1score \leq 10$, and $-1 \leq IoU \leq 10$. Also, the first experiment's results for all cases show an increase in the correct indication of natural gas. These results demonstrate that the proposed method allows a better generation of the DL model, even when there is a loss of precision. Likewise, it shows a marked reduction of the necessary training seismic images, which indicates a

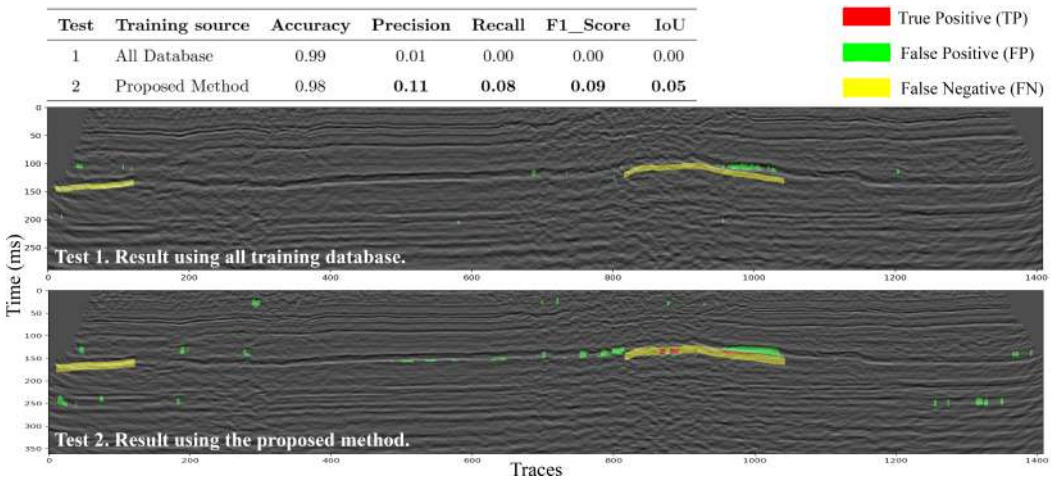


Figure 5.11: First experiment example of no significant improvement in generalization.

Table 5.3: Operational hyper parameters for the DL model recommended for each experiment.

Field Target	Experiment	Gas pixel	ROI pixel	ROI size	Batch size	Balance
Gavião Azul	First	17	12	225	38	True
	Second	17	12	266	162	True
	Third	17	16	264	144	True
Gavião Belo	First	18	12	271	195	True
	Second	15	16	34	201	True
	Third	11	19	272	155	False
Gavião Branco	First	10	14	250	85	True
	Second	12	18	265	83	True
	Third	10	16	247	128	True
Gavião Caboclo	First	16	16	92	274	True
	Second	16	16	83	265	True
	Third	19	15	227	17	True
Gavião Carijo	First	17	11	274	92	False
	Second	18	16	266	162	True
	Third	13	18	271	76	True
Gavião Preto	First	12	14	271	53	True
	Second	15	13	265	83	True
	Third	14	15	267	151	True
Gavião Real	First	19	16	266	39	True
	Second	18	16	266	162	True
	Third	10	12	272	155	True
Gavião Tesoura	First	16	14	250	85	False
	Second	15	15	265	83	False
	Third	15	13	275	110	True
Gavião Vermelho	First	10	18	202	33	False
	Second	12	16	200	32	False
	Third	15	13	275	110	True

correct selection of training seismic images that have the relevant features for the indication of gas in target seismic images.

5.2.2 Second Experiment

The second experiment differs in the training database pre processing, in which it delimits the seismic images to the regions that are within the limits of the exploration fields, that is, for each trace of each training seismic image an analysis of its geographical location is carried out, discarding all those traces that are not contained within any of the nine exploration fields described in Section 4.2.1.2, as an example, Figure 5.12 shows the selected training set for the Gavião Tesoura field.

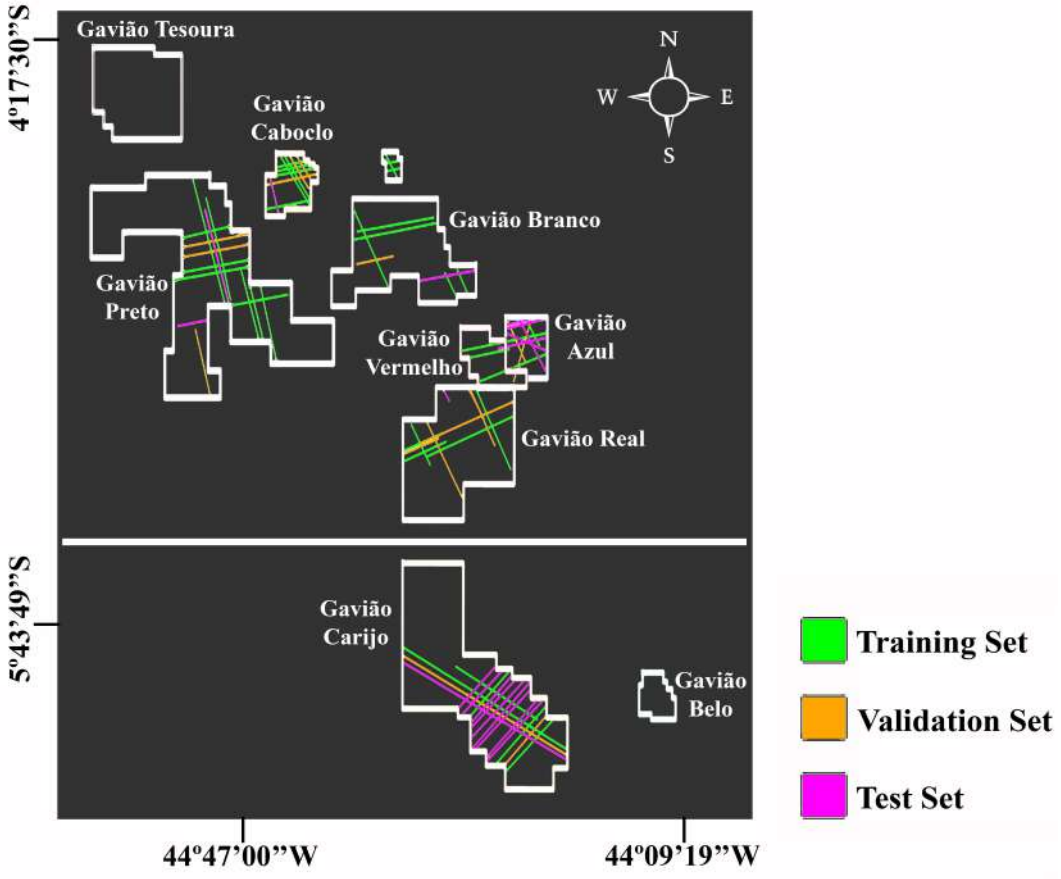


Figure 5.12: Training cut images recommendation for Gavião Tesoura field.

This experiment is carried out to reduce the uncertainty that exists in the labeling of the "No gas" class, since, as mentioned in Section 4.2.1.1, there is only a high level of confidence in the labels that indicate the position of natural gas in the 2D data, however for the other areas it is not guaranteed that there is no natural gas, it is only known that they are areas that have not been analyzed by geoscientists.

The second experiment uses the results obtained in the first stage of the first experiment as a comparison baseline, then the method proposed in method 2 is used to train the method proposed by Andrade et al. (2021) based

on an LSTM for each field separately, using the new training seismic image cut base.

Table 5.4: Second experiment table results.

Field Target	Training source	Accuracy	Precision	Recall	F1_Score	IoU
Gavião Azul	All Database	0.99	**0.45**	0.49	0.43	0.30
	Proposed Method/Split	0.99	0.39	**0.71**	**0.47**	**0.34**
Gavião Belo	All Database	0.99	**0.54**	0.45	**0.47**	**0.35**
	Proposed Method/Split	0.98	0.31	**0.70**	0.41	0.29
Gavião Branco	All Database	0.99	**0.45**	0.39	0.39	0.26
	Proposed Method/Split	0.99	0.38	**0.69**	**0.46**	**0.32**
Gavião Caboclo	All Database	0.99	**0.30**	0.38	**0.29**	**0.21**
	Proposed Method/Split	0.99	0.22	**0.47**	0.28	0.20
Gavião Carijo	All Database	0.99	**0.29**	0.17	0.21	0.13
	Proposed Method/Split	0.98	0.26	**0.50**	**0.31**	**0.21**
Gavião Preto	All Database	0.99	**0.28**	0.22	0.22	0.15
	Proposed Method/Split	0.99	0.23	**0.42**	**0.28**	**0.18**
Gavião Real	All Database	0.99	**0.28**	0.24	0.23	0.16
	Proposed Method/Split	0.99	0.25	**0.52**	**0.30**	**0.19**
Gavião Tesoura	All Database	0.99	**0.23**	0.28	0.24	0.15
	Proposed Method/Split	0.99	0.18	**0.60**	**0.27**	**0.17**
Gavião Vermelho	All Database	0.99	**0.45**	0.50	**0.42**	**0.29**
	Proposed Method/Split	0.98	0.29	**0.64**	0.37	0.25

Table 5.4 presents the results of the second experiment. Since the original training seismic images have been cut, it is not possible to make a comparison between the number of images used for training, so the "Database size" column is not included.

To facilitate the identification of the improvement in the performance of the metrics when using the proposed method, Table 5.5 is presented. In this experiment it is not possible to make a comparison regarding the number of seismic images used for training, because this experiment performs a division of the original seismic images. The metrics show the difference between the performance obtained when using the proposed method and the baseline, those highlighted in blue represent an improvement, while those highlighted in pink present an equal or worse result.

When using the proposed method in the Gavião Azul, Gavião Branco, Gavião Carijo, Gavião Preto, Gavião Real, and Gavião Tesoura fields, an improvement in the correct indication of natural gas according to the marking labels of the geoscientific expert is obtained, with a percentage of 22%, 30%, 33%, 20%, 28%, and 32%, with a deterioration in accuracy that implies an increase in false positives of 6%, 7%, 3%, 5%, 3 %, and 5%, the results overall show an improvement of 4%, 7%, 10%, 6%, 7% and 3% respectively based on the F1 score metric.

The Gavião Belo, Gavião Caboclo and Gavião Vermelho fields present a similar trend in terms of results when using the proposed method, for these fields an improvement in the indication of natural gas of 25%, 9% and 14%

Table 5.5: Method 2 second experiment, improvement of metrics in relation to results using all available data for training.

Field Target	Precision	Recall	F1_Score	IoU
Gavião Azul	-0.06	0.22	0.04	0.04
Gavião Belo	-0.22	0.25	-0.07	-0.06
Gavião Branco	-0.07	0.3	0.07	0.06
Gavião Caboclo	-0.08	0.09	-0.01	-0.01
Gavião Carijo	-0.03	0.33	0.1	0.08
Gavião Preto	-0.05	0.2	0.06	0.03
Gavião Real	-0.03	0.28	0.07	0.03
Gavião Tesoura	-0.05	0.32	0.03	0.02
Gavião Vermelho	-0.16	0.14	-0.05	-0.04

is obtained, with a deterioration of 22%, 8% and 16%, leading to an overall performance loss of 7%, 1% and 5% respectively of the F1 score metric.

For the second experiment, a new recommendation of operational hyper parameters is made for the method based on the LSTM model. Table 5.3 shows the resulting recommendation.

In general, the results of the second experiment show that using the proposed method it is possible to increase the correct indication of natural gas in all cases, however, it also shows that cutting the training seismic images produces a loss of representative traces, which produces a significant loss of precision in three fields.

5.2.3 Third Experiment

This experiment aims to prove that the method proposed in this work can be used independently of the DL model used to indicate natural gas in 2D seismic images. This experiment uses the GRU neural network, which replaces the LSTM used in the first two experiments.

As in the first experiment, it comprises two stages. The first stage obtains the comparison baseline using the entire available training database with the new GRU network for each exploration field. This test ensured that the training set used no seismic images belonging to the target field.

In the second stage, the proposed method defines the hyper parameters for the new DL model that uses the GRU neural network. The recommended

datasets that perform the training for each field are not changed concerning the first experiment (see Section 5.2.1). This occurs because the training dataset is defined regardless of the natural gas indication technique.

Table 5.6: Third experiment table results

Field Target	Training source	Accuracy	Precision	Recall	F1_Score	IoU
Gavião Azul	All Database	0.99	**0.43**	0.52	0.43	0.30
	Proposed Method/GRU	0.99	0.41	**0.73**	**0.48**	**0.35**
Gavião Belo	All Database	0.99	**0.54**	0.46	0.48	0.36
	Proposed Method/GRU	0.99	**0.54**	**0.54**	**0.51**	**0.40**
Gavião Branco	All Database	0.99	**0.46**	0.41	0.40	0.27
	Proposed Method/GRU	0.99	0.40	**0.66**	**0.48**	**0.34**
Gavião Caboclo	All Database	0.99	**0.30**	0.42	**0.32**	**0.23**
	Proposed Method/GRU	0.99	0.26	**0.54**	**0.32**	**0.23**
Gavião Carijo	All Database	0.99	**0.25**	0.14	0.17	0.10
	Proposed Method/GRU	0.99	0.23	**0.41**	**0.27**	**0.17**
Gavião Preto	All Database	0.99	**0.32**	0.19	0.21	0.14
	Proposed Method/GRU	0.99	0.25	**0.42**	**0.28**	**0.19**
Gavião Real	All Database	0.99	**0.30**	0.34	0.28	0.18
	Proposed Method/GRU	0.99	0.27	**0.50**	**0.31**	**0.20**
Gavião Tesoura	All Database	0.99	**0.21**	0.23	0.20	0.12
	Proposed Method/GRU	0.99	0.16	**0.60**	**0.23**	**0.15**
Gavião Vermelho	All Database	0.99	**0.49**	0.50	0.44	0.31
	Proposed Method/GRU	0.99	0.42	**0.70**	**0.47**	**0.34**

Table 5.6 shows the results of the second experiment for all fields. There is an increase in the correct indication of natural gas according to the geoscientist's marking labels, which vary from $8\% \leq Recall \leq 37\%$. However, there is an increment of false positives that varies from $-7\% \leq Precision \leq 0\%$, and these results conclude in a generalization improvement according to the variability of the metric of $1\% \leq F1score \leq 11\%$ and $0\% \leq IoU \leq 7\%$.

Table 5.3 presents the recommended hyper parameters for the third experiment, which uses a GRU network method.

To facilitate the identification of the improvement in the performance of the metrics when using the proposed method, Table 5.7 is presented. In this experiment, the results of the Train Size column do not present changes in relation to the first experiment, since only the gas inference method is changed. The other metrics show the difference between the performance obtained when using the proposed method and the baseline, those highlighted in blue represent an improvement, while those highlighted in pink present an equal or worse result.

In general, the results of the third experiment show a single pattern, in which an improvement in the generalization of the DL model is obtained when using the proposed method and with a reduction in the size of the required training data set.

Table 5.7: Method 2 third experiment, improvement of metrics in relation to results using all available data for training.

Field Target	Train size	Precision	Recall	F1_Score	IoU
Gavião Azul	0.34	-0.02	0.21	0.05	0.05
Gavião Belo	0.1	0	0.08	0.03	0.05
Gavião Branco	0.37	-0.06	0.25	0.09	0.07
Gavião Caboclo	0.26	-0.04	0.12	0.01	0
Gavião Carijo	0.23	-0.02	0.27	0.11	0.07
Gavião Preto	0.44	-0.07	0.23	0.07	0.05
Gavião Real	0.19	-0.03	0.16	0.03	0.03
Gavião Tesoura	0.3	-0.06	0.37	0.03	0.02
Gavião Vermelho	0.29	-0.07	0.2	0.03	0.02

5.3 Discussion

This section presents important aspects of the proposed method and analyzes the implications of the results of each experiment.

5.3.1 First Experiment Discussion

The results of the first experiment show that using the proposed method increases the correct indication of natural gas that agrees with the geoscientist's marking labels. In addition, there is a decrease in precision, which increases the false positives. Even so, for almost all cases, there is an improvement in the generalization of the DL model. The analysis of the precision loss concludes that there are two possible causes. The first indicates traces whose feature differences allow the classification of the "Gas" and "No gas" classes to be too subtle. Therefore, they evade the learning process, indicating that the DL model prioritizes the shape of rock structures. The second cause is related to the uncertainty in the marking labels since, as explained in Section 4.2.1.1, there is no certainty if the zones outside the marking of the "Gas" class contain reserves of natural gas and can be areas where the analysis has not been performed if so, the loss of precision is due to a lack of marking and not an indication error of the DL model.

The Gavião Tesoura field shows the only case in which the proposed method obtains a reduction in the F1 score metric, although it is only 1%. As in the previous fields, it improves the correct indication of natural gas. However, for this field, the loss of precision causes a reduction in the general DL model

performance. When analyzing the causes of this loss, the most likely reason is that the training set contains images with a large amount of unanalyzed terrain. However, for the DL model, these zones are marked as part of the "No gas" class, which may lead to the observed behavior. Figure 5.13 shows the assignment of seismic images that are used as training, in this, it can be seen how there are lines that are considerably far from the fields, sections in which there is no certainty of the presence of gas.

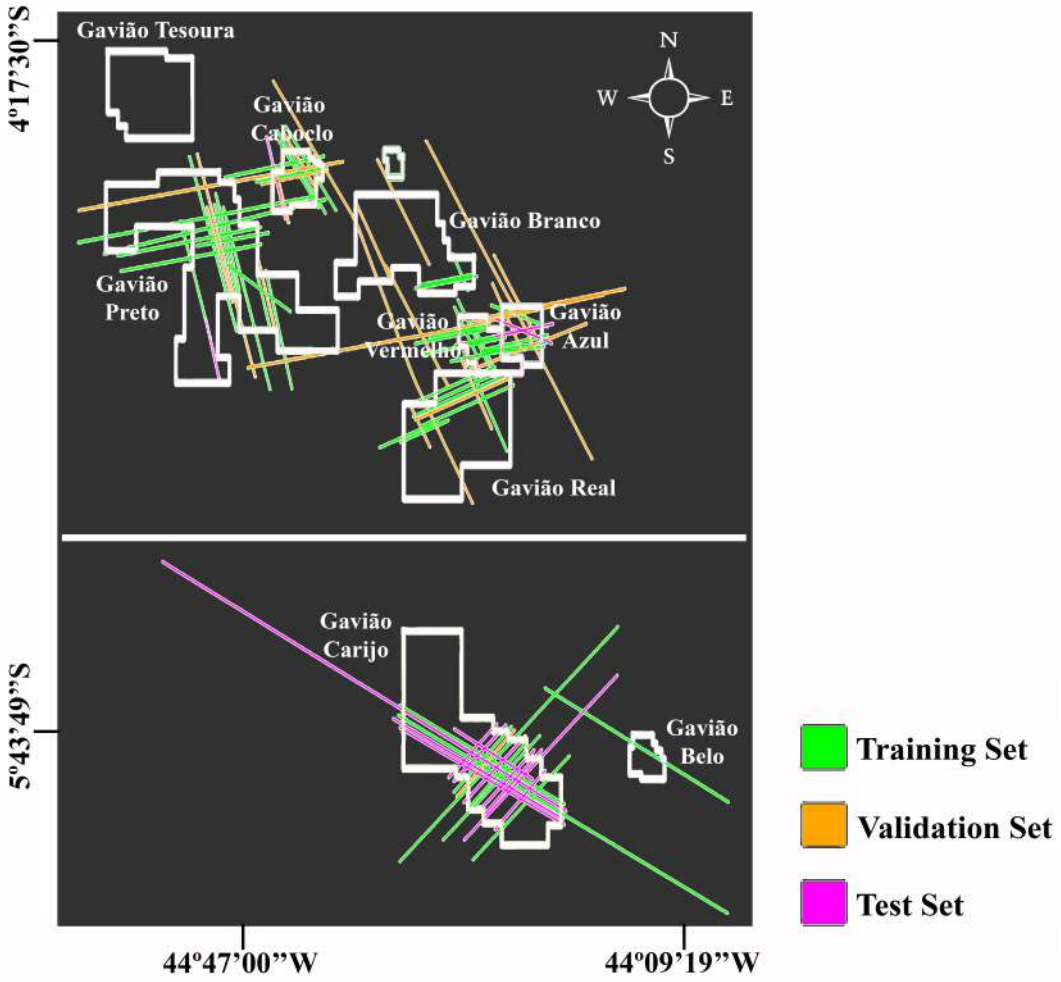


Figure 5.13: Training seismic images recommendation for Gavião Tesoura field.

