Coastal CTD Vertical Profile Outlier Detection by AI

Young Min Choi[1], Chaewook Lim[2], Jin Il Song[3], Seung-Buhm Woo[1,*]
[1]Department of Ocean Sciences, Inha University, South Korea, [2]Artificial Intelligence Convergence Research Center, Inha University
[3]Gyeonggi-Incheon SeaGrant Center, South Korea
E-mail: ymchoi@mpl.inha.ac.kr, sbwoo@inha.ac.kr*

1 Introduction

- CTD observations provide vertical profiles of temperature and salinity, serving as essential data for analyzing water masses, mixing, and density structures.
- In practice, CTD data often contain errors due to sensor instability, operator mistakes, or contact with the seafloor. Systematic error detection is essential for quality control of oceanographic data.
- With advances in AI, machine learning-based outlier detection methods have gained attention. Among them, Isolation Forest (IF) is efficient due to its simple structure and fast computation. This study aims to improve detection accuracy by refining IF-based error detection through preprocessing.

2 Methodology



Figure 1. Location of the CTD casting observation stations. (Incheon, New Port)

- This study uses 344 CTD casts collected off New Incheon Port, with 10 to 20 observations conducted at each casting point.
- Each CTD profile includes depth, temperature, salinity, and turbidity, with frequent errors near the surface and bottom due to sensor instability and seafloor contact.

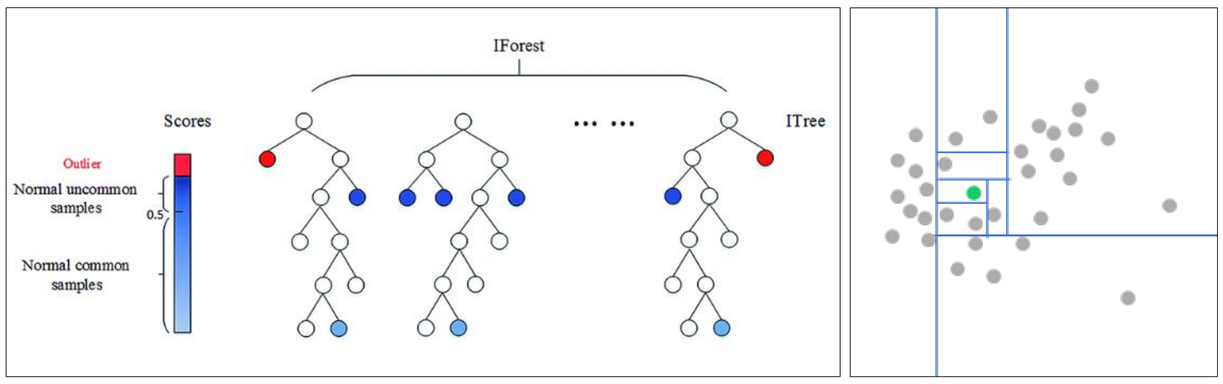


Figure 2. Schematic Diagram of Isolation Forest.

- The Isolation Forest (IF) is an unsupervised model that isolates anomalies by randomly partitioning data using multiple decision trees, without predefined error labels.
- In this study, both global and local anomaly scores were combined to detect outliers in CTD data. Hyperparameters were set empirically, and the anomaly threshold was determined based on the known error ratio.

Pres(P)	Density
Temperature(f)	Feature_diff
Salinity(f)	Feature_grad
Turbidity