Anomaly Detection Using Threshold-Based Approach in Seismic Data

Muhammad Talha Khan, Syed Maisum Abbas Rizvi, Hafiz Muhammad Ibraheem

Abstract—**The analysis of seismic data is crucial in geological exploration specifically in oil and gas industry where noise-free and high-resolution imaging is important for useful resource extraction. However, issues such as noise contamination, frequency anomalies and data incompleteness often complicate the seismic interpretations. The study focuses on detecting the frequency anomalies introduced during compensation process. This is a critical step that can generate the frequency artifacts that affect the seismic data integrity. The research aims to provide an automated anomaly detection approach to identify and mitigate it, ultimately, enhancing the reliability of seismic data interpretations. To detect the anomalies in data, threshold-based approach is used in which lower and upper threshold is defined as normal data boundary, others are considered as anomalies. After that, decision tree supervised machine learning algorithm is used to train and test the model which performed well as compared to the rest of the models for anomaly detection during training and testing in both phases. In future, the type-based anomaly detection can be considered for further research purposes.**

Index Terms—**seismic data, anomaly detection, oil and gas, threshold-based approach, decision tree, supervised machine learning, classification**

I. INTRODUCTION

The seismic-based data plays an important role in geological exploration, especially for oil and gas industries, where the importance of this has a huge impact on extracting an effective resource. The current challenges of resolution enhancement and reducing seismic data noise are due to limitations such as incomplete data and noise contamination. The recent advancements in the field especially with Shearlet Denoising and Compressed Sensing (CS) Techniques have shown significant progress in addressing these specific issues.

The Compressed Sensing theory processes data by enabling signal reconstruction using the Nyquist theorem which uses fewer measurements as compared to traditionally required which is basically a transformative approach. This is good for parse signals such as in cases like getting seismic-reflection coefficients, where the data acquisition can be time-sensitive and costly. This has proven effective in different studies for improving the overall reconstruction of lost data during acquisition along with frequency coverage [1], [3].

Furthermore, to preserve the characteristics of a signal while maintaining noise minimization, denoising is a critical aspect that needs to be focused on. The Shearlet transformation is here as a superior way that uses directional capabilities along with multi-scaling which ultimately provides a robust framework to decompose and denoise the frequency [1].

Even with these advancements, the challenges persist, especially related to introducing frequency anomalies that appear during the frequency compensation process. This stage in which the artifacts that get generated can affect the seismic interpretation integrity ultimately leading to skewed-based geological assessments. This paper aims to reduce the gap by focusing on the frequency anomaly during the frequency compensated process based on seismic data. The primary aim is to come up with a method and test an anomaly detection method that can be able to detect frequency artifacts efficiently, ultimately improving the reliability of exploratory seismic data.

II. LITERATURE REVIEW

Various studies have put an emphasis on high-resolution based subsurface imaging focusing on wide-bandwidth seismic data. The low-frequency components are important for subsurface structure imaging accuracy and deeper penetration of subsurface structures. On the other hand, high-frequency based components contribute to the improvements of resolution and detailed imaging [1], [2]. The methods used for enhancing the seismic data such as Q-filtering have been utilized but usually struggle with the stability and preservation of the signal integrity [1].

CS has revolutionized the seismic data acquisition and its processing by doing the data recovery from under-sampled data. This is beneficial in reducing the data acquisition costs and also overcoming the practical limitations posed via traditional sampling rates [1]. The foundation for CS lies in the reconstruction of signals that are sparse in a transform domain. For seismic-based applications, this also signal can be reconstructed accurately by optimizing the coefficients of the L1 norm, given that the data meet the Restricted Isometry Property (RIP) [1].

The studies have given a demonstration of the CS application for merging the low-frequency-based extensions along with the original high-frequency data, that provides an output in enhanced bandwidth of seismic data [1]. In one study, it extended this approach by CS integration with the denoising algorithm to prove the feasibility of using CS for the comprehensive frequency compensation [1]. However, these techniques can introduce frequencies that can complicate the seismic interpretation.