5.3.2 Second Experiment Discussion

The second experiment shows two types of trends concerning the results. For six of the nine fields, when using the proposed method, a better generalization of the DL model is achieved according to the metrics obtained, in all these fields a significant improvement is achieved in the correct natural gas indication, which coincides with the marking of the specialist. A loss in precision

is also noted, however, this is small compared to the improvement in Recall leading to better overall performance compared to baseline.

The second type of behavior is observed in the Gavião Belo, Gavião Caboclo, and Gavião Vermelho fields, for which using the proposed method causes a deterioration in the generalization of the DL model. When analyzing the cause of this loss, it is found that it is due to the marked deterioration in precision, even though in all fields an increase in the correct indication of natural gas is observed.

One of the possible causes of the loss of precision is because, by cutting the seismic images to preserve only the traces within the fields, the number of samples of the "No gas" class is reduced, since the samples with cut contain few representative data of this class, to exemplify this situation Figure 5.14 is presented, in which a common case with most of the traces belong to the "Gas" class is observed. Overall, these results emphasize the challenge of handling the "No gas" class uncertainty of the training data.

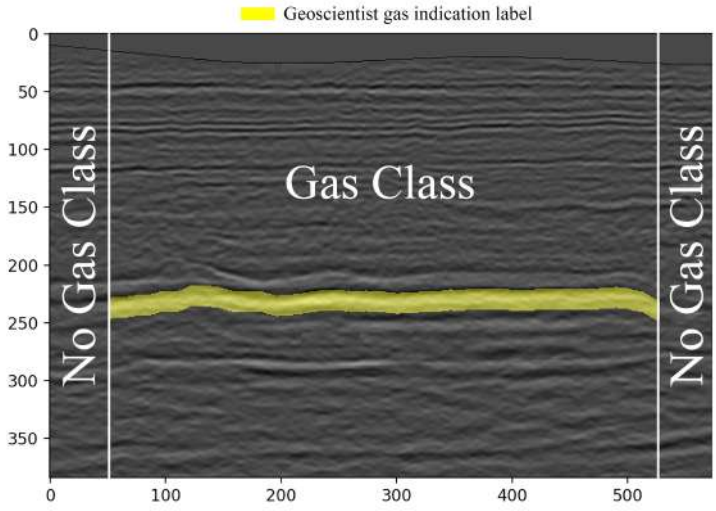


Figure 5.14: Seismic cut seismic images and class example.

In those cases where the training cut seismic images offer enough representation for the DL model learning from both classes, it is observed that the precision increases, Figure 5.15 shows an example of this case.

Figure 5.16 shows the case in which using the original traces amount achieves a better result.

5.3.3 Third Experiment Discussion

This experiment shows a unique type of behavior that demonstrates that using the proposed method obtains a better generalization performance of the DL model with GRU neural network for the indication of natural gas.

Test	Training source	Accuracy	Precision	Recall	F1_Score	IoU
1	Original seismics	0.99	0.13	0.97	0.23	0.13
2	Cut seismics	0.99	**0.29**	**0.98**	**0.44**	**0.29**

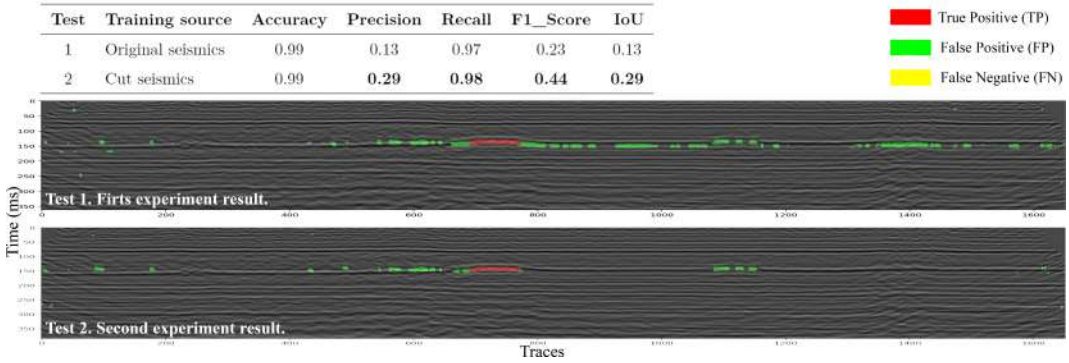


Figure 5.15: Example of improvement when using the trace cut in seismic.

Test	Training source	Accuracy	Precision	Recall	F1_Score	IoU
1	Original seismics	**0.99**	**0.87**	**0.82**	**0.84**	**0.73**
2	Cut seismics	0.98	0.39	0.89	0.55	0.37

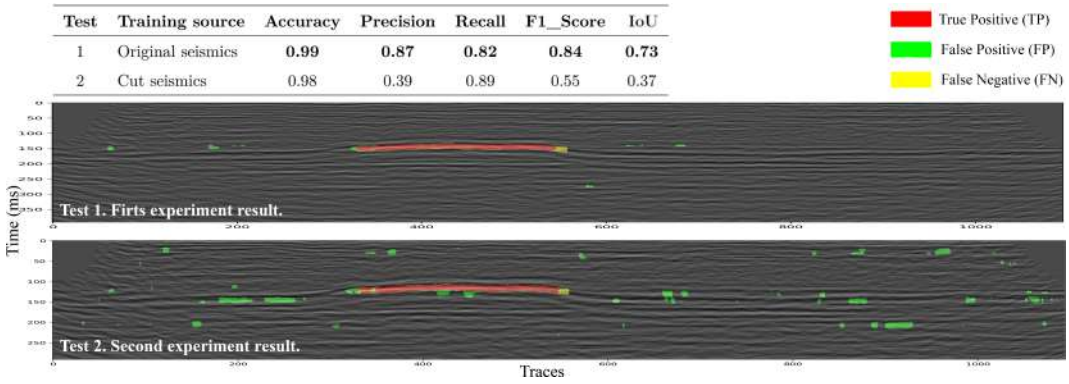


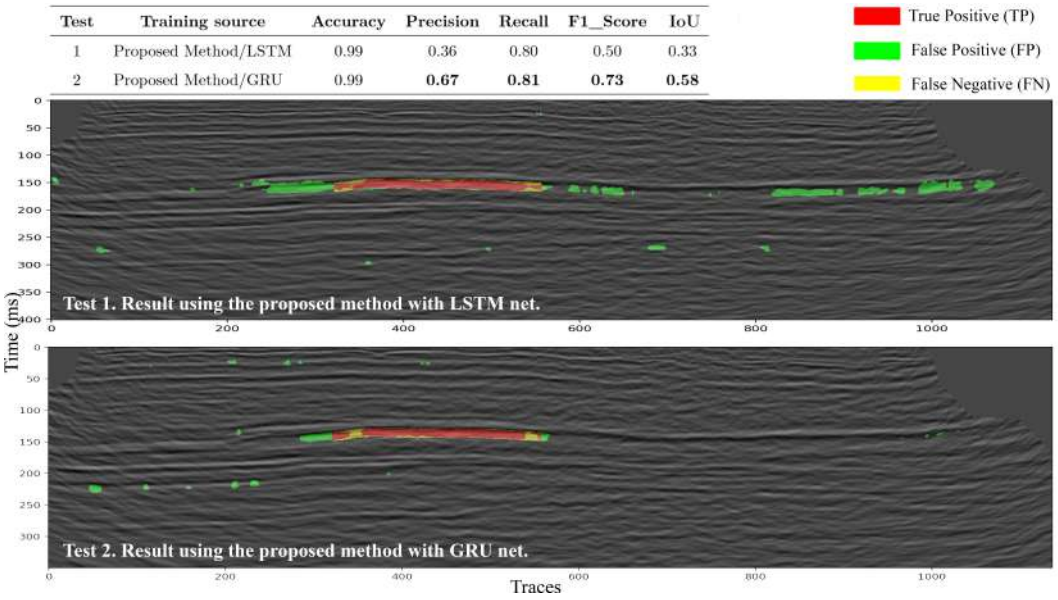
Figure 5.16: Example of deterioration in generalization when using seismic trace cutting.

Comparing the performance of the neural networks shows that GRU has a lower loss of precision with a similar gain of Recall, which produced an increase in the F1 score. Figure 5.17 presents an example of the comparison of the performance of both networks. These results occur when considering all the fields, but there are fields for which the LSTM network has better performance. Figure 5.18 shows an example of this behavior.

The results of the third experiment indicate that the recommendation of training data based on feature similarity works regardless of the technique used to perform the gas indication. Thus, the recommended datasets are valid for subsequent experiments and do not require reprocessing. However, the same is untrue for operational hyper parameters since each gas indication DL model requires tuning.

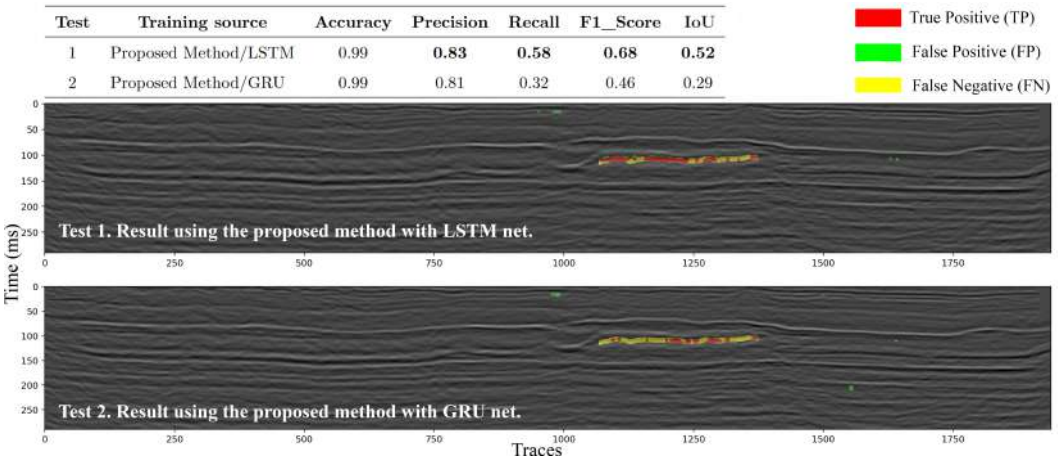
5.3.4 General Analysis of Results

The comparison of features from seismic images allows the creation of several clusters based on similarity, each containing seismic images whose features are similar. This similarity makes them especially representative to



Test	Training source	Accuracy	Precision	Recall	F1_Score	IoU
1	Proposed Method/LSTM	0.99	0.36	0.80	0.50	0.33
2	Proposed Method/GRU	0.99	**0.67**	**0.81**	**0.73**	**0.58**

Figure 5.17: Example of improvement in the indication of natural gas using the GRU network.



Test	Training source	Accuracy	Precision	Recall	F1_Score	IoU
1	Proposed Method/LSTM	0.99	**0.83**	**0.58**	**0.68**	**0.52**
2	Proposed Method/GRU	0.99	0.81	0.32	0.46	0.29

Figure 5.18: Example of improvement in the indication of natural gas using the LSTM network.

be used in the training of the DL model, whose purpose is to process data with features like those of the cluster.

Considering the properties inherent to seismic images, various combinations of rock layers cause many possible feature distributions. In addition, the equipment and capture process and the type of processing used in the data add particularities. Despite this, it is possible to find patterns that establish similarities. The results show that it is effective to use an unsupervised approach to extract features that establish similarities between the seismic images and create training sets that are more representative of the new target seismic images.

The experiments show that the proposed method can increase the correct

indication of natural gas reservoirs, leading to a better generalization of the DL model. This result implies that using training images that share features with those of the new seismic images performs the DL model better, although the number of seismic images is much smaller compared to training with all available data. Likewise, it shows that using a hyper parameter search approach for the DL model based on an ML technique is effective and eliminates the dependency on the user experience.

The results also indicate a loss of precision, but for most experiments, it does not imply a loss in the DL model generalization. Even so, the analysis of the causes of this increase in false positives reveals two possible factors. The first is the restriction of not modifying the number of traces of the original seismic images. The second is the uncertainty in the label of the "No gas" class since there is no certainty that the false positive indication is correct.

When comparing the results of the first and second experiments, it is found that, although the recommendation of the training seismic images allows a better generalization of the model, not completely controlling the traces that are used affects the performance, this is evidenced when comparing the results of the Gavião Vermelho field. In other words, using recommended seismic images whose number of traces is equal to the original seismic image implies using traces that were not recommended (they were only part of a image whose higher number of traces was recommended), that is, non-recommended traces interfere with learning the DL model.

On the other hand, performing a cut of the seismic images to preserve only the traces that are within the exploration fields causes a significant reduction in the number of traces of the "No Gas" class, and depending on the recommended training set they lead to a significant increase in false positives.

In the context of the experiments, generalization performance implies a better performance of the DL models for indicating natural gas when used in new seismic images. For this purpose, the proposed method achieves generalization by specialization. It uses the same architecture as the DL model but is specifically trained to recognize the relevant domains for new seismic images. The identification of these domains within the training data is especially significant, considering that there is no ground truth.

Finally, the results when using the proposed method show an improvement in the generalization of the DL model for natural gas indication, which does not imply a modification of the architecture and is independent of it.

5.3.5 Important Aspects of the Proposed Method 2

The proposed method contains various techniques based on both ML and DL, which perform pre processing, feature extraction, clustering, recommendation of both training data and hyper parameters, and natural gas indication in 2D seismic images, highlighting the main advantages found in the development of the method 2:

1. The proposed method allows a better generalization of the DL models for the indication of natural gas, regardless of the selected DL model.

2. Both the creation of a baseline of seismic features comparison and the recommendation of the training dataset do not require marking labels, which implies an unsupervised approach.

3. From an ML point of view, using an unsupervised approach makes better use of the training data, as it is possible to use all available data to train the Autoencoder-based feature extraction model.

4. Although the training data directly impacts the performance of the DL models, just as important are the hyper parameters that configure the behavior of the models, that is, it is necessary to adjust the hyper parameters so that the DL model can properly learn from the training data.

5. There is no reliance on user experience to determine the training set or choose the operational hyper parameters of the network, which leads to less user operation time required.

6. The proposed method can be applied regardless of the selected natural gas indication technique, which implies that modification of the original architecture is not required to achieve a better generalization.

7. The proposed method presents a comparison basis of seismic features, allowing the creation of representative sets without modifying the original data.

8. The proposed method demonstrates the importance of the representativeness of the training seismic images over their quantity.

9. According to the results of the state of the art analysis, method 2 is the first to use different ML and DL techniques to improve the generalization for the indication of natural gas in 2D seismic images.

In the same way, multiple limitations were identified, among them the following stand out:

1. For each new seismic dataset, it is necessary to perform a new DL model recommendation.

2. The proposed method improves the generalization based on the seismic images within the training set that are more representative of new seismic images. However, if these images do not exist, the generalization will not present improvements.

3. A complete seismic image is recommended based on most traces belonging to a cluster. This fact implies that not everything recommended as training data is controlled.

4. Performing the hyper parameter search using the PSO technique implies an average processing time of sixteen hours.

5.3.6 Research implications

This subsection provides information on how the proposed method will influence current research trends in this area.

Method 2 shows that the application of the analysis of the training data used in the DL model affects the generalization, demonstrating that the selection of representative seismic images about the new study data offers better performance in the indication of natural gas reserves in 2D seismic images.

In the same way, the effectiveness of the operational hyper parameters used in the DL models depends on the domains represented by the features of the training set. That is, the different sets of features require specific hyper parameters that allow recognition of the patterns to be learned by the DL model.

From the point of view of the daily application of the indication of natural gas, the present work offers an automatic method that allows professionals to obtain training images that offer greater representativeness for the new target seismic images. It also selects the hyper parameters used in the DL model. This data otherwise must be determined manually by professionals.

5.4 Conclusion

Method 2 demonstrated that within the seismic training data, there are often unknown domains, and the use of the entire training set does not guarantee that the DL model learns the patterns of each domain. In this context, the selection specifies that the training data relative to the features of the new target seismic images provide better overall model performance. However, this performance depends on the existence of seismic images that are representative of the training set. Otherwise, the performance does not present a significant improvement.

Within the context of generalization performance, the proposed method does not require the modification of the original data or the DL model for the indication of natural gas, which shows the importance of analyzing the available training data. A random selection of the training data in an automatic learning context does not guarantee the selection of the seismic images that offer greater representativeness than the objective data.

6
Improving generalization performance in gas inference DL models for 2D seismic image by recommending both training seismic patches set and DL model training operational hyper parameters - Method 3

This chapter presents a method that separates the seismic image to create a set of standard-sized patches, this seeks to overcome the limitation presented in Chapter 5 in which a complete seismic image is recommended as part of the training set, which implies that some traces used to DL model training were not recommended for the target seismic image. This method is shown independently of the previous ones, presenting a new DL-based feature extraction approach focused on patch processing.

The specific contributions of method 3 to state of the art are: First, it introduces an unsupervised DL patch feature extraction method that establishes a basis for comparing seismic training data and allow the creation of clusters with concentrated representativeness based on pattern similarity. Second, a new training data recommendation method for natural gas inference DL models is presented based on patches that standardize the seismic size. Finally, the proposed method 3 improved the generalization performance of the DL model.

6.1
Proposed Method

This section presents the proposed method to improve the generalization performance of the DL model for gas reservoir inference in 2D seismic images. This method presents a new perspective that uses standard-sized seismic patches to recommend both the training data set and operational hyper parameters to train the DL model according to the patterns in the target seismic image.

As in Chapter 5, the input data is a 2D seismic image set with ground truth for gas reservoirs, and without ground truth indicating the grouping of seismic images based on their features domains, as indicated in Section 4.2.1.2.

The proposed method 3 comprises six sub processes that allow analyzing the seismic images and recommending the training dataset based on the

similarity of the seismic features with those of the target set. This process is carried out based on standard-sized seismic patches that seek to overcome the original size variation between the seismic images. Figure 6.1 provides a high-level description of the proposed method 3, the sub processes are explained in Section 6.1.1 to Section 6.1.6.