Denoising is an important step when it comes to seismic-based data preprocessing as residual noise can affect analysis quality, especially the frequency compensation. Shearlet transform has been highly effective for this very purpose, which outperforms the traditional wavelet and curvelet methods. Also, unlike the wavelets, which are only used to represent the isotropic features, this excels in capturing the anisotropic structures along with directional features of the seismic-based data [1], [2]. This property also makes it viable for enhancing the seismic signal sparsity and mitigating the noise more effectively.

The studies comparing the curvelets and wavelet transformations have also confirmed that the Shearlet transformation provides superior denoising and preserves the critical signal components. Like, a study highlighted the robustness of noise suppression and retention by employing denoising combined with CS for frequency compensation [1].

Even though the Shearlet denoising and CS methods provide substantial advantages in seismic-based data preprocessing, detecting anomalies still remains persistent. These anomalies can be set as artifacts during the frequency compensation process which may to misinterpretation of geological structures. Being able to identify such anomalies effectively can be critical to ensure data quality and reliability. The research that is present in this paper focuses on developing a framework that is tailored to seismic data.

The study utilizes the Facies Classification Benchmark dataset [4] using the SEGY format to detect anomalies in seismic data. Using the Decision Tree Algorithm (DTA), the data will be analyzed for its potential to enhance anomaly detection and to contribute towards an accurate geological interpretation.

III. Research Methodology

This study goal is to identify the anomalies in seismic dataset for accurate analysis and interpretation, by implementing supervised machine learning using threshold-based approach via decision tree model. Machine learning technique, like Decision Trees, are well-suited for classification problem and non-linear relationship. Along with that, this technique is suited to handle complex dataset like seismic data. Additionally, SMOTE (Synthetic Minority Oversampling Technique) is used to balance the dataset.

The dataset (Facies Classification Benchmark - SEGY) [4] which is used in this research is 3d seismic data from Netherlands F3 block, consists of 3 dimensions (Number of crosslines x Number of in-lines (range) x Number of time samples) having values (255 x 901 x 601). This dataset is then converted into 2d (Number of crossline x Number of time sample). For better understanding, here 20 in-line slices are considered for this research and each slice consists of 255 x 601.

To detect anomalies in data, threshold-based approach is applied and determine anomalies while the remaining are normal in this selected 20 in-line slices. Those points that were not lie in given upper and lower threshold are anomalies. By using this approach, 5.64% data points are determined as anomalies.

As the data was unbalanced so, the SMOTE is used to balance both classes by resampling the actual data. It ensures that all classes are balanced properly for training the machine learning model.

To train the model, balanced dataset is split into training and testing, for both x and y (normal and anomaly) classes with ratio of 80 and 20.

A Decision Tree Classifier is implemented to train the model because of its accountability and strongest over large data.

Confusion Matrix and Performance Matrix that includes Accuracy, Precision, Recall, Specifity, and F1-Score (prioritized for anomaly detection as it gives the highlight between precision and recall) are used to monitor the model's performance for both training and testing scenarios.

The challenging point was that the resampling using SMOTE could not generate unrealistic pattern for given data.

This experiment was performed at a Google Colab Platform. Python 3.0 is used for code writing with its libraries like NumPy, pandas, seaborn, scikit-learn, matplotlib, and segyio to apply preprocessing, visualization, and performing seismic operations and so on.

IV. Results And Discussions

The proposed approach brings the accuracy of 96%, average precision of 94%, recall of 98% and F1-score of 96% in the test dataset. Table 1 and Table 2 shows the results of training and Testing datasets. Fig. 1 shows the detection of anomalies align with the region of high amplitude fluctuations. Such findings indicate the irregularities (faults, fractures) in the subsurface based of prior studies.