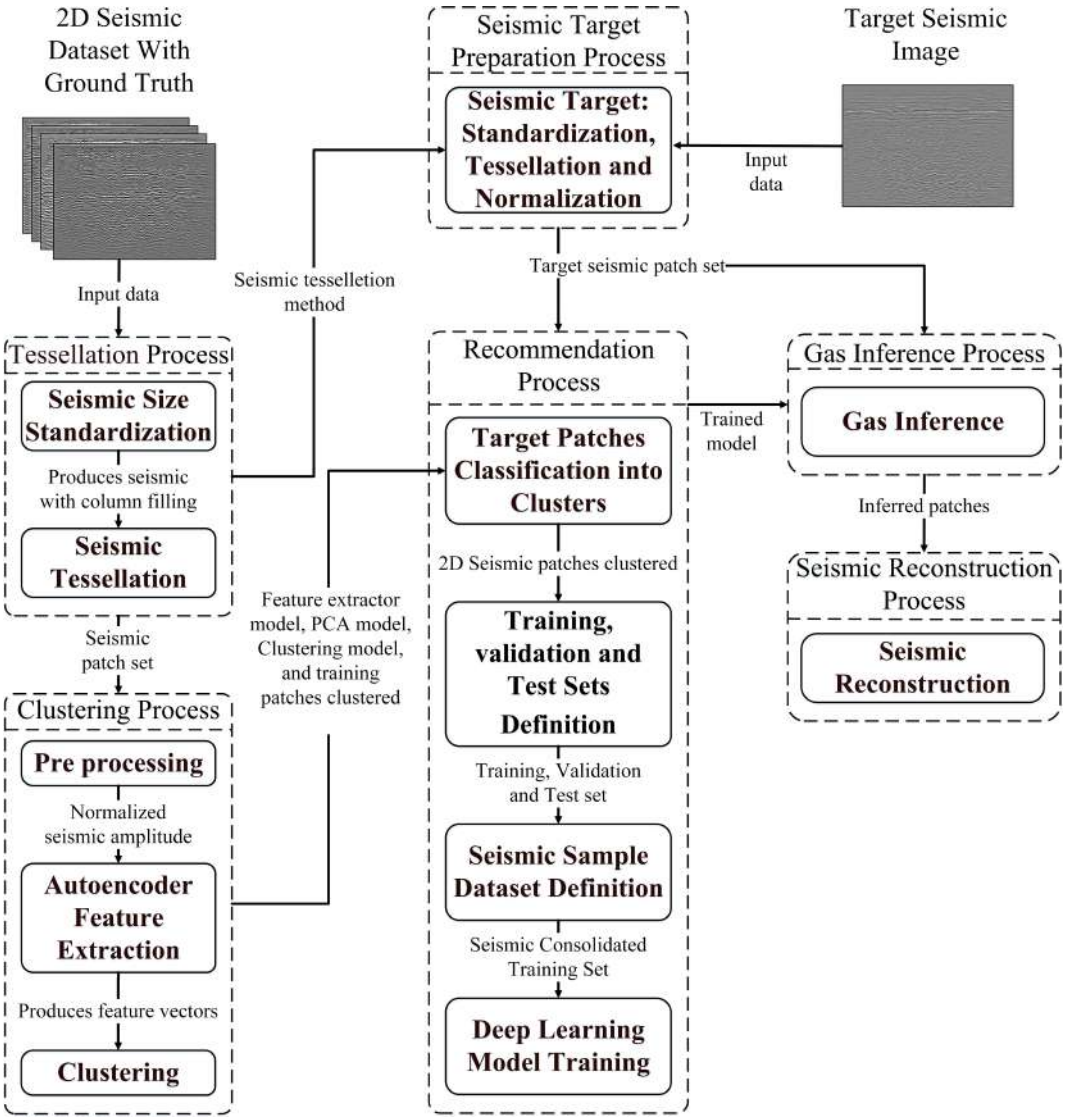


Figure 6.1: Proposed Method 3.

6.1.1 Tessellation Process

This sub process aims to transform the original seismic images training set into a collection of standard-size patches, this allows a comparison based on seismic images with an equal number of traces, which overcomes the variable size problem that normally exists in seismic images.

This step takes the 2D seismic image set to transform it into a set of standard-sized 2D patches. This process requires the segy seismic files as input data and returns as output a set of standard-sized segy seismic files that preserve the original indices information following the SEG standard.

The tessellation process consists of two sequential steps, as shown in Figure 6.2, explained in Section 6.1.1.1 and Section 6.1.1.2.

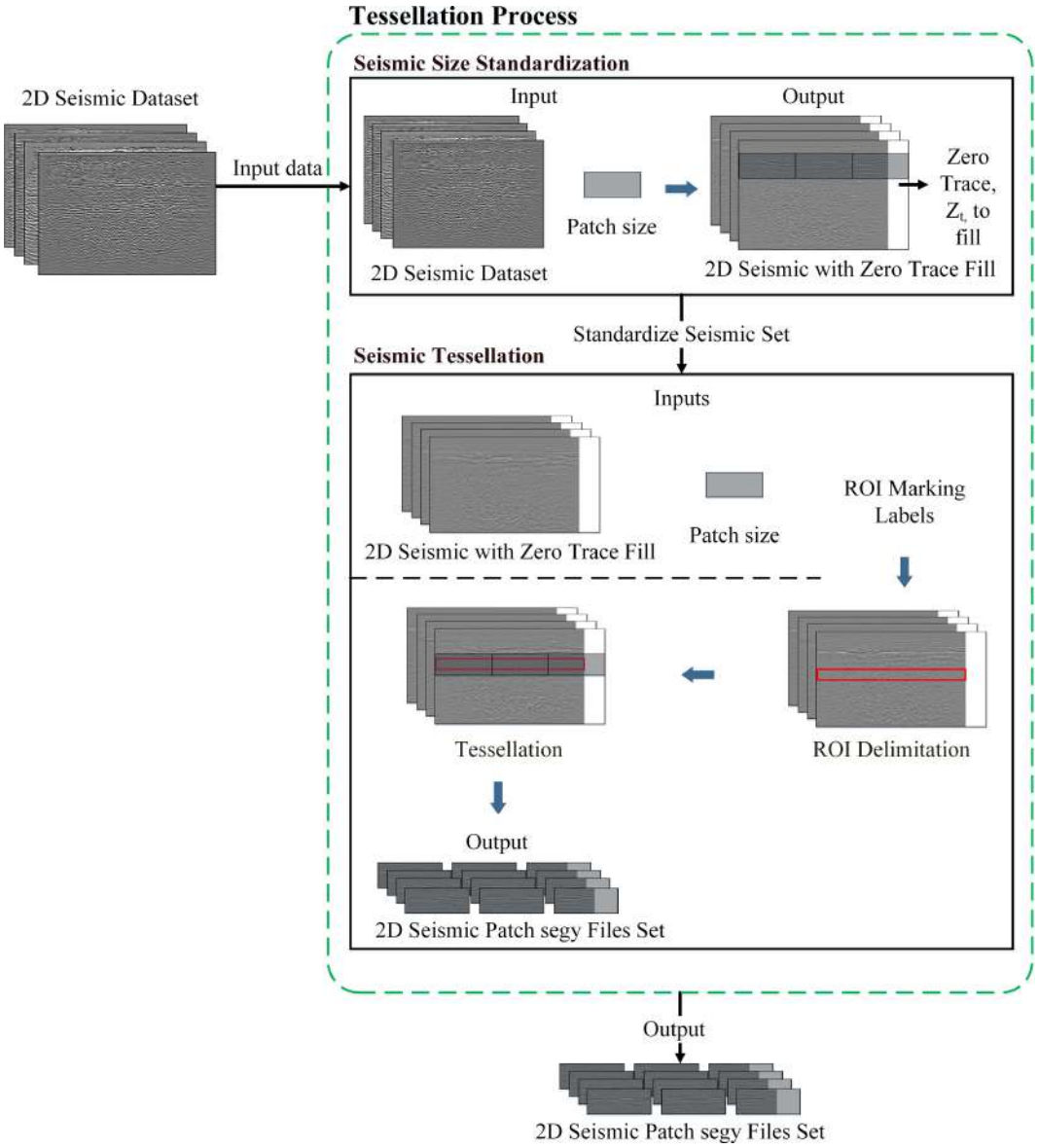


Figure 6.2: Tessellation Process pipeline.

6.1.1.1 Seismic Size Standardization

This step aims to increase the number of traces of the original seismic image to allow the extraction of standard-size patches. This means that depending on the size of the patch to be extracted, it may be necessary to

add traces with a zero value to the end of the seismic image to avoid the final patch having a different size due to the trace's lack.

This step takes as input the original set of training seismic images and the patch size (defined by the user). The result is a collection of seismic images with zero-value traces at the end of each image.

Each original 2D seismic image is modified by adding traces (columns, See Section 2.1) with zero value, this is necessary because there is no indication of the trace number which a seismic file can have.

The amount of zero traces, Z_t, to be added to each seismic image depends on the patch size, this means that it is necessary to add as many Z_t equal to the division modulus between the trace amount of the seismic image, $f\,|t|$, and the traces amount of the patch size, $Patch_{size}\,(t)$, i.e., the value of Z_t is calculate according to the Equation (6-1).

$$Z_{ti} = f_i\,|t| \quad mod \quad Patch_{size}\,(t) \qquad \forall i \in F \tag{6-1}$$

where F is the seismic images set.

6.1.1.2 Seismic Tessellation

This step aims to extract a set of seismic patches from the seismic images to create a training standard-size set, this allows to overcome the problem of seismic comparison with different sizes. This step takes as input the standardized seismic set with zero trace filling, the specification of the patch size, and the indication of the ROI marking labels (see Section 4.2.1.1). As output, a set of seismic image patches of equal size to the ROI is created, this means that the patches are only extracted from the region delimited by the ROI, respecting the size defined by the patch size.

Since the ROI size in the depth component is not too extensive, tessellation is only done relative to the number of traces, i.e., each standardized seismic image is divided into m patches according to Equation (6-2).

$$m_i = \frac{f_i|t|}{Patch_{size}(t)} \quad \forall i \in F \tag{6-2}$$

Note that the number of seismic patches, m, created from each seismic image may differ and depend on each original seismic size.

As a result of this step, a set of seismic patches is created, each saved as a unique segy file, identified by using the key name of the original seismic image from which it came, and adding a unique consecutive number for each patch.

6.1.2 Clusterization Process

This section aims to create several clusters that contain the training patches grouped by the similarity of their features, which allows the creation of groups with concentrated representativeness, which are useful when selecting training data for new seismic targets. This section is composed of three sub processes that create different models for feature extraction, feature analysis, and patch clustering. The set of training 2D seismic patch set is taken as input and, as output, three models are created, in addition to the training patch clusters.

The clustering process consists of three sequential steps, as shown in Figure 6.3, explained in Section 6.1.2.1 to Section 6.1.2.3.

6.1.2.1 Pre processing

This step normalizes the amplitude value of all traces to allow a comparison between patches regardless of the field they come from. The set of seismic image patches is taken as input and a set of normalized seismic image patches is created as output.

Pre processing is necessary since the seismic amplitude varies according to unknown parameters and characteristics in the seismic image acquisition (see Section 4.2.1). The normalization is applied to put each seismic patch on the same scale, using a function that has a response interval of $[-1, +1]$ (Equation (4-1)), allowing the disparity of positive and negative values of each trace to be preserved within a defined representation space.

6.1.2.2 Autoencoder Feature Extraction

This step has two objectives, the first is to train a patch feature extraction model based on Autoencoder, the second is to extract the features from the training patch set. The set of normalized seismic patches is taken as input and, as output, a encoder feature extractor model is created, and the set of features from the training set is obtained. This step allows extracting features from seismic image patches without ground truth. In other words, this step allows extracting features that can be used to perform patch comparisons of seismic images, which constitutes the basis of the present method.

The use of an Autoencoder approach for feature extraction is because it does not require marking labels, which is necessary since there are no seismic similarity labels that would allow cluster creation. In addition, as presented

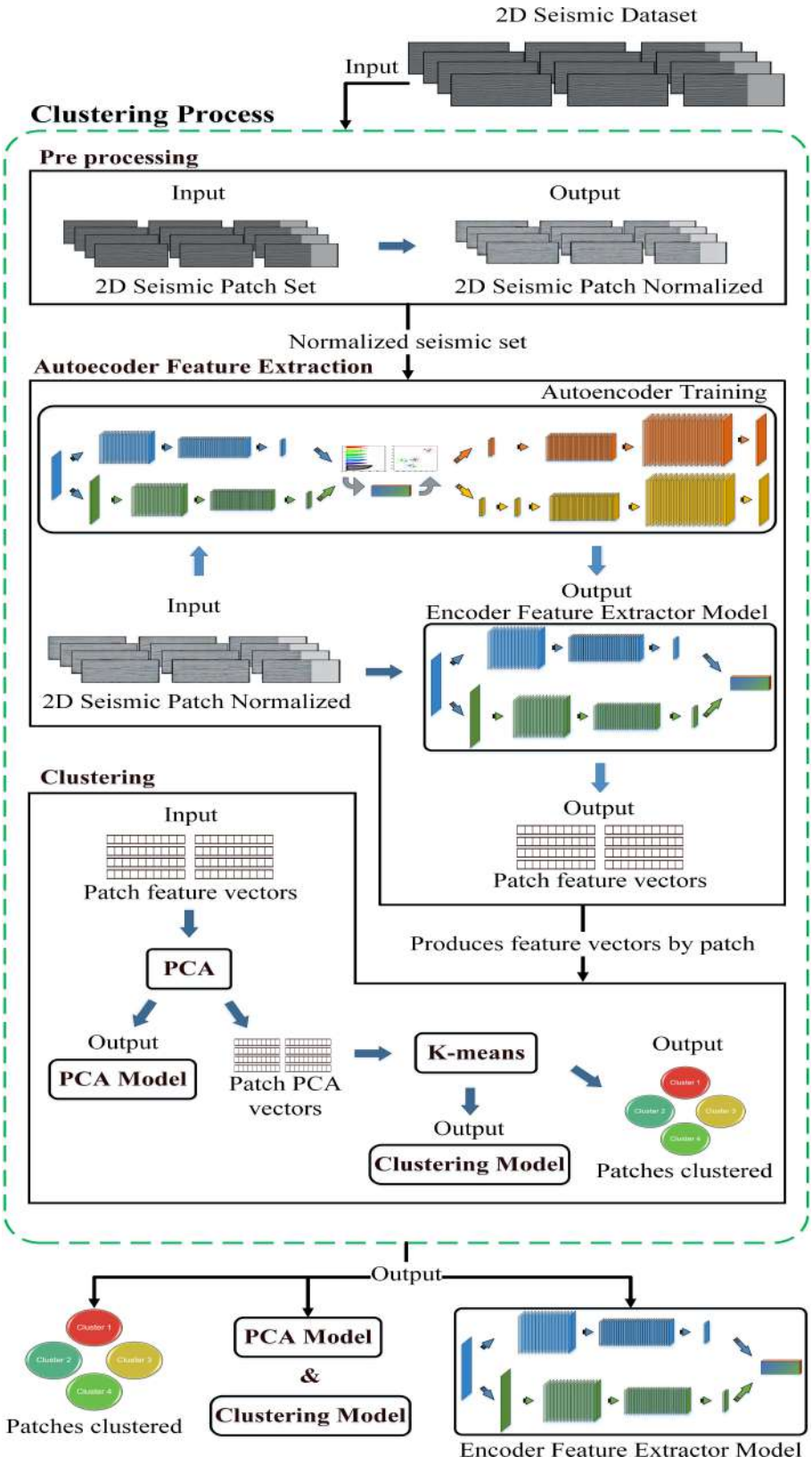


Figure 6.3: Clusterization Process pipeline.

in Section 4.2.1, there are different and unknown domains that according to Section 2.2.2 can be considered as DA in which a DL can help recognize the

different patterns within the seismic patches.

This step builds a feature extraction model based on an Autoencoder concept. However, the architecture is modified by adding two decoding branches plus batch clustering quality evaluation, this new decoder allows evaluation of the latent space using different approaches. Furthermore, two branches are created in the encoder stage to improve feature extraction.

Figure 6.4 presents the Autoencoder-based feature extraction network used in method 3. The encoder step is performed in two branches that receive the normalized 2D seismic patch, the first branch applies three convolution blocks consecutive, which transform the input image patch to a low-dimensional representation space.

The second branch applies a Fourier transform to the original patch before running three consecutive convolution blocks. The Fourier transform is used given its compatibility with seismic data, since these are time series. In addition, the Fourier transform can be inverted, which adapts to the encoder-decoder concept used for the Autoencoder. Using a second coding branch allows features to be extracted from the same seismic patch, but in a different representation space, enriching the latent space.

At the end of the convolution block, both branches are united into a single vector that contains the seismic features within a latent representation space. The idea behind this encoder architecture is to allow the extraction of representative features of the seismic patch by viewing it from different spaces (time and signal).

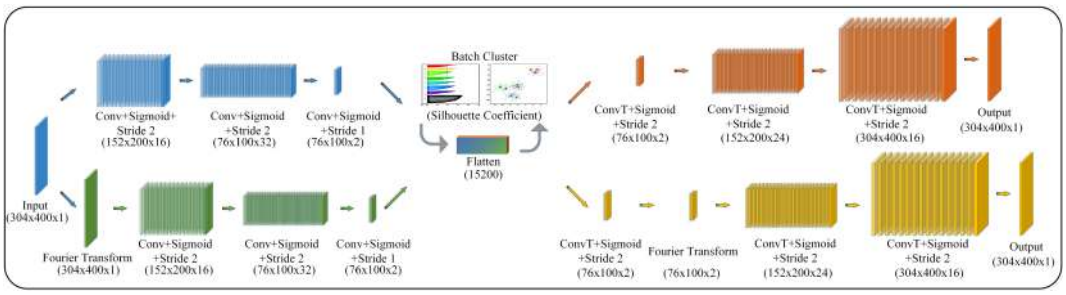


Figure 6.4: Autoencoder (Encoder-Decoder) model architecture for method 3.

The decoder step has two branches, both take the feature vector created in the encoder step as input, the first branch applies three consecutive blocks of transposed convolutions with the aim of recreating the input seismic patch. The second branch applies a transposed convolution and then applies a Fourier transform, then applies two consecutive transposed convolution blocks with the aim of recreating the input seismic patch with the Fourier transform.

The use of two branches allows evaluating the latent space created by

the encoder from two different spaces. The main idea is that the representative space must contain the main features that allow the original entrance to be recreated but expressed in two different spaces. The first branch performs a reconstruction of the original seismic patch and its error is quantified using the MSE loss function. The second branch performs a Fourier transform seismic reconstruction of the patch and the error is quantified using the MSE loss function.

There is another performance metric that is applied to the batch of seismic patches, after a batch is processed, the features extracted by the encoder are used to create different clusters, using the silhouette coefficient (see Section 2.6) to evaluate the quality of the grouping. The number of clusters to be created is automatically calculated based on the silhouette score.

This third metric evaluates the encoder's ability to extract features that can be separated based on their similarity. In other words, the ability of the latent space to extract clusterable features is evaluated and included in the loss function.

From a high level of abstraction, the loss function is composed of the sum of two components starting with the MSE. The first corresponds to the loss of the first branch of the decoder section in which the seismic is represented as time series plus the clustering score. The second corresponds to the second branch of the decoder which is the Fourier branch, and which is also affected by the clustering score. Equation (6-3) present the loss function explained in a high level of abstraction.

$$\begin{aligned}\mathcal{L}(Time, Fourier) = & \, MSE\,(Time) * \alpha\,(Time) \\ & + (1 - \alpha\,(Time)) * (\text{Cluster Score})^{\beta} \\ & + MSE\,(Fourier) * \alpha\,(Fourier) \\ & + (1 - \alpha\,(Fourier)) * (\text{Cluster Score})^{\beta}\end{aligned} \tag{6-3}$$

where α and β indicate how much the MSE and clustering score affect the final loss function, respectively.