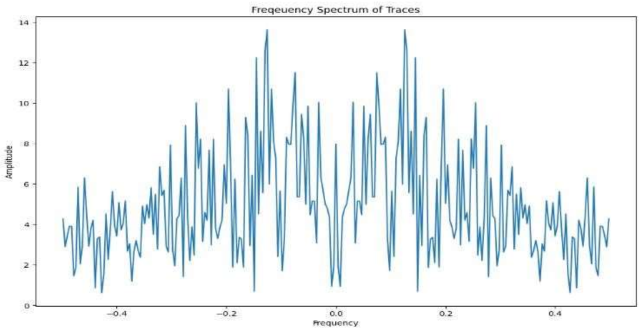

Fig. 1. Frequency Spectrum of Traces

The amplitude distribution is mostly normal that it can be visualize in Fig. 2.

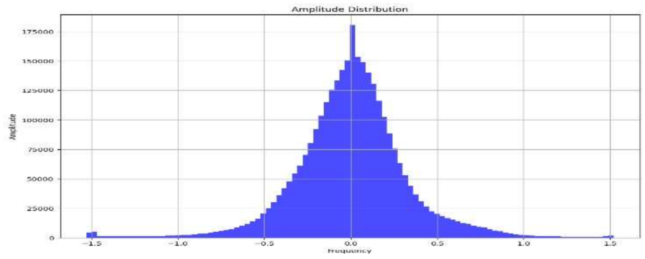

Fig. 2. Amplitude Distribution

The count of anomalies detected across each trace in selected 20 in-line slices starting from 101 to 120 are shown in Fig. 3.

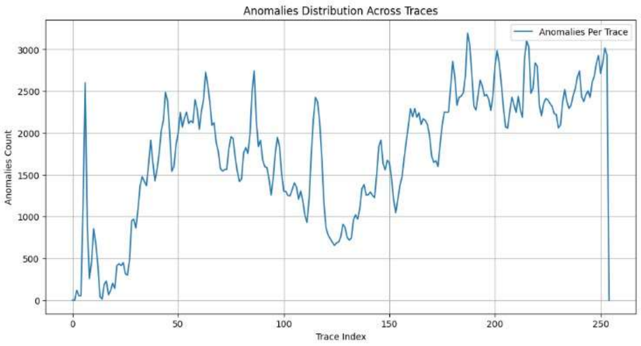

Fig. 3. Anomalies Distribution Across Traces

The count of anomalies detected across each time sample in selected 20 in-line slices starting from 101 to 120 are shown in Fig. 4.

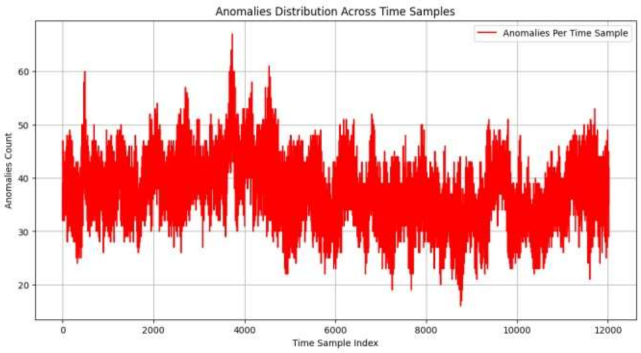

Fig. 4. Anomalies Distribution Across Time Samples

Relative to traditional statistical methods at threshold-based anomaly detection, the proposed approach proved much more sensitive to even subtle anomalies when applied in complex geological formations. The model re-discovered anomalies that could not be detected using conventional methods on all traces, suggesting that these improvement prospect to establish more accurate exploration. Fig. 5 shows detected anomalies in selected in-line slices data. It includes 20 in-line slices which has 250 traces and had 601 times samples per in-line slice.

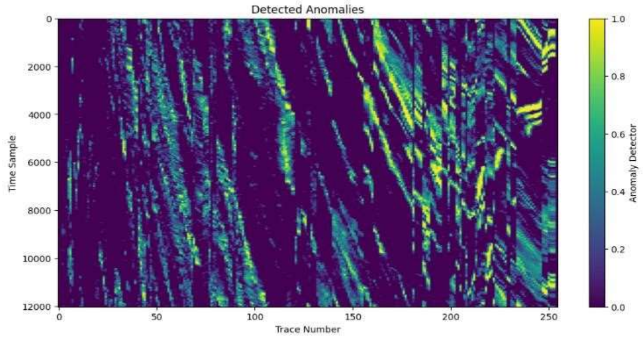

Fig. 5. Detected Anomalies

As the data was imbalanced, that can be seen in Fig. 6. There were just 14.68% anomalies detected in the training data. Ultimately the results were so poor because of mismatch data size in both classes. Therefore, SMOTE was applied to balance the normal and anomaly classes.

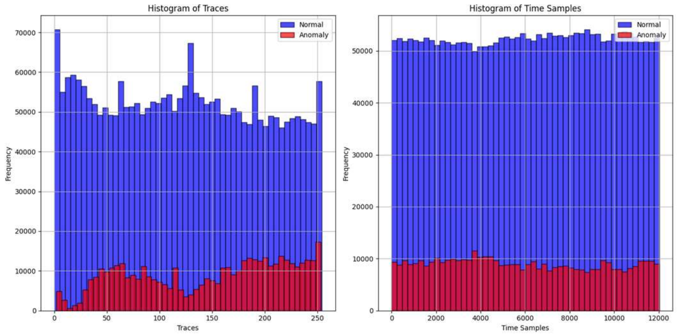

Fig. 6. Before SMOTE: Data Distribution

After applying SMOTE, data distribution is again visualized and anomaly percentage is calculated to check whether the both classes are balanced or not. Now, the training data is balanced by 50% normal and 50% anomaly. Fig. 7 shows the data distribution after applying SMOTE.

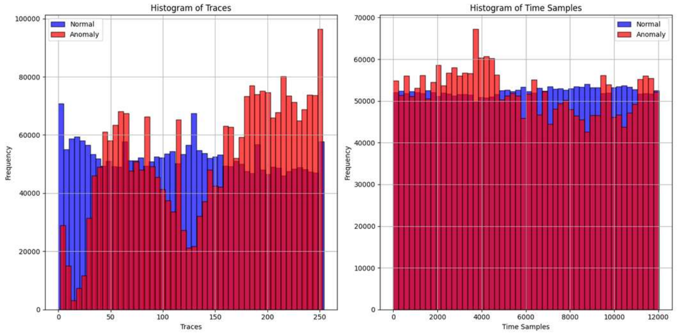

Fig. 7. After SMOTE: Data Distribution

Applying the Decision Tree supervised machine leaning model, it performed well during training, results are shown in Fig. 8 and Performance Measures are shown in Table 1.

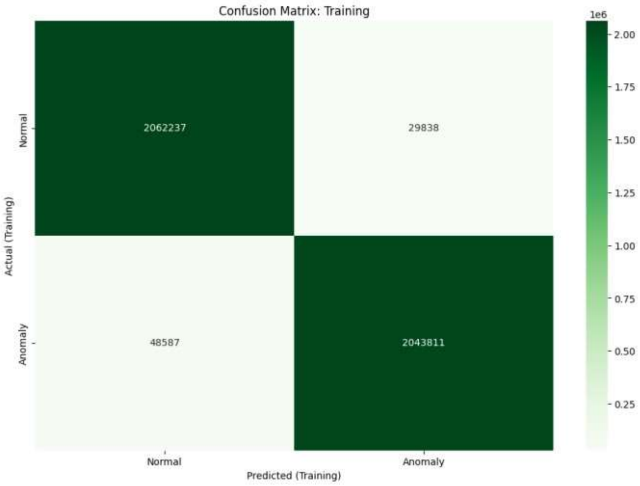

Fig. 8. Confusion Matrix: Training

Results of Training Dataset	
Accuracy	98%
Precision	99%
Recall	98%
Specifity	99%
F1-Score	98%

Table 1. Results of Training Dataset

After training, the model is tested on unseen dataset, it performed good as expected during testing, results are shown in Fig. 9 and Performance Measures are shown in Table 2.