The cluster score presents two possible behaviors depending on the value of the silhouette coefficient. If the silhouette value is close to -1, it is considered that it is not possible to group the samples based on the extracted features, and consequently it is necessary for the loss function to cause large changes. On the contrary, if the value of the silhouette is close to 1, it means that clustering is possible and the loss function should cause a small change.

Equation (6-4) presents the loss function for the Autoencoder training process in a more detailed way:

$$\mathcal{L}(x^1, x^2) = \frac{1}{n}\sum_{i=1}^{n}\left(Y_i^{x^1} - \hat{Y_i^{x^1}}\right) * \alpha^{x^1} + \left(1 - \alpha^{x^1}\right) * \begin{Bmatrix} \left(1 - \left(0.6 - \left|S_c^{x^1}\right|\right)\right)^{\beta} & \Longrightarrow & S_c^{x^1} \leq 0.0 \\ \left(1 - \left(S_c^{x^1}\right)^{-1}\right)^{\beta} & \Longrightarrow & S_c^{x^1} > 0.0 \end{Bmatrix} + \frac{1}{n}\sum_{i=1}^{n}\left(Y_i^{x^2} - \hat{Y_i^{x^2}}\right) * \alpha + \left(1 - \alpha^{x^2}\right) * \begin{Bmatrix} \left(1 - \left(0.6 - \left|S_c^{x^2}\right|\right)\right)^{\beta} & \Longrightarrow & S_c^{x^2} \leq 0.0 \\ \left(1 - \left(S_c^{x^2}\right)^{-1}\right)^{\beta} & \Longrightarrow & S_c^{x^2} > 0.0 \end{Bmatrix} \tag{6-4}$$

where x^1 is the original seismic patch, x^2 is the seismic Fourier transformed patch, Y is the reconstructed seismic patch, $\hat{Y_i^{x^1}}$ is the original seismic patch and $\hat{Y_i^{x^2}}$ is the seismic Fourier transform, S_C is the silhouette coefficient, and n is the number of seismic patches.

If the silhouette coefficient is negative, its effect on the loss function is reduced by a threshold value of 0.6, this allows the network to learn even in cases where the extracted features do not allow clustering to be performed. Also, the silhouette coefficient value is subtracted from one in order to ensure that small values have a greater effect on the loss function.

Training the Autoencoder model uses an early stopping approach, with an Adam optimizer and fifteen epochs. The values α and β define the effect each part has on the loss function. however, a value of 0.5 for α and 0.4 for β is recommended.

After the creation of the encoder feature extraction model, the normalized set of 2D seismic patches is processed to extract their features. For each patch a feature vector is created, producing the set of patch feature vectors.

6.1.2.3 Clustering

This step aims to group the seismic image patches into clusters based on the feature vectors, two models are also created that analyze the features and create the clusters. It is in this step where the comparison between seismic patches is made, seeking to identify similarities between them. This models are used in the training set recommendation according to the features of the target seismic images.

To obtain the most relevant features from the vectors set, a feature analysis is applied using the PCA (Section 2.5) technique. This also allows the size reduction of each patch feature vector, which implies a reduction in the seismic image representation space, and facilitates the clustering process.

This step also creates a PCA model that can be used to analyze the target seismic image that is represented in the same feature space.

The next step uses the K-means technique to assign each training PCA patch feature vector to different clusters, and create a clustering model that can be used to identify clusters that are most similar to the target seismic image. This step does not present changes with that one presented in Section 5.1.1.3.

6.1.3 Seismic Target Preparation Process

This process aims to transform the target seismic image into seismic image patches that are in the same representation space as the training set. As a result of this step, the target seismic patch set is created. This step allows the target seismic image to be transformed to compare it with the training set.

This step receives the target seismic image and transforms it using the following steps described in Section 6.1.1. First, size standardization is applied to the seismic image, allowing standard-sized patches to be extracted. Then, using the ROI labels and patch size indication, the standardized seismic image is tessellated, transforming it into a set of seismic patches. Finally the patches are normalized as in the same way as Section 6.1.2.1.

Figure 6.5 presents the pipeline used to transform the target seismic image into seismic patches that will be used in both the recommendation process and the gas inference process.

6.1.4 Recommendation Process

This step aims to train a 2D seismic image gas inference DL model, using a set of recommended seismic patches that come from clusters that have similar features to the target seismic image patches, and also using the recommended operational hyper parameters. It is in this step where the features of the target seismic image are identified and compared with those of the training set, to recommend the most appropriate training data and hyper parameters to be used by the DL model.

The inputs are the set of target seismic patch set created in Section 6.5, the training clusters and the PCA and clustering methods created in Section 6.1.2.3 and, finally, the encoder feature model created in Section 6.1.2.2.

The recommendation process can be applied to a one or more seismic images represented as set of patches and includes five steps that produce a trained DL model as a response, Figure 6.6 presents the pipeline.

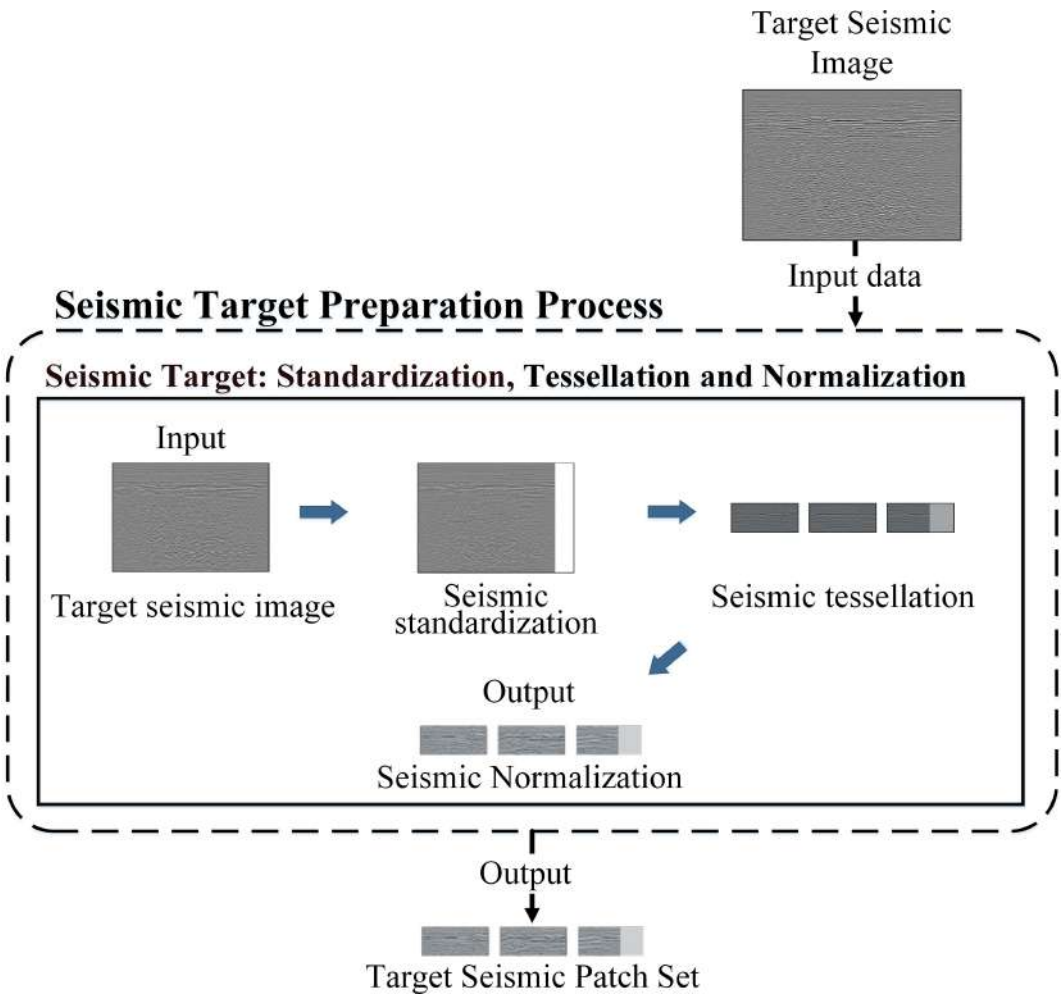


Figure 6.5: Seismic target preparation process pipeline.

6.1.4.1 Target Patches Classification into Clusters

This step aims to assign each target seismic image patch to one of the training clusters, allowing the identification of the training seismic image patches that are most similar to the target image patches.

This step takes the set of target seismic patches to extract their features using the encoder feature extractor model, creating the feature vectors. They are then analyzed using the PCA model, this allows the target patches to be placed in the same representation space of the training set.

The next step uses the clustering model to assign each PCA feature target vector to the most similar cluster. In other words, It identifies the cluster with the most similar features and assigns the target patch.

Finally, the original target seismic images patches are linked to the training seismic patches in the selected cluster, this means that the patch feature vectors are now replaced by the original seismic image patches they represented.

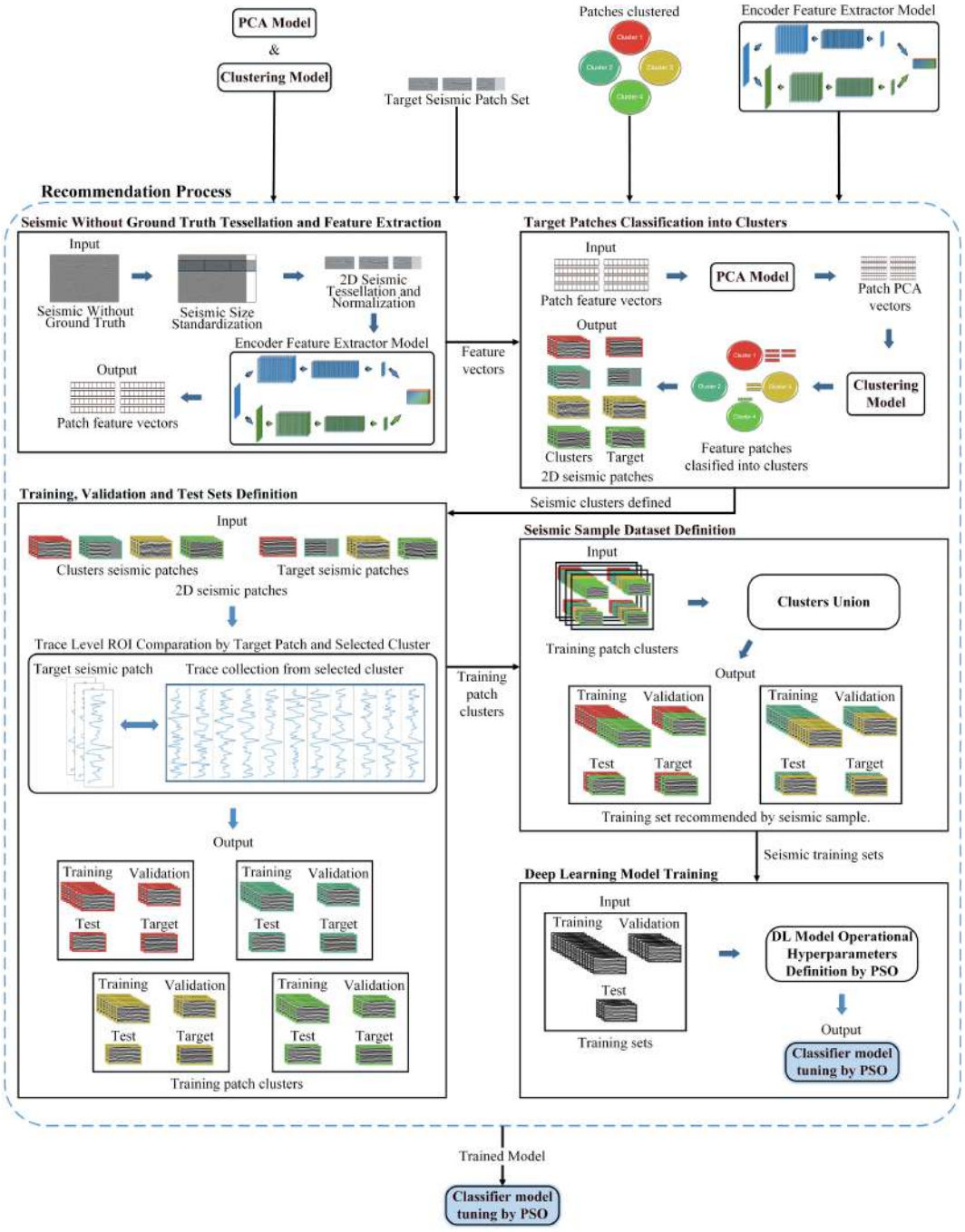


Figure 6.6: Method 3 recommendation process pipeline.

6.1.4.2 Training, validation and Test Sets Definition

This step aims to separate the seismic image patches from the selected clusters to create the training, validation and test subsets, and all the target patches that were assigned to the same cluster are grouped into a set called target. In other words, this step prepares the different sets necessary to train a DL model at the patch level.

The cluster assigned for each seismic patch is separated into training,

validation, and testing subsets, with seventy, twenty, and ten percent of training patches respectively. The training and validation subsets are chosen randomly, but the seismic patches for the test set are selected based on their similarity to each target seismic patch.

The seismic image patches from the selected cluster are similar to the target seismic image patch since the clustering model selected them. However, some patches within the cluster have a trace distribution more like the target seismic patch. This stage compares them at the trace level so that the test data is as similar as possible to those of the target seismic image patch. Similarity is evaluated through the MAE metric.

This process is similar to that in Section 5.1.2.3, and is done because selecting the test set as similar as possible to the target set allows recommending a training set for the DL model completely focused on the target seismic image.

6.1.4.3 Seismic Sample Dataset Definition

This step aims to join the clusters that have patches that come from the same target seismic image. This process allows joining all the clusters and their subsets to train the DL model at the patch level, but which is also representative for the complete target seismic image.

Up to this step, there are training, validation, test, and target subsets for each cluster in which a target seismic patch was recommended, this means that all recommendations are made at the patch level.

In this step, the subsets belonging to different clusters are joined if the recommended target patches come from the same original seismic image, i.e., the patches coming from a single seismic image can be assigned to different clusters, so to recommend a training set for a complete seismic image, it is necessary to join all clusters that have patches that come from the same seismic image.

At the end of this step, training set recommendations are made for each seismic target image.

6.1.4.4 Deep Learning Model Training

This step aims to train the gas inference DL model using the recommended training set, but also recommends operational hyper parameters that allow the DL model to identify the representative pattern within the training set.

This step introduces the operational hyper parameters of a specific DL model to indicate gas reservoirs in 2D seismic image. The result of this step is a trained DL model that can be used on the target seismic image.

The choice of operational hyper parameters refers to configuration parameters that affect how the DL model works without affecting the architecture. The hyper parameters to be tuned are the same selected for method 2 presented in Section 5.1.2.4.

When the search for hyper parameters is manual, the result and the time required will depend on the user's experience, and the complexity of the task increases when considering the multiple possible combinations.

For this reason, the proposed method 3 uses an optimization technique that allows the search to occur within a multidimensional space with an iterative adjustment, an adjustable learning rate, a defined number of iterations, and a focus on reducing the cost function that allows the production of a reasonable solution. Considering all these characteristics, the Particle Swarm Optimization, PSO, (George et al., 2020; Ma et al., 2022; Muisyo et al., 2022; Shi et al., 2022) is selected.

PSO performs a defined number of iterations to test various hyper parameters on the DL model, which implies training and testing the model for the indication of gas reservoirs. This process uses all the defined datasets in Section 6.1.4.3. The training and validation sets teach to the DL model, and the test set measures performance and provides feedback to the PSO, which updates the hyper parameters before starting a new iteration.

In method 3, the GRU neural network (Section 2.3.2) proposed in Section 5 is used to perform gas inference, it is selected because it presents a better generalization performance according to the results presented in Sections 5.3.3.

At the end of the iterative PSO process, the hyper parameters for DL gas inference model training are recommended. It also obtains a specific trained DL model for the target seismic image.

6.1.5 Gas Inference Process

The aim of this step is to use the trained gas inference DL model on the target seismic image patches, i.e., this process runs inference on the set of target patches to obtain a gas indication within each seismic image patch.

This step uses the trained model from Section 6.1.4.4 specific for the patches that belong to the same seismic image, which is represented as patches that come from Section 6.1.3. This DL model was trained using the clusters specifically recommended for the target seismic patches (see Section 6.1.4.3),

this means that for each seismic image, a different gas inference model may exist.

The output of the DL inference model is a set of seismic patches with gas indication. Figure 6.7 presents the pipeline.

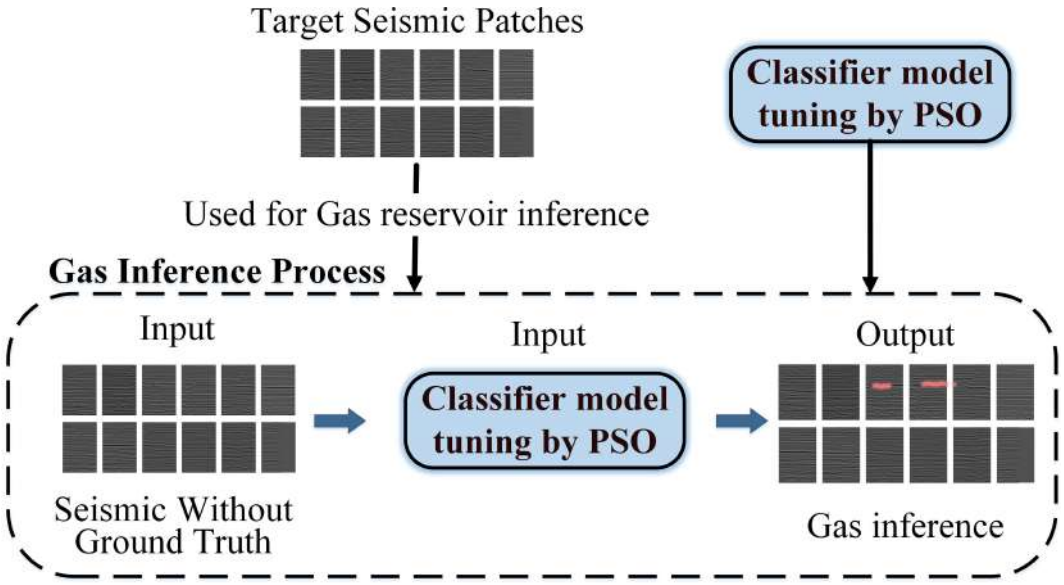


Figure 6.7: Gas indication process method 3 pipeline.

6.1.6 Seismic Reconstruction Process

As the input data for the gas inference process are seismic patches, the outputs are also a set of seismic patches, that is the reason why it is necessary to apply a seismic reconstruction process to concatenate all the target patches into a single seismic image.

To concatenate the patches that come from the same seismic image, the key name described in Section 6.1.1.2 of each patch is used. The patch key name is separated into the seismic key name and the consecutive identification number. The name of the seismic key is used to group all the patches that come from the same seismic image, then the consecutive number indicates the order to concatenate each patch. Figure 6.8 presents the seismic reconstruction pipeline.

6.2 Experiments and Results

This section presents the experiments carried out and their results, following the performance comparison when using a gas inference DL model in 2D seismic image trained with all datasets available versus the results when

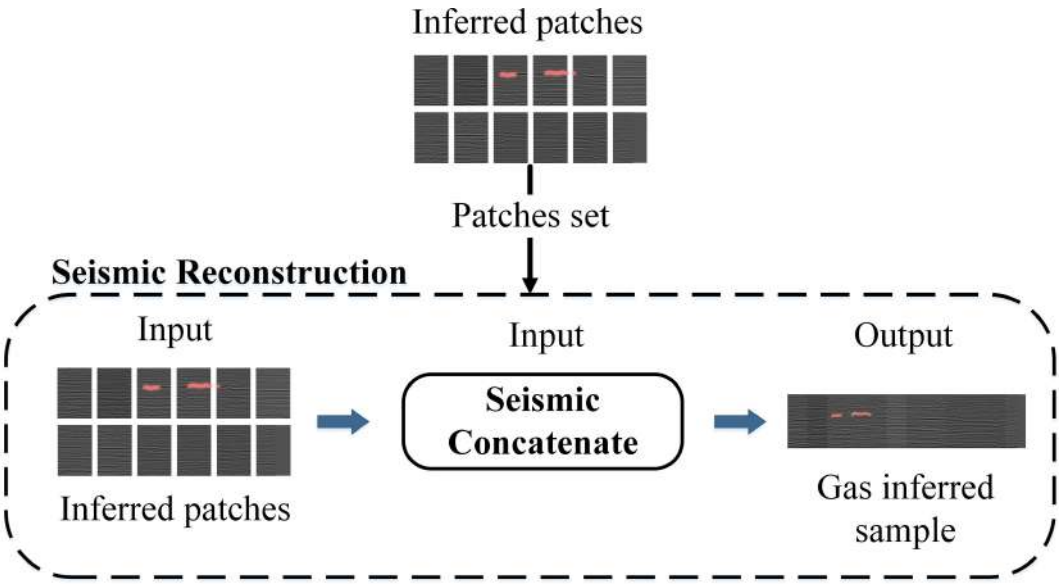


Figure 6.8: Seismic reconstruction process pipeline.

using the same DL model architecture trained by the proposed recommendation method. The performance metrics used in the experiments are described in Section 2.6.

The experiments perform two-stage tests using the gas inference method proposed by Andrade et al. (2021) but change the DL network to a GRU network as described in Section 6.1.4.4.

The first stage obtains the comparison baseline using the entire training database available with the GRU network for five randomly selected target fields, this stage is similar to Section 5.2.3. This test is performed taking care that the DL model does not use any seismic images from the target field for training.

The baseline performs nine tests, in each one it takes a single field as the target and uses the others as training, in this way it uses the traditional approach that randomly assigns the seismic images into three subsets to train the DL model, that is, 70% as training, 20% as validation and 10% as testing.

In the second stage, the proposed method 3 defines both the hyper parameters of the DL model used by the GRU network and the clusters of recommended seismic patches from the training set that will be used according to the target seismic image. In other words, the proposed method (see Section 6.1) is used which takes all the available training data and the gas inference model to make a recommendation that adapts the model to the target seismic images.

The proposed method 3 uses equal-size patches as a comparison basis, this means that all training seismic images have a standard size. Figure 6.9

shows an example of the content of the sets recommended for training and validation for a single seismic belonging to the Gavião Branco field.

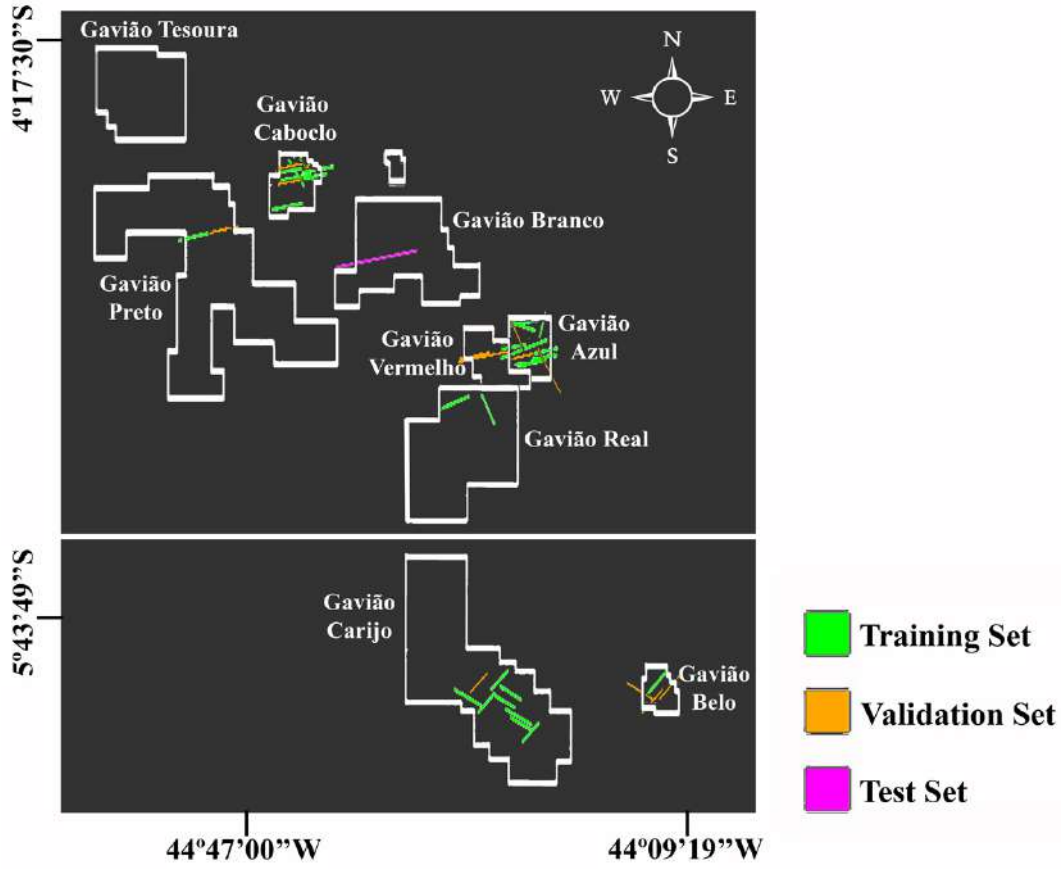


Figure 6.9: Patches recommended for DL model training.

The table 6.1 presents the two-stage test result of the experiment. The table includes the "Patch Amount" column, which refers to the number of patches used in each test.

For all fields, a reduction in the number of patches required to train the DL model is observed, varying between 84% and 87%. In almost all tests there is a reduction in precision which means that the DL gas inference model reduces its ability to recognize "No gas" class patterns, however, there is a significant improvement in recall which means that the DL model improves in the identification of the class "Gas". These results show an improvement in generalization performance based on the F1 score ranging from 2% to 21%, or between 1% and 15% based on the IoU score.

To facilitate the identification of the improvement in the performance of the metrics when using the proposed method, Table 6.2 is presented. The Train Size column shows the percentage of the training set that was used for each field; for example, for Gavião Azul only 14% of all available seismic patches were used for training. The other metrics show the difference between

Table 6.1: Experiment results

Field Target	Training source	Patch Amount	Accuracy	Precision	Recall	F1_Score	IoU
Gavião Azul	All Database	891	0.99	**0.43**	0.52	0.43	0.30
	Proposed Method 3	**126**	0.99	0.40	**0.77**	**0.50**	**0.37**
Gavião Belo	All Database	968	0.99	0.54	0.46	0.48	0.36
	Proposed Method 3	**133**	0.99	**0.55**	**0.70**	**0.60**	**0.49**
Gavião Branco	All Database	666	0.99	**0.46**	0.41	0.40	0.27
	Proposed Method 3	**95**	0.99	0.40	**0.76**	**0.50**	**0.35**
Gavião Caboclo	All Database	892	0.99	**0.30**	0.42	0.32	0.23
	Proposed Method 3	**131**	0.99	0.27	**0.57**	**0.34**	**0.24**
Gavião Carijo	All Database	797	0.99	0.25	0.14	0.17	0.10
	Proposed Method 3	**123**	0.99	**0.33**	**0.53**	**0.38**	**0.25**
Gavião Preto	All Database	1027	0.99	**0.32**	0.19	0.21	0.14
	Proposed Method 3	**161**	0.99	0.30	**0.54**	**0.35**	**0.24**
Gavião Real	All Database	1088	0.99	**0.30**	0.34	0.28	0.18
	Proposed Method 3	**196**	0.99	0.28	**0.63**	**0.35**	**0.23**
Gavião Tesoura	All Database	1017	0.99	**0.21**	0.23	0.20	0.12
	Proposed Method 3	**128**	0.99	0.20	**0.59**	**0.27**	**0.17**
Gavião Vermelho	All Database	1102	0.99	**0.49**	0.50	0.44	0.31
	Proposed Method 3	203	0.99	0.38	**0.78**	**0.47**	**0.33**

the performance obtained when using the proposed method and the baseline, those highlighted in blue represent an improvement, while those highlighted in pink present an equal or worse result.

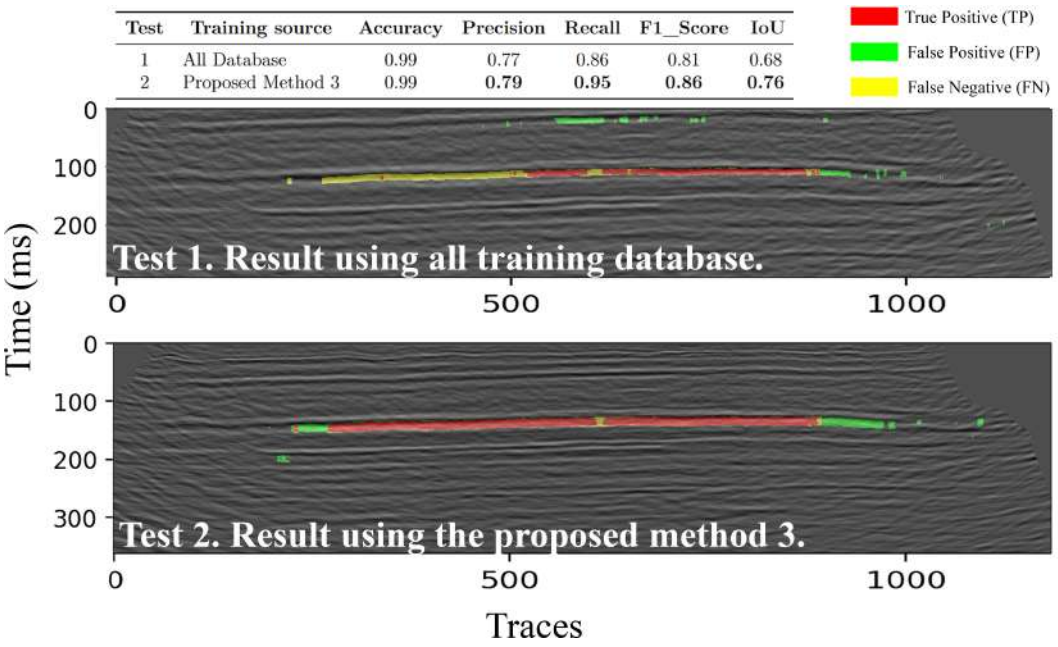
Table 6.2: Method 3 first experiment, improvement of metrics in relation to results using all available data for training.

Field Target	Train size	Precision	Recall	F1_Score	IoU
Gavião Azul	0.14	-0.03	0.25	0.07	0.07
Gavião Belo	0.14	0.01	0.24	0.12	0.14
Gavião Branco	0.14	-0.07	0.35	0.1	0.08
Gavião Caboclo	0.15	-0.03	0.15	0.02	0.01
Gavião Carijo	0.15	0.08	0.39	0.21	0.15
Gavião Preto	0.16	-0.02	0.35	0.14	0.1
Gavião Real	0.18	-0.03	0.3	0.07	0.06
Gavião Tesoura	0.13	-0.01	0.36	0.07	0.05
Gavião Vermelho	0.18	-0.12	0.28	0.03	0.02

Three images illustrate the effects of the proposed method 3 on the seismic images. Figure 6.10 shows an example of the ideal case in which an improvement occurs in all metrics.

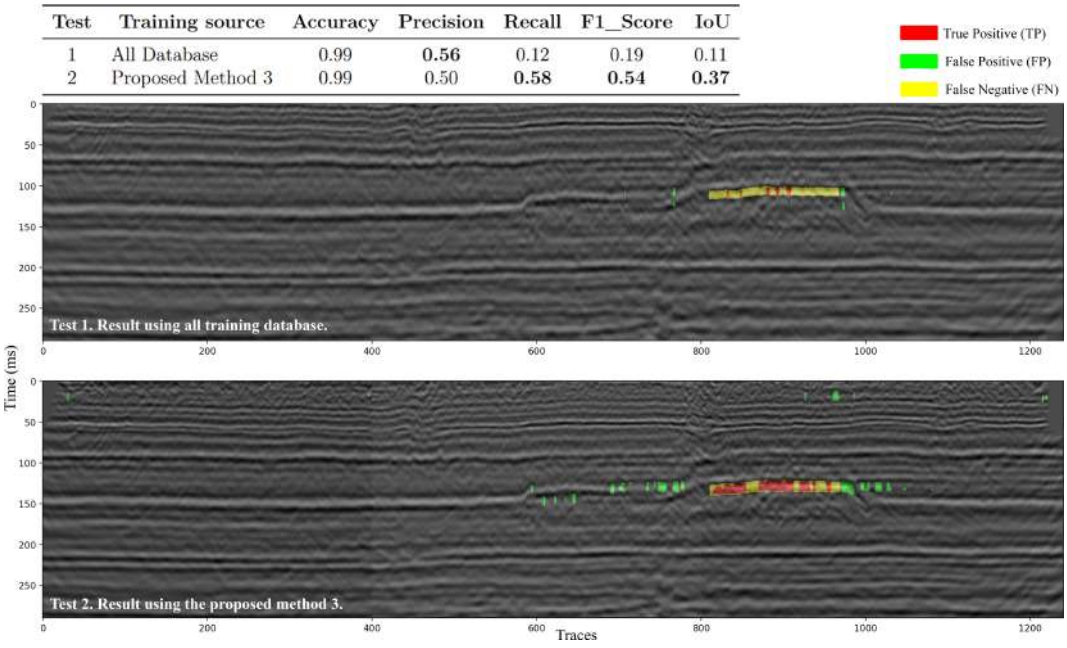
Figure 6.11 shows the case in which an increase in the correct marking of natural gas is obtained but with a loss of precision, which is the most common result according to Table 6.1.

Figure 6.12 shows the case where there are no representative seismic images within the training set with respect to the target seismic, in these



Test	Training source	Accuracy	Precision	Recall	F1_Score	IoU
1	All Database	0.99	0.77	0.86	0.81	0.68
2	Proposed Method 3	0.99	**0.79**	**0.95**	**0.86**	**0.76**

Figure 6.10: Example of improvement in the indication of natural gas.



Test	Training source	Accuracy	Precision	Recall	F1_Score	IoU
1	All Database	0.99	**0.56**	0.12	0.19	0.11
2	Proposed Method 3	0.99	0.50	**0.58**	**0.54**	**0.37**

Figure 6.11: Example of Gas indication improvement and precision loss.

cases, the result does not present an improvement in any metric.

Finally, Table 6.3 presents the recommended operational hyper parameter ranges for DL model training. The recommendation is expressed by ranges since method 3 makes a recommendation for each seismic image, this means that there are as many recommendations as there are seismic images in each field.

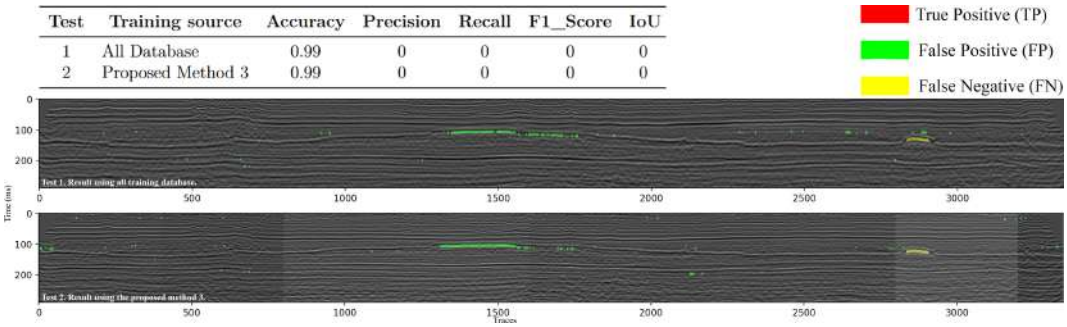


Figure 6.12: Example of no significant improvement in generalization performance.

Table 6.3: Hyper parameter ranges selected per field

Field Target	Gas pixel	ROI pixel	ROI size	Batch size	Balance	# Epoch
Gavião Azul	10 -18	11 - 18	227 - 276	23 - 587	True - False	10 - 55
Gavião Belo	10 - 20	10 - 20	200 - 280	16 - 800	True - False	10 - 200
Gavião Branco	11 - 16	10 - 17	223 - 278	38 - 799	True - False	20 - 110
Gavião Caboclo	12 - 18	11 - 19	230 - 278	23 - 186	True - False	25 - 200
Gavião Carijo	13 - 19	10 - 19	244 - 278	60 - 192	True - False	20 - 95
Gavião Preto	11 - 18	11 - 18	230 - 273	64 - 188	True - False	20 - 45
Gavião Real	14 - 19	10 - 18	244 - 275	37 - 180	True - False	15 - 200
Gavião Tesoura	14 - 19	12 - 19	232 - 279	90 - 737	True - False	10 - 200
Gavião Vermelho	11 - 19	10 - 19	244 - 273	37 - 172	True - False	10 - 200

6.2.1 Discussion

The experiment results show that the use of the proposed method 3 allows a significant increase in the correct indication of gas reserves in 2D seismic images, ranging from $15\% \leq Recall \leq 39\%$. This result demonstrates that within the training set, there are seismic images that are more representative according to the seismic objective, and it is possible to identify them using an unsupervised approach.

However, there is an increase in false positives ranging from $-7\% \leq Precision \leq 8\%$, indicating that there is a persistent challenge in identifying the "No gas" class, when analyzing the results, different possible causes emerge. The first is introduced by method 3 itself, the tessellation process allows seismic comparison with a standard size, but the tessellation does not take into account the position of the gas reserves within the training seismic images, this means that there is no criterion that allows balancing the number of pixels of the "Gas" and "No gas" classes within the patches. The second is related to the class size imbalance, and the reliability issue with the "No Gas" class, as described in Section 4.2.1. The reliability problem arises because within each seismic image all pixels are considered to represent the "No Gas" class unless the specialty market label indicates otherwise, but this may not be true for the entire seismic image, since the Pixels of the class "No Gas" may actually belong to an not

analyzed area and there are no labels that identify them. On the other hand, the "No Gas" class is dominant in the seismic images, this means that there are many more different features that the DL model needs to learn, i.e., there can be many domains within the seismic image and they all belong to the “No gas” class.

6.2.2 General Analysis of Results

Seismic feature analysis allows the identification of different patterns that can be used to create clusters containing similar seismic features, that is, each cluster contains seismic image that is representative of a specific feature domain. Using a cluster to train DL models results in better generalization performance when used on target seismics images that have similar features to the cluster, compared to using a DL models that was trained using the entire seismic image database.