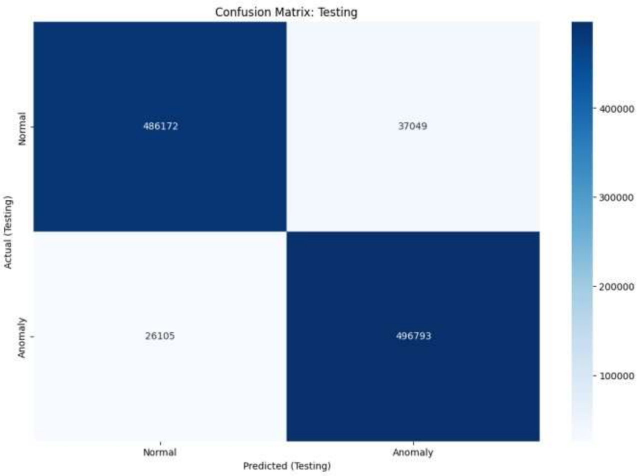

Fig. 9. Confusion Matrix: Testing

Results of Testing Dataset	
Accuracy	94%
Precision	93%
Recall	95%
Specifity	93%
F1-Score	94%

Table 2. Results of Testing Dataset

Performance Measures shown that the model is useful for detecting anomalies in large data like seismic data as it performed very well to identify both of the classes "Normal" and "Anomaly" with high accuracy and stability. Table 3 shows the final Testing Data Classification Report.

Testing Data Classification Report				
	Precision	Recall	F1-Score	Support
Normal	95%	93%	94%	523221
Anomaly	93%	95%	94%	522898
Accuracy			94%	1046119
Macro Avg.	94%	94%	94%	1046119
Weighted Avg.	94%	94%	94%	1046119

Table 3. Testing Data Classification Report

V. Conclusion

This research focused on the development and the assessment for anomaly detection model. The data was focused on seismic activity specifically on oil and gas to address the associated challenges with processing large datasets. The metrics achieved in this research highlights the ability of the model to detect anomalies in a consistent manner while also minimizing the misclassifications. The results also put focus on the sensitivity of the model in seismic traces for refined variations which plays a critical role in identifying the subsurface irregularities.

Some limitations were also observed even with model's effectiveness. Implementing advanced data techniques such as adaptive noise filtering in frequency domain or feature extraction could possibly enhance the anomaly detection. Furthermore, integrating the domain-based knowledge to the model could also reduce noise-based errors and improve it further.

The approach based on Machine Learning proved to be more superior in capturing the complex patterns as compared to the traditional threshold-based approach for geologically diverse regions. This makes the model a useful tool to detect anomalies in real-time seismic exploration enabling better and accurate results.

Future research can focus on type-based classification of anomalies along with scaling the model to larger datasets to explore the application in other domains.

References

[1] D. Wang *et al.*, "Anti-noise Full Frequency Expansion for Seismic Data with Compressed Sensing," *IEEE Transactions on Geoscience and Remote Sensing*, 2024, doi: 10.1109/TGRS.2024.3471815.

[2] M. H. Rahma Putra, M. Hermana, I. B. S. Yogi, T. M. Hossain, M. F. Abdurrachman, and S. J. A. Kadir, "Reservoir porosity assessment and anomaly identification from seismic attributes using Gaussian process machine learning," *Earth Sci Inform*, vol. 17, no. 2, pp. 1315–1327, Apr. 2024, doi: 10.1007/s12145-024-01240-7.

[3] S. Tatyana, S. Alexander, S. Maria, and C. Boris, "Seismic Anomalies in the Geothermal District Revealed by the Relaxation Algorithm of Selected Coordinate Descent," in *Proceedings of 2021 14th International Conference Management of Large-Scale System Development, MLSD 2021*, Institute of Electrical and Electronics Engineers Inc., 2021. doi: 10.1109/MLSD52249.2021.9600169.

[4] Yazeed Alaudah, Patrycja Michałowicz, Motaz Alfarraj, and Ghassan AlRegib, (2019), "A machinelearning benchmark for facies classification," Interpretation 7: SE175-SE187