By using a tessellation approach, it is possible to make a comparison between seismics with the same size, which makes it easier to identify the "Gas" class, but presents a challenge for the identification of the "No Gas" class. This is because only patches containing the "Gas" class are used in the recommendation process, in this way, the uncertainty that exists in the marking labels of the "No gas" class is reduced (see Section 4.2.1.1), but there are several implications. First, since there is no indication of balance on the classes in each patch, it is possible that the "No Gas" class has weak representation. Secondly, the spatial context can be lost, i.e. the size of the patch and the position in which it is applied can cut off the layer structure that helps identify the spatial context of the gas reservoir. Third, since tessellation works without overlap, there are fewer patches, which reduces training time but can also reduce the number of domains that can be represented. Fourth, using only patches with class "Gas" eliminates the reliability problem since the areas around the gas reserve have extensive studies carried out by the specialists who make the marking labels. Fifth, tests show that to extract features from seismic patches, the traditional Autoencoder approach in which the original image is reconstructed is not sufficient. To obtain representative features within the gas inference task, it is necessary to modify the architecture and use new latent space evaluation criteria.

The reduction in the amount of patch needed to improve generalization performance indicates that there are patches containing patterns that better represent the new seismic target. However, there are cases in which there are no representative patches for the new seismic target. In these cases the results

are similar to those obtained for training using the entire database.

Finally, in the specific case of gas inference using a DL model on 2D seismic image, there is a generalization performance losses when a model trained on a specific field is used on a new seismic field. The experiment shows that the selection of specific training images affects the performance of the DL model, which means that comparing seismic image for clustering improves the overall performance and can be used on seismic image coming from new explorations. However, new training is necessary and effectiveness depends on the existence of representative data.

6.2.3 Important Aspects of the Proposed Method 3

The proposed method contains various techniques based on both ML and DL, which perform pre processing, feature extraction, clustering, recommendation of both training data and hyper parameters, and natural gas indication in 2D seismic image, highlighting the main advantages found in the development of the method 3:

1. The proposed method creates a base of standard-sized seismic images that can be used for feature comparison.

2. The experiment demonstrates that the analysis and selection of the training dataset have a significant effect on the generalization performance of the DL model used in seismic images.

3. The proposed method enables better generalization performance of the DL model for natural gas reservoir indication compared to the traditional approach using the entire database with random selection for training, validation and testing sets.

4. Creating the Autoencoder feature extraction and the clustering models does not require label marking, i.e., an unsupervised learning process.

5. The proposed method introduces a new Autoencoder-based feature extraction model that uses a three-validation approach, including a clustering performance metric.

6. From an ML point of view, using an unsupervised approach makes better use of the training data, as it is possible to use all available data to train the Autoencoder-based feature extraction model.

7. Conducting the experiment demonstrated that the operational hyper parameters affect the ability of the DL model to identify patterns within

the training set, that is, it is necessary to identify the appropriate operational hyper parameters that enable the learning process, and these depend on the training set.

8. The use of tessellation makes it possible to eliminate the problem of uncertainty in the seismic image because it is possible to use only the patches that contain gas labels, which are areas with extensive study by experts who carry out the labeling.

9. The proposed method does not depend on user experience to determine a DL model training set or the operational hyper parameters selection.

10. The proposed method presents a comparison basis of seismic features, allowing the creation of representative sets without modifying the original seismic traces.

11. The proposed method demonstrates the importance of the representativeness of the training seismic images over their quantity.

12. Based on the state of the art analysis results, method 3 is the first to use a tessellation approach to make a dataset and hyperparameter recommendation to improve the generalization performance for natural gas indication in 2D seismic images.

In the same way, multiple limitations were identified, among them the following stand out:

1. For each target seismic image, it is necessary to train a new DL model and perform a new search for operational hyper parameters.

2. The proposed method depends on the existence of representative seismic images within the seismic image database to make a recommendation that improves the generalization performance of the DL model on target seismic images. If these images do not exist, generalization will not present improvements.

3. There is no tessellation policy that ensures a balanced number of pixels of the "Gas" and "No Gas" classes within each patch. This creates an imbalanced database of training seismic image patches.

4. To eliminate the uncertainty created by the “No gas” class, all patches without gas marking labels are discarded.

5. Performing the hyper parameter search using the PSO technique implies an average processing time of sixteen hours.

6.2.4 Research implications

This subsection provides information on how the proposed method will influence current research trends in this area.

Proposed method 3 demonstrates that analysis of the training database allows the identification of seismic image features which can be used to create clusters with similar representative seismic image patches.

The experiment shows that it is possible to improve the generalization performance of the gas inference model by only using seismic image patches that include gas marking labels, meaning that it is possible to discard training seismic patches that represent areas that have not been analyzed, and so it cannot be known with certainty if they have gas reserves or not.

From the DL point of view, method 3 demonstrates that it is necessary to tune the operational hyper parameters to correctly recognize patterns within the set of training patches, that is, the performance of the DL model is susceptible to the operational hyper parameters.

For the geoscience area that works on the analysis of 2D seismic images to indicate gas reservoirs, the present work offers a technique that can be integrated with DL-based methods to improve the performance of data coming from new exploration campaigns. Offering comparative analyses of training and target data automatically and that do not depend on the experience of professionals.

6.3 Conclusion

Method 3 demonstrated that the use of standard-size seismic image patches allows the identification of seismic features and the creation of clusters based on their similarity. Also, the use of this cluster for the recommendation of the training set for the gas inference DL model based on the comparison with target seismic images improves the generalization performance. However, the improvement depends on the existence of representative patches within the training set. Otherwise, the performance does not present a significant improvement.

7
The Three Methods Comparison

In this chapter, a comparison is made between the three proposed methods from the point of view of their operational application and the results obtained.

7.1
Operational Application

This section presents the advantages and weaknesses of each method about its operability from the point of view of its application.

Advantages of the method 1.

- It allows the user more control over the DL model since method 1 only recommends the training data based on a fixed cluster set. This implies that the user can manually configure the DL model's hyper parameters to indicate natural gas reserves.
- The execution time to obtain a recommended training set is less compared to the other methods. This is because method 1 extracts features that represent seismic images using four techniques that do not require training. Furthermore, the extracted features represent a complete seismic image and not just a patch, which means that the size of the database to be clustered is the same as that of the original seismic images.
- Since the size of the original seismic images training database remains constant (no tessellation or data augmentation processing is performed), it is possible to use a clustering method that does not require an auxiliary technique to determine the number of clusters to create.

Disadvantages of method 1.

- There is no automatic process that verifies the quality of the features extracted by each technique. This means that no quantification determines the contribution made by each feature extraction technique to perform clustering.
- Each feature extraction technique requires a manual hyper parameter tuning process, this implies that the user needs to perform consecutive tests.

- Method 1 has no automatic operative hyper parameter searching for the gas inference DL model, which means that for each new set of seismic images, a manual hyper parameter search is required, which can take months and is completely dependent on user experience.
- Because the feature extractor takes the seismic image as a single element, seismic images with different numbers of traces are processed similarly, which implies shallow feature extraction since seismic images do not have a standard size.

Advantages of method 2.

- For the dataset recommendation process, the seismic image is divided into patches of the same size, this allows to overcome the problem created by the different amounts of traces within the seismic images and allows to consider the structure of each seismic patch for the feature extraction.
- The use of an Autoencoder DL model as a feature extractor for clustering makes it unnecessary to search for hyperparameters when performing new training.
- The feature extraction process based on Autoencoder allows evaluating the quality of the features, through the evaluation of the reconstruction of seismic images.
- An automatic recommendation of the operational hyper parameters for the gas inference DL model training is performed, which implies that for each new seismic image set, it is possible to get both a training set and DL model hyper parameters recommendation in a quantified period of time.

Disadvantages of method 2.

- The user loses the ability to intervene in the tuning processes of the DL models, that is, the user can only define the acceptable ranges for the hyperparameters, and then method 2 makes a recommendation that does not require user intervention, unlike method 1 in which the user participates in each stage of the process.
- Although the training set recommendation is done using a patch approach, the gas inference process still uses a complete seismic image as a training sample, this means that the seismic images used in gas inference contain traces that were not recommended.

- It is necessary to use a technique to automatically define the appropriate number of clusters, which increases processing time. This technique is necessary due to the use of patches, which increase the size of the database, making it impossible to use the clustering technique used in method 1.

Advantages of method 3.

- The original seismic image database is transformed into patches by a tessellation process, the new set of patches is used for both recommendation and gas inference processes. This means that all patches used for the gas inference process were selected by the recommendation process according to the traces belonging to the target seismic image.
- Method 3 allows using only patches that have gas marking labels, this limits the learning process to traces in which there is an analysis carried out by a specialist.
- Autoencoder's new feature extraction model enables evaluation of the representation space using three validation approaches.

Disadvantages of method 3.

- The recommendation of the operational hyperparameters for the gas inference DL model is made for each new seismic image, which means that it is necessary to train the DL model for each seismic image.
- There is no tessellation policy to balance classes within each patch. This produces an imbalanced training set with few samples of the "No Gas" class.
- The user loses the ability to intervene in the tuning processes of the DL models, that is, the user can only define the acceptable ranges for the hyperparameters, and then method 2 makes a recommendation that does not require user intervention, unlike method 1 in which the user participates in each stage of the process.
- It is necessary to use a technique to automatically define the appropriate number of clusters, which increases processing time. This technique is necessary due to the use of patches, which increase the size of the database, making it impossible to use the clustering technique used in method 1.

Table 7.1 presents a comparison of the general characteristics of the proposed methods.

Table 7.1: General comparison of the proposed methods.

Criterion	Method 1	Method 2	Method 3
User intervention. Indicates the level of interaction required with the user.	**High.** The user is responsible for the definition of hyper parameters in both the feature extraction process and the DL gas model inference process.	**Low.** The user only defines the allowed parameter ranges in all processes.	**Low.** The user only defines the allowed parameter ranges in all processes.
		Feature Extraction Process	
Sample size. It refers to the image size from which the method can extract features.	**Non-standard.** The original size of the seismic image is used, it may not be standard.	**Patch of 360 rows and 16 columns.**	**Patch of 304 rows and 400 columns.**
Complexity of feature extraction method. It refers to the complexity of the algorithms that make up the feature extraction process.	$O(n^3)$. Non-DL based method.	**5.611 trainable parameters.** Autoencoder Model.	**18.902 trainable parameters.** Autoencoder Model.
Feature extraction model training time.	**N/A.**	**6 Hours.**	**8.5 Hours**
Quality of the feature extraction process. It refers to whether there is a technique to quantify the quality of the extracted features.	**N/A.**	**Based on Autoencoder reconstruction.**	**Based on Autoencoder two-branch reconstruction and clustering score.**
		Clustering Process	
Complexity of clustering method. It refers to the complexity of the algorithms that make up theclustering process.	$O(nlogn)$	$O(n^2)$	$O(n^2)$
Number of images processed. It refers to the number of seismic images to be clustered.	**< 300**	**< 4.000**	**< 1.100**
Clusters amount definition. Refers to how the number of cluster created from the training data is defined.	**User-defined.**	**Automatic.** Based on the silhouette coefficient	**Automatic.** Based on the silhouette coefficient
Recommended sample. Refers to the recommended seismic image size to train the DL gas inference model.	**Original seismic images without resizing.**	**Original seismic images without resizing.**	**Standard size seismic image patches.**
		Gas Inference DL Model Operative Hyper parameters Definition	
Complexity of hyper parameter definition method. It refers to the complexity of the algorithms that search for the operational hyper parameters of the DL gas inference model.	**N/A.**	$O(n)$	$O(n)$
Hyper parameter definition time. Refers to the time required to select the recommended operating hyper parameters for the DL gas inference model based on the recommended training set.	**Months.** It depends on the user experience and involves performing multiple tests that can take months.	**16 Hours.** Based on PSO technique.	**16 Hours.** Based on PSO technique.

7.2 Metric Result Comparison

In this section, a comparison of the results obtained when applying each method is made.

Table 7.2 presents the results of the methods, where it shows that method 3 presents in all cases an improvement in the correct indication of natural gas according to Recall metric, likewise it presents a deterioration in precision, yet for most fields, it shows better generalization performance according to F1 score and IoU metrics.

Figure 7.1 presents an example where method 2 achieves better performance compared to method 1, but method 3 presents a higher metric regarding correct gas indication that produces better generalization. However, method 2 has a higher precision that produces a cleaner result, related to the presence

Table 7.2: Comparison of results between methods.

Field Target	Training source	Accuracy	Precision	Recall	F1_Score	IoU
Gavião Azul	Method 1	0.99	**0.46**	0.52	0.45	0.32
	Method 2	0.98	0.32	0.76	0.43	0.29
	Method 3	0.99	0.40	**0.77**	**0.50**	**0.37**
Gavião Belo	Method 1	0.99	**0.58**	0.55	0.54	0.43
	Method 2	0.99	0.56	0.63	0.56	0.44
	Method 3	0.99	0.55	**0.70**	**0.60**	**0.49**
Gavião Branco	Method 1	0.99	**0.47**	0.42	0.40	0.27
	Method 2	0.99	0.40	0.72	0.49	**0.35**
	Method 3	0.99	0.40	**0.76**	**0.50**	**0.35**
Gavião Caboclo	Method 1	0.99	**0.28**	0.43	0.29	0.20
	Method 2	0.99	0.23	0.50	0.29	0.21
	Method 3	0.99	0.27	**0.57**	**0.34**	**0.24**
Gavião Carijo	Method 1	0.98	0.31	0.22	0.24	0.16
	Method 2	0.99	0.22	0.33	0.25	0.16
	Method 3	0.99	**0.33**	**0.53**	**0.38**	**0.25**
Gavião Preto	Method 1	0.99	0.29	0.23	0.22	0.15
	Method 2	0.99	0.26	0.41	0.28	0.19
	Method 3	0.99	**0.30**	**0.54**	**0.35**	**0.24**
Gavião Real	Method 1	0.99	**0.36**	0.26	0.26	0.17
	Method 2	0.99	0.27	0.49	0.30	0.20
	Method 3	0.99	0.28	**0.63**	**0.35**	**0.23**
Gavião Tesoura	Method 1	0.99	**0.28**	0.39	**0.31**	**0.21**
	Method 2	0.99	0.16	0.57	0.23	0.14
	Method 3	0.99	0.20	**0.59**	0.27	0.17
Gavião Vermelho	Method 1	0.99	**0.49**	0.50	0.43	0.30
	Method 2	0.98	0.35	0.70	0.43	0.30
	Method 3	0.99	0.38	**0.78**	**0.47**	**0.33**

of false positives as shown in the figure. In this case the metrics indicate that method 3 presents better performance, but the seismic image suggests that method 2 may be more suitable for the geological study.

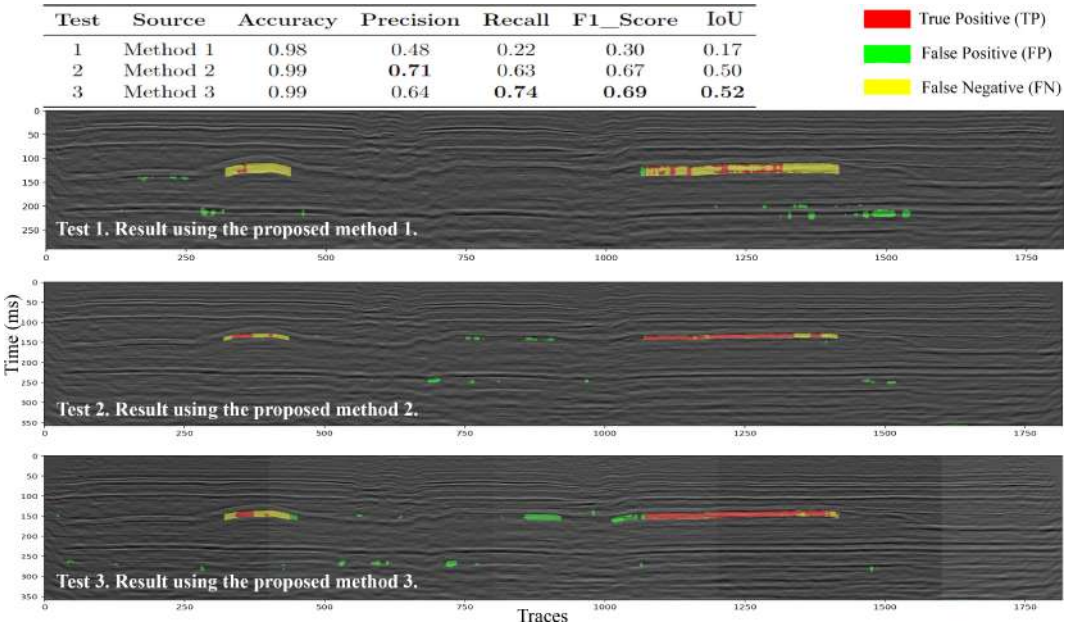


Figure 7.1: First performance comparison.

Figure 7.2 shows a case where method 2 achieves the highest gas reservoir indication based on the recall metric, but also has the lowest precision value. method 3 presents a better precision value than method 2 and a higher recall value compared to method 1. However, since method 1 has few false positives, it offers better generalization based on F1 score and IoU metrics. In this case, method 1 presents a better inference, which is corroborated when analyzing the figure, since there is no presence of false positive blocks.

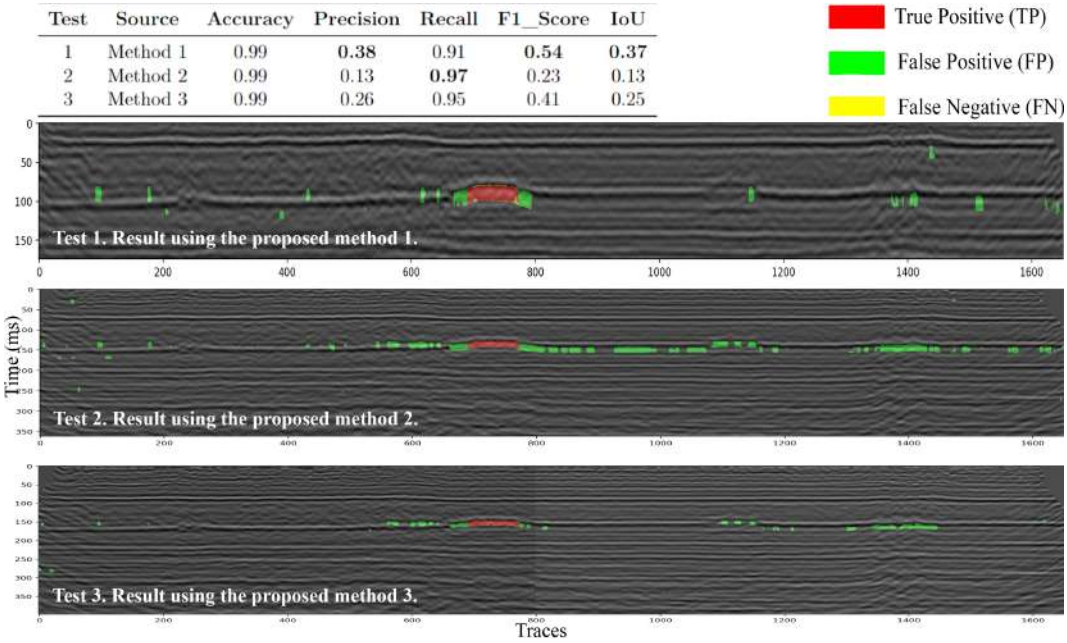


Figure 7.2: Second performance comparison.

Figure 7.3 presents an example where method 3 achieves a higher gas reservoir indication compared to the other methods, but has more false positives than method 2. Despite this, the overall performance shows that method 3 presents a better result.

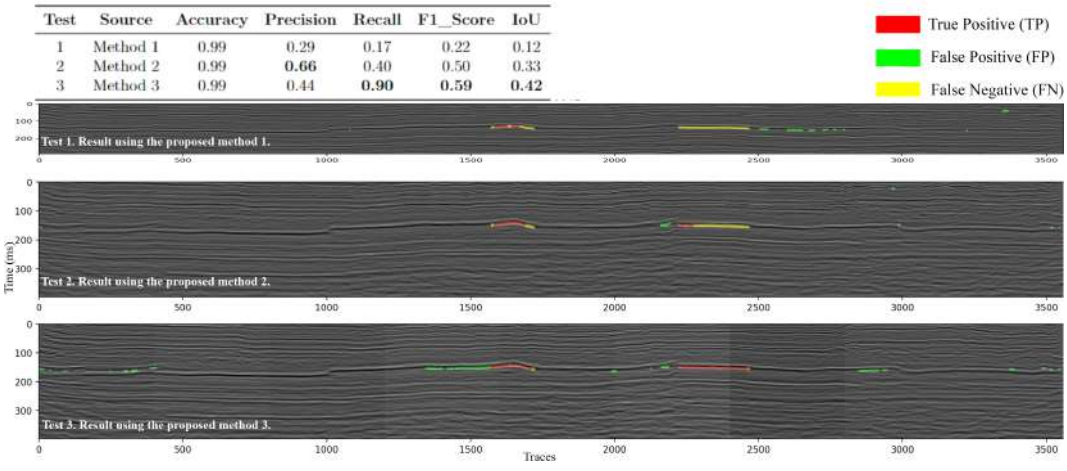
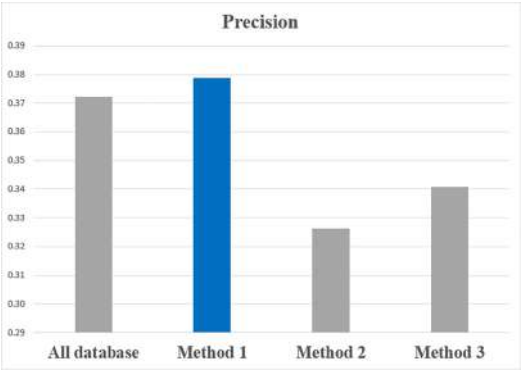


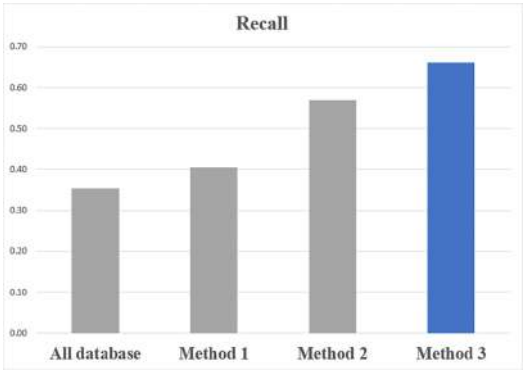
Figure 7.3: Example of metric improvement using method 3.

In general, the methods presented focus on improving DL generalization performance, and each new method seeks to overcome the limitations recognized in the previous one. Figure 7.4 presents the comparison of overall metrics for all methods, including the baseline that uses the entire database as training for the DL model. The result shows that overall method 3 has higher performance, except for the Precision metric in which method 1 achieves better performance, this could happen because method 1 is not a fully automatic method, which means that the user was responsible for setting the hyper parameters.

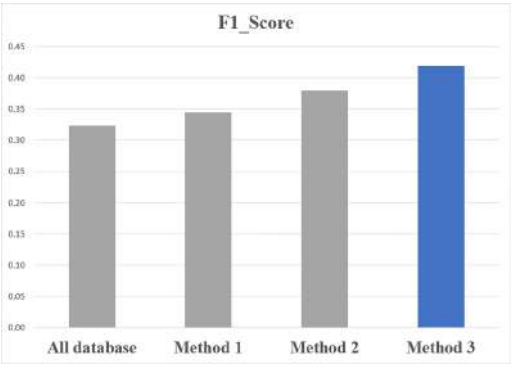
The results generally show a trend of improvement in the generalization performance of the DL model with each new method. In particular, the identification of gas reservoirs presents a significant improvement. However, the increase in false positives demonstrates that it is necessary to focus new research on the identification of the "No Gas" class.



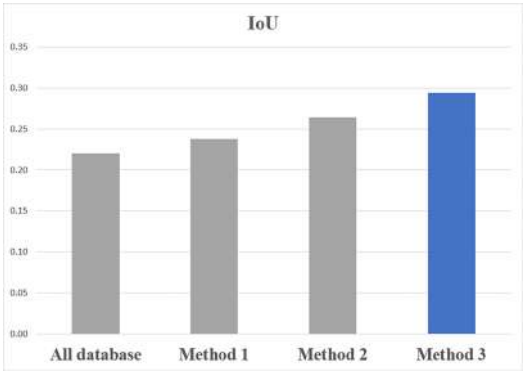
7.4(a): Precision.



7.4(b): Recall.



7.4(c): F1 Score.



7.4(d): IoU.

Figure 7.4: Overall metrics comparison.

8
Conclusions

In this work, various methods and techniques related to seismic imaging, time series, gas indication, generalization performance, performance metrics, deep learning, and machine learning were discussed. All of them were analyzed with the aim of improving the performance of deep learning models designed for gas inference in 2D seismic images, when used in images that come from new exploration fields.

Experiments carried out using the Gavião Cluster of the Paleozoic Parnaíba Basin located in northeastern Brazil, allow us to affirm that within the seismic images there are different patterns that can be considered domains. These different patterns are created in the collection process and are caused by various reasons such as equipment, parameterization, terrain properties, time, and exploration equipment, however, from a computational point of view, most of the reasons are unknown, i.e., no ground truth identifies each domain or which seismic images belong to each.

Analysis of seismic images showed that it is possible to extract representative features that allow comparison even when no ground truth exists. These features allow the creation of clusters that contain seismic images that concentrate representativeness, that is, it is possible to group seismic images that are similar based on the extracted features. These clusters can be considered to contain domains that are related. This method allows images from new explorations to be analyzed to identify the cluster that has the most similar features. In this way, from the available database it is possible to identify more representative images for new seismic images.

Experiments show that using recommended clusters allows better generalization of gas inference DL models, compared to the default DL model trained with all available data. In other words, the results present evidence that it is possible to use available seismic data to find representative samples for images from new fields and that using these samples to train DL models allows better performance to be achieved.

8.1 Contributions

This work proposes three methods with the aim of improving the generalization performance of DL models in the inference of gas reserves, the main contributions of these methods are listed.

- The proposed methods allow improving the generalization performance of the gas inference DL models without altering the network architecture or transforming the original seismic images.
- The proposed methods can be used independently of the selected gas inference DL model.
- A seismic image comparison basis is establish, which is essential for the DL model generalization performance improvement.
- Three methods are proposed that allow the extraction of unsupervised seismic features to compare similarities.
- The proposed methods makes it possible to improve the generalization performance of DL models using the available seismic images. This is made possible by a specialization of the model by using a training set specifically selected for the target seismic images, which helps maintain or improve model performance.
- Three training data set recommendation processes are proposed that allow identifying seismic images that are more representative for a target seismic image, this significantly reduces the training set and in the case of the first proposed process it is possible to identify the cases in which representative seismic images are not available.
- Two of the proposed methods does not depend on user experience to determine the operational hyper parameters to be used based on the features of the specific target seismic image.
- The proposed method demonstrates the importance of the representativeness of the training images over their quantity.
- Proposed method 3 presents a method to obtain standard size seismic images that can be used in the gas inference process and then reconstructed to their original size.
- The proposed method 3 introduces a new Autoencoder-based feature extraction model that uses a three-validation approach, including a clustering performance metric.

- The use of tessellation makes it possible to eliminate the problem of uncertainty in the seismic image because it is possible to use only the patches that contain gas labels, which are areas with extensive study by experts who carry out the labeling.

8.2 Answer to Research Questions

Main Question:

How to alter DL-based methods for gas inference in 2D seismic images to adapt them to specific patterns of new seismic images allowing better generalization performance?

Answer:

To adapt these DL-based methods to the specific patterns of new seismic images, it is possible to add a new process at the beginning of each method that analyzes the training set to separate it into clusters with similar features.

This process makes it possible to compare the patterns of the new seismic images with each cluster and select the most representative one. Using this cluster as a training set for DL-based methods allows learning patterns that are representative for gas inference in the new seismic image, improving generalization performance.

Subsequent Questions:

1. Compared to the results obtained with the default DL model on seismic images, how can the available data be used to allow better generalization performance of the DL model? This question addresses the use and manipulation of the available data, without modifying the existing DL model, that is, it seeks to improve the model's performance on new data but without altering it.

   Answer:

   The available data can be used to create clusters that provide concentrated representativeness based on features extracted from seismic images. By comparing the new seismic image features with the clusters it is possible to recommend a more representative training set that allows better generalization performance, compared to the results of the same DL model using all the data available for training.

   The clustering and recommendation process is independent of the DL model for gas inference, so it does not imply a modification of the model but rather a careful analysis and selection of the training data.

2. How to identify patterns within seismic images to allow a comparison that establishes similarities or domains, which improve the indication of natural gas reserves? This question refers to how to extract features that allow different seismic images to be compared, and that these features are also relevant for identifying natural gas.

   Answer:

   It is possible to use DL-based methods for feature extraction that represent seismic image patterns. By grouping these features using clustering techniques such as HDBSCAN or K-means, clusters with concentrated representation are obtained, which in turn can be considered domains.

   By selecting seismic images that belong to the same domain as the target seismic images, a training set is obtained that allows the DL gas inference model to learn more representative patterns, improving generalization performance.

8.3 Future Works

In the research carried out, challenges arose that were faced by prioritizing the indication for gas and improving the generalization performance of the DL gas inference model.

Each of the proposed methods is presented in order to overcome the challenges encountered, method 3 being the one that is presented in order to overcome the limitation given by the uncertainty in the labeling of the "No gas" class, in addition to the difficulties that arise from the comparison between seismic images of different sizes.

However, it is found that it is necessary as next work to prioritize the recommendation of representative samples for the "No gas" class. This can be developed through the use of a tessellation model with superposition, which would allow preserving the elimination of samples whose uncertainty is high, but at the same time the number of representative traces would increase. Likewise, a recommendation approach can be considered that focuses on the search for a single class, which would imply a double parallel process that would focus on each class independently.

8.4 Scientific Productions

Table 8.1 presents the scientific articles that are produced based on present work.

Table 8.1: Published Articles Based on Proposed Methods

Paper	**Classification**	**Status**
Sepulveda, L. F. M., Gattass, M., Silva, A. C., Quevedo, R., Michelon, D., Siedschlag, C., and Ribeiro, R. (2023a). Generalization of deep learning models for natural gas indication in 2d seismic data. *Pattern Recognition*, 141	A1	Published
Sepulveda, L. F. M., Gattass, M., Silva, A. C., Quevedo, R., Michelon, D., Siedschlag, C., and Ribeiro, R. (2023b). Seismic data classification for natural gas detection using training dataset recommendation and deep learning. *Geoenergy Science and Engineering*, 228. The journal was known until 2022 as Journal of Petroleum Science and Engineering.	A1	Published
Improving generalization performance in gas inference DL models for 2D seismic image by recommending both training seismic patches set and DL model training operational hyper parameters	A1	In progress

Bibliography

Abelha, M., Petersohn, E., Bastos, G., and Araújo, D. (2018). New insights into the Parnaíba Basin: Results of investments by the Brazilian National Petroleum Agency. *Geological Society Special Publication*, 472(1):361–366.

Ahmadi, H. R., Mahdavi, N., and Bayat, M. (2021). A novel damage identification method based on short time Fourier transform and a new efficient index. *Structures*, 33(August 2020):3605–3614.

Ahmed, S. M., Raychaudhuri, D. S., Paul, S., Oymak, S., and Roy-Chowdhury, A. K. (2021). Unsupervised multi-source domain adaptation without access to source data. pages 10098–10107. IEEE.

Ahn, Y. and Choe, J. (2022). Reliable channel reservoir characterization and uncertainty quantification using variational autoencoder and ensemble smoother with multiple data assimilation. *Journal of Petroleum Science and Engineering*, 209.

Alfarhan, M., Deriche, M., Maalej, A., Alregib, G., and Al-Marzouqi, H. (2020a). Multiple events detection in seismic structures using a novel u-net variant. *Proceedings - International Conference on Image Processing, ICIP*, 2020-October:2900–2904.

Alfarhan, M., Maalej, A., and Deriche, M. (2020b). Concurrent detection of salt domes and faults using resnet with u-net. *Proceedings - 2020 6th Conference on Data Science and Machine Learning Applications, CDMA 2020*, pages 118–122.

Alsadi, H. N. (2017). *Seismic Hydrocarbon Exploration*. Springer International Publishing.

Amirabadi, M. A., Kahaei, M. H., and Nezamalhosseini, S. A. (2020). Novel suboptimal approaches for hyperparameter tuning of deep neural network [under the shelf of optical communication]. *Physical Communication*, 41.

Andrade, F., Fernando Santos, L., Gattass, M., Quevedo, R., Michelon, D., Siedschlag, C., and Ribeiro, R. (2021). Gas reservoir segmentation in 2D onshore seismics using LSTM-AutoEncoder. In *First International*

Meeting for Applied Geoscience & Energy Expanded Abstracts, pages 1651–1655. Society of Exploration Geophysicists.

ANP (2012). Anp - agência nacional do petróleo gás natural e biocombustíveis. geoanp – mapa de dados georreferenciados. `http://geo.anp.gov.br/mapview`, Last accessed on 2022-01-19.

Azzam, S. S. S., Elkady, H. H., and Rabea, T. M. M. (2018). The impact of seismic interpretation on the hydrocarbon trapping at Falak field, Meleiha, Western Desert, Egypt. *Egyptian Journal of Petroleum*, 27(4):785–793.

Bai, T. and Tahmasebi, P. (2021). Attention-based lstm-fcn for earthquake detection and location. *Geophysical Journal International*, 228:1568–1576.

Baxevanis, A. D. and Ouellette, B. F. F. (2004). *BIOINFORMATICS A Practical Guide to the Analysis of Genes and Proteins SECOND EDITION*. Wiley edition.

Bhandari, N. and Pahwa, P. (2023). Evaluating partitioning based clustering methods for extended non-negative matrix factorization (nmf). *Intelligent Automation and Soft Computing*, 35:2043–2055.

Canchumuni, S. W., Emerick, A. A., and Pacheco, M. A. C. (2019). Towards a robust parameterization for conditioning facies models using deep variational autoencoders and ensemble smoother. *Computers and Geosciences*, 128:87–102.

Cao, Z., Ma, L., Long, M., and Wang, J. (2018). Partial adversarial domain adaptation.

Caruana, R. (1997). Multitask learning. *Machine Learning*, 28:41–75.

Chang, J., Li, J., Kang, Y., Lv, W., Xu, T., Li, Z., Xing Zheng, W., Han, H., and Liu, H. (2021). Unsupervised domain adaptation using maximum mean discrepancy optimization for lithology identification. *GEOPHYSICS*, 86(2):ID19–ID30.

Chattopadhyay, R., Sun, Q., Fan, W., Davidson, I., Panchanathan, S., and Ye, J. (2012). Multisource domain adaptation and its application to early detection of fatigue. *ACM Transactions on Knowledge Discovery from Data*, 6(4).

Cho, K., van Merrienboer, B., Gulcehre, C., Bahdanau, D., Bougares, F., Schwenk, H., and Bengio, Y. (2014). Learning phrase representations using rnn encoder-decoder for statistical machine translation.

Chui, C. K., Jiang, Q., Li, L., and Lu, J. (2021). Analysis of an adaptive short-time Fourier transform-based multicomponent signal separation method derived from linear chirp local approximation. *Journal of Computational and Applied Mathematics*, 396:113607.

da Cruz, L. B., Souza, J. C., de Sousa, J. A., Santos, A. M., de Paiva, A. C., de Almeida, J. D. S., Silva, A. C., Junior, G. B., and Gattass, M. (2020). Interferometer eye image classification for dry eye categorization using phylogenetic diversity indexes for texture analysis. *Computer Methods and Programs in Biomedicine*, 188:105269.

de Carvalho Filho, A. O., Silva, A. C., de Paiva, A. C., Nunes, R. A., and Gattass, M. (2018). Classification of patterns of benignity and malignancy based on CT using topology-based phylogenetic diversity index and convolutional neural network. *Pattern Recognition*, 81:200–212.

De Miranda, F. S., Vettorazzi, A. L., Cunha, P. R. C., Aragão, F. B., Michelon, D., Caldeira, J. L., Porsche, E., Martins, C., Ribeiro, R. B., Vilela, A. F., Corrêa, J. R., Silveira, L. S., and Andreola, K. (2018). Atypical igneous-sedimentary petroleum systems of the Parnaíba Basin, Brazil: Seismic, well logs and cores. *Geological Society Special Publication*, 472(1):341–360.

de Sousa Costa, R. W., da Silva, G. L. F., de Carvalho Filho, A. O., Silva, A. C., de Paiva, A. C., and Gattass, M. (2018). Classification of malignant and benign lung nodules using taxonomic diversity index and phylogenetic distance. *Medical and Biological Engineering and Computing*, 56(11):2125–2136.

Delgado, J. M. D. and Oyedele, L. (2021). Deep learning with small datasets: using autoencoders to address limited datasets in construction management. *Applied Soft Computing*, 112.

Dell, S., Walda, J., Hoelker, A., and Gajewski, D. (2020). Categorizing and correlating diffractivity attributes with seismic-reflection attributes using autoencoder networks. *Geophysics*, 85:O59–O70.

Duong, L. T., Le, N. H., Tran, T. B., Ngo, V. M., and Nguyen, P. T. (2021). Detection of tuberculosis from chest X-ray images : Boosting the performance with vision transformer and transfer learning. *Expert Systems With Applications*, 184(July):115519.

Fei, G., Wang, S., and Liu, B. (2016). Learning cumulatively to become more knowledgeable. volume 13-17-August-2016, pages 1565–1574. Association for Computing Machinery.

Fei-Fei, L., Fergus, R., and Perona, P. (2006). One-shot learning of object categories. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 28:594–611.

Fernando Santos, L., Gattass, M., Correa Silva, A., Miranda, F., Siedschlag, C., and Ribeiro, R. (2020). Natural gas detection in onshore data using transfer learning from a LSTM pre-trained with offshore data. In *SEG Technical Program Expanded Abstracts 2020*, pages 1190–1195. Society of Exploration Geophysicists.

Gatti, F. and Clouteau, D. (2020). Towards blending physics-based numerical simulations and seismic databases using generative adversarial network. *Computer Methods in Applied Mechanics and Engineering*, 372:113421.

Geng, C., Huang, S.-J., and Chen, S. (2021). Recent advances in open set recognition: A survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 43:3614–3631.

George, S. T., Subathra, M. S., Sairamya, N. J., Susmitha, L., and Premkumar, M. J. (2020). Classification of epileptic eeg signals using pso based artificial neural network and tunable-q wavelet transform. *Biocybernetics and Biomedical Engineering*, 40:709–728.

Ghosh, S., Singh, R., Vatsa, M., Ratha, N., and Patel, V. M. (2020). *Domain Adaptation for Visual Understanding*. Springer International Publishing.

Gonzalez, R. C. and Woods, R. E. (2008). *Digital image processing*. Prentice Hall, Upper Saddle River, N.J.

Goodfellow, I. J., Bengio, Y., and Courville, A. (2016). *Deep Learning*. MIT Press, Cambridge, MA, USA. `http://www.deeplearningbook.org`.

Gupta, H. K., editor (2021). *Encyclopedia of Solid Earth Geophysics*. Springer International Publishing.

Han, J., Kamber, M., and Pei, J. (2012). Advanced cluster analysis.

Hermessi, H., Mourali, O., and Zagrouba, E. (2019). Deep feature learning for soft tissue sarcoma classification in MR images via transfer learning. *Expert Systems with Applications*, 120:116–127.

Hochreiter, S. and Schmidhuber, J. (1997). Long short-term memory. *Neural Computation*, 9:1735–1780.

Hu, K. and Wu, X. (2022). Mode shape prediction based on convolutional neural network and autoencoder. *Structures*, 40:127–137.

Huang, Y., Xu, J., Wu, Q., Zheng, Z., Zhang, Z., and Zhang, J. (2019). Multi-Pseudo Regularized Label for Generated Data in Person Re-Identification. *IEEE Transactions on Image Processing*, 28(3):1391–1403.

Jiang, J., Ren, H., and Zhang, M. (2022). A convolutional autoencoder method for simultaneous seismic data reconstruction and denoising. *IEEE Geoscience and Remote Sensing Letters*, 19.

Jiang, Z., Zhang, S., Turnadge, C., and Xu, T. (2021). Combining autoencoder neural network and bayesian inversion to estimate heterogeneous permeability distributions in enhanced geothermal reservoir: Model development and verification. *Geothermics*, 97.

Jin, C., Cheng, Y., Yang, X., Li, S., Hu, J., and Lan, G. (2022). Adaptive classification of aggregate morphologies using clustering for investigation of correlation with contact characteristics of aggregates. *Construction and Building Materials*, 349.

Jin, Q., Cui, H., Sun, C., Meng, Z., Wei, L., and Su, R. (2021). Domain adaptation based self-correction model for COVID-19 infection segmentation in CT images. *Expert Systems with Applications*, 176(January):114848.

Jolliffe, I. T. (2002). *Principal Component Analysis, Second Edition*.

Kandanaarachchi, S. (2022). Unsupervised anomaly detection ensembles using item response theory. *Information Sciences*, 587:142–163.

Kannammal, G. R., Sivamalar, P., Santhi, P., Vetriselvi, T., Kalpana, V., and Nithya, T. (2022). Prediction of quality in production using optimized hyper-parameter tuning based deep learning model. *Materials Today: Proceedings*.

Kar, C. and Banerjee, S. (2021). Intensity prediction of tropical cyclone using multilayer multi-block local binary pattern and tree-based classifiers over North Indian Ocean. *Computers and Geosciences*, 154(February):104798.

Kim, M. and Song, J. (2022). Near-real-time identification of seismic damage using unsupervised deep neural network. *Journal of Engineering Mechanics*, 148.

Kislov, K. V. and Gravirov, V. V. (2018). Deep artificial neural networks as a tool for the analysis of seismic data. *Seismic Instruments*, 54:8–16.

Kislov, K. V., Gravirov, V. V., and Vinberg, F. E. (2020). Possibilities of seismic data preprocessing for deep neural network analysis. *Izvestiya, Physics of the Solid Earth*, 56:133–144.

Kong, Q., Chiang, A., Aguiar, A. C., Fernández-Godino, M. G., Myers, S. C., and Lucas, D. D. (2021). Deep convolutional autoencoders as generic feature extractors in seismological applications. *Artificial Intelligence in Geosciences*, 2:96–106.

Lee, K., Lim, J., Ahn, S., and Kim, J. (2018). Feature extraction using a deep learning algorithm for uncertainty quantification of channelized reservoirs. *Journal of Petroleum Science and Engineering*, 171:1007–1022.

Lei, X., Xia, Y., Wang, A., Jian, X., Zhong, H., and Sun, L. (2023). Mutual information based anomaly detection of monitoring data with attention mechanism and residual learning. *Mechanical Systems and Signal Processing*, 182.

Leng, D., Zheng, L., Wen, Y., Zhang, Y., Wu, L., Wang, J., Wang, M., Zhang, Z., He, S., and Bo, X. (2022). A benchmark study of deep learning-based multi-omics data fusion methods for cancer. *Genome Biology*, 23.

Lentzakis, A. F., Seshadri, R., Akkinepally, A., Vu, V.-a., and Ben-akiva, M. (2020). Hierarchical density-based clustering methods for tolling zone definition and their impact on distance-based toll optimization. *Transportation Research Part C*, 118(August):102685.

Li, K., Chen, S., and Hu, G. (2020). Seismic labeled data expansion using variational autoencoders. *Artificial Intelligence in Geosciences*, 1:24–30.

Li, K., Zong, J., Fei, Y., Liang, J., and Hu, G. (2022). Simultaneous seismic deep attribute extraction and attribute fusion. *IEEE Transactions on Geoscience and Remote Sensing*, 60.

Li, W., Yue, D., Colombera, L., Du, Y., Zhang, S., Liu, R., and Wang, W. (2021). Quantitative prediction of fluvial sandbodies by combining seismic attributes of neighboring zones. *Journal of Petroleum Science and Engineering*, 196:107749.

Li, Y., Yuan, L., and Vasconcelos, N. (2019). Bidirectional Learning for Domain Adaptation of Semantic Segmentation. In *2019 IEEE/CVF Conference*

on Computer Vision and Pattern Recognition (CVPR), pages 6929–6938. IEEE.

Lin, C.-h., Hsu, K.-c., Johnson, K. R., Luby, M., and Fann, Y. C. (2019). International Journal of Medical Informatics Applying density-based outlier identifications using multiple datasets for validation of stroke clinical outcomes. *International Journal of Medical Informatics*, 132(August):103988.

Liu, X., Li, B., Li, J., Chen, X., Li, Q., and Chen, Y. (2021). Semi-supervised deep autoencoder for seismic facies classification. *Geophysical Prospecting*, 69:1295–1315.

Liu, X., Shao, G., Liu, Y., Liu, X., Li, J., Chen, X., and Chen, Y. (2022). Deep classified autoencoder for lithofacies identification. *IEEE Transactions on Geoscience and Remote Sensing*, 60.

Lloyd, S. (1982). Least squares quantization in PCM. *IEEE Transactions on Information Theory*, 28(2):129–137.

Long, M., Wang, J., Ding, G., Sun, J., and Yu, P. S. (2013). Transfer feature learning with joint distribution adaptation. pages 2200–2207. IEEE.

Long, M., Wang, J., Ding, G., Sun, J., and Yu, P. S. (2014). Transfer joint matching for unsupervised domain adaptation. pages 1410–1417. IEEE.

Lou, Y., Li, S., Liu, N., and Liu, R. (2022). Seismic volumetric dip estimation via a supervised deep learning model by integrating realistic synthetic data sets. *Journal of Petroleum Science and Engineering*, 218.

Lu, C., Jiang, H., Yang, J., Wang, Z., Zhang, M., and Li, J. (2022). Shale oil production prediction and fracturing optimization based on machine learning. *Journal of Petroleum Science and Engineering*, 217.

Ma, Y., Liang, F., Zhu, M., Chen, C., Chen, C., and Lv, X. (2022). Ft-ir combined with pso-cnn algorithm for rapid screening of cervical tumors. *Photodiagnosis and Photodynamic Therapy*, 39.

Magurran, A. E. (2004). *Measuring biological diversity: Rejoinder*.

Maharjan, S., Guidio, B., Fathi, A., and Jeong, C. (2022). Deep and convolutional neural networks for identifying vertically-propagating incoming seismic wave motion into a heterogeneous, damped soil column. *Soil Dynamics and Earthquake Engineering*, 162.

Mehmet Bilal, E. R. (2021). Heart sounds classification using convolutional neural network with 1D-local binary pattern and 1D-local ternary pattern features. *Applied Acoustics*, 180:108152.

Meng, F., Fan, Q., and Li, Y. (2022). Self-supervised learning for seismic data reconstruction and denoising. *IEEE Geoscience and Remote Sensing Letters*, 19:1–5.

Mousavi, S. M., Zhu, W., Ellsworth, W., and Beroza, G. (2019). Unsupervised clustering of seismic signals using deep convolutional autoencoders. *IEEE Geoscience and Remote Sensing Letters*, 16:1693–1697.

Muisyo, I. N., Muriithi, C. M., and Kamau, S. I. (2022). Enhancing low voltage ride through capability of grid connected dfig based wecs using wca-pso tuned statcom controller. *Heliyon*, 8.

Mustafa, A. and AlRegib, G. (2021). Man-recon: Manifold learning for reconstruction with deep autoencoder for smart seismic interpretation. pages 2953–2957. IEEE.

Nanda, N. C. (2016). *Seismic Data Interpretation and Evaluation for Hydrocarbon Exploration and Production*. Springer International Publishing.

Nematzadeh, S., Kiani, F., Torkamanian-Afshar, M., and Aydin, N. (2022). Tuning hyperparameters of machine learning algorithms and deep neural networks using metaheuristics: A bioinformatics study on biomedical and biological cases. *Computational Biology and Chemistry*, 97.

Nikolopoulos, S., Kalogeris, I., and Papadopoulos, V. (2022). Non-intrusive surrogate modeling for parametrized time-dependent partial differential equations using convolutional autoencoders. *Engineering Applications of Artificial Intelligence*, 109.

Osgood, B. (2019). *Lectures on the Fourier Transform and Its Applications*. Pure and Applied Undergraduate Texts. American Mathematical Society.

Pan, S., Chen, K., Chen, J., Qin, Z., Cui, Q., and Li, J. (2020). A partial convolution-based deep-learning network for seismic data regularization1. *Computers and Geosciences*, 145.

Pan, S. J. and Yang, Q. (2010). A survey on transfer learning.

Pan, W., Torres-Verdín, C., and Pyrcz, M. J. (2021a). Stochastic pix2pix: A new machine learning method for geophysical and well conditioning of

rule-based channel reservoir models. *Natural Resources Research*, 30:1319–1345.

Pan, Z., Hu, S., Wu, X., and Wang, P. (2021b). Adaptive center pixel selection strategy in Local Binary Pattern for texture classification. *Expert Systems with Applications*, 180(March 2020):115123.

Powers, D. M. W. (2011). Evaluation: From precision, recall and f-measure to roc., informedness, markedness & correlation. *Journal of Machine Learning Technologies*, 2(1):37–63.

Pradhan, A. and Mukerji, T. (2020). Seismic bayesian evidential learning: estimation and uncertainty quantification of sub-resolution reservoir properties. *Computational Geosciences*, 24:1121–1140.

Qi, Y., Yang, L., Liu, B., Liu, L., Liu, Y., Zheng, Q., Liu, D., and Luo, J. (2021). Accurate diagnosis of lung tissues for 2D Raman spectrogram by deep learning based on short-time Fourier transform. *Analytica Chimica Acta*, 1179:338821.

Qian, F., Liu, Z., Wang, Y., Liao, S., Pan, S., and Hu, G. (2022). Dtae: Deep tensor autoencoder for 3-d seismic data interpolation. *IEEE Transactions on Geoscience and Remote Sensing*, 60.

Qian, F., Yin, M., Liu, X.-Y., Wang, Y.-J., Lu, C., and Hu, G.-M. (2018). Unsupervised seismic facies analysis via deep convolutional autoencoders. *GEOPHYSICS*, 83:A39–A43.

Quamer Nasim, M., Maiti, T., Rifat Arefin, M., Mei, J., and Singh, T. (2020). Seismic Facies Analysis : A deep domain adaptation approach. *Ieee Transactions on Geoscience and Remote Sensing*, pages 1–15.

Ranjbar, I. and Toufigh, V. (2022). Deep long short-term memory (lstm) networks for ultrasonic-based distributed damage assessment in concrete. *Cement and Concrete Research*, 162.

Rezatofighi, H., Tsoi, N., Gwak, J., Sadeghian, A., Reid, I., and Savarese, S. (2019). Generalized intersection over union: A metric and a loss for bounding box regression. *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 2019-June:658–666.

Rollmann, K., Soriano-Vargas, A., Cirne, M., Davolio, A., Schiozer, D. J., and Rocha, A. (2021). A three-way convolutional network to compare 4d

seismic data and reservoir simulation models in different domains. *Journal of Petroleum Science and Engineering*, page 109260.

Rollmann, K., Soriano-Vargas, A., Cirne, M., Davolio, A., Schiozer, D. J., and Rocha, A. (2022). A three-way convolutional network to compare 4d seismic data and reservoir simulation models in different domains. *Journal of Petroleum Science and Engineering*, 208:109260.

Rousseeuw, P. J. (1987). Silhouettes: a graphical aid to the interpretation and validation of cluster analysis.

Saito, K., Yamamoto, S., Ushiku, Y., and Harada, T. (2018). Open set domain adaptation by backpropagation.

Sammut, C. and Webb, G. I. (2017a). *Encyclopedia of Machine Learning and Data Mining*. Springer US.

Sammut, C. and Webb, G. I. (2017b). *Encyclopedia of Machine Learning and Data Mining*. Springer Publishing Company, Incorporated, 2nd edition.

Sandwell, D. T. (2021). *Advanced Geodynamics*. Cambridge University Press.

Sanodiya, R. K., Mathew, J., Aditya, R., Jacob, A., and Nayanar, B. (2021). Kernelized Unified Domain Adaptation on Geometrical Manifolds. *Expert Systems with Applications*, 167(January 2020):114078.

Sarhan, M. A. and Safa, M. G. (2019). 2D seismic interpretation and hydrocarbon prospects for the Neogene-Quaternary succession in the Temsah Field, offshore Nile Delta Basin, Egypt. *Journal of African Earth Sciences*, 155(February):1–12.

Sepulveda, L. F. M., Gattass, M., Silva, A. C., Quevedo, R., Michelon, D., Siedschlag, C., and Ribeiro, R. (2023a). Generalization of deep learning models for natural gas indication in 2d seismic data. *Pattern Recognition*, 141.

Sepulveda, L. F. M., Gattass, M., Silva, A. C., Quevedo, R., Michelon, D., Siedschlag, C., and Ribeiro, R. (2023b). Seismic data classification for natural gas detection using training dataset recommendation and deep learning. *Geoenergy Science and Engineering*, 228.

Shi, L., Gong, J., and Zhai, C. (2022). Application of a hybrid pso-ga optimization algorithm in determining pyrolysis kinetics of biomass. *Fuel*, 323.

Shu, X., Song, Z., Shi, J., Huang, S., and Wu, X. J. (2021). Multiple channels local binary pattern for color texture representation and classification. *Signal Processing: Image Communication*, 98(July):116392.

Soudani, A. and Barhoumi, W. (2019). An image-based segmentation recommender using crowdsourcing and transfer learning for skin lesion extraction. *Expert Systems with Applications*, 118:400–410.

Sudharshan, P. J., Petitjean, C., Spanhol, F., Oliveira, L. E., Heutte, L., and Honeine, P. (2019). Multiple instance learning for histopathological breast cancer image classification. *Expert Systems with Applications*, 117:103–111.

Theodoridis, S. (2020). *Machine Learning*. Elsevier.

Trani, L., Pagani, G. A., Zanetti, J. P. P., Chapeland, C., and Evers, L. (2022). Deepquake — an application of cnn for seismo-acoustic event classification in the netherlands. *Computers Geosciences*, 159:104980.

Trapp, M., Bogoclu, C., Nestorović, T., and Roos, D. (2019). Intelligent optimization and machine learning algorithms for structural anomaly detection using seismic signals. *Mechanical Systems and Signal Processing*, 133:106250.

Veillard, A., Morère, O., Grout, M., and Gruffeille, J. (2018). Fast 3d seismic interpretation with unsupervised deep learning: Application to a potash network in the north sea. volume 2018, pages 1–5. European Association of Geoscientists and Engineers, EAGE.

Venkateswara, H., Lade, P., Ye, J., and Panchanathan, S. (2015). Coupled support vector machines for supervised domain adaptation. pages 1295–1298. ACM.

Venkateswara, H. and Panchanathan, S. (2020). *Domain Adaptation in Computer Vision with Deep Learning*. Springer International Publishing.

Wang, Y., Wang, B., Tu, N., and Geng, J. (2020a). Seismic trace interpolation for irregularly spatial sampled data using convolutional autoencoder. *GEOPHYSICS*, 85:V119–V130.

Wang, Y., Wang, C., Xue, H., and Chen, S. (2022). Self-corrected unsupervised domain adaptation. *Frontiers of Computer Science*, 16(5).

Wang, Z., Di, H., Shafiq, M. A., Alaudah, Y., and Alregib, G. (2018). Successful leveraging of image processing and machine learning in seismic structural interpretation: A review. *https://doi.org/10.1190/tle37060451.1*, 37:451–461.

Wang, Z., Du, B., and Guo, Y. (2020b). Domain adaptation with neural embedding matching. *IEEE Transactions on Neural Networks and Learning Systems*, 31:2387–2397.

Waqas, U. and Ahmed, M. F. (2022). Investigation of strength behavior of thermally deteriorated sedimentary rocks subjected to dynamic cyclic loading. *International Journal of Rock Mechanics and Mining Sciences*, 158.

Wu, T., Shen, L., and Xu, Y. (2021). Fixed-point proximity algorithms solving an incomplete Fourier transform model for seismic wavefield modeling. *Journal of Computational and Applied Mathematics*, 385:113208.

Xu, P., Lu, W., and Wang, B. (2019). A semi-supervised learning framework for gas chimney detection based on sparse autoencoder and tsvm. *Journal of Geophysics and Engineering*, 16:52–61.

Xu, T., Zhang, W., Li, J., Liu, H., Kang, Y., and Lv, W. (2022). Domain generalization using contrastive domain discrepancy optimization for interpretation-while-drilling.

Yang, N., Zhang, Z., Yang, J., and Hong, Z. (2022). Mineral prospectivity prediction by integration of convolutional autoencoder network and random forest. *Natural Resources Research*, 31:1103–1119.

Yu, Y., Gong, Z., Wang, C., and Zhong, P. (2017). An Unsupervised Convolutional Feature Fusion Network for Deep Representation of Remote Sensing Images. *IEEE Geoscience and Remote Sensing Letters*, pages 1–5.

Zhang, G., Lin, C., Ren, L., Li, S., Cui, S., Wang, K., and Sun, Y. (2022a). Seismic characterization of deeply buried paleocaves based on bayesian deep learning. *Journal of Natural Gas Science and Engineering*, 97.

Zhang, K., Lin, N., Tian, G., Yang, J., Wang, D., and Jin, Z. (2022b). Unsupervised-learning based self-organizing neural network using multi-component seismic data: Application to xujiahe tight-sand gas reservoir in china. *Journal of Petroleum Science and Engineering*, 209:109964.

Zhang, S. B., Si, H. J., Wu, X. M., and Yan, S. S. (2022c). A comparison of deep learning methods for seismic impedance inversion. *Petroleum Science*, 19:1019–1030.

Zhao, S., Li, B., Reed, C., Xu, P., and Keutzer, K. (2020). Multi-source domain adaptation in the deep learning era: A systematic survey.

Zhao, Y. X., Li, Y., and Wu, N. (2022). Unsupervised dual learning for seismic data denoising in the absence of labelled data. *Geophysical Prospecting*, 70:262–279.

Zhu, W., Mousavi, S. M., and Beroza, G. C. (2020). Seismic signal augmentation to improve generalization of deep neural networks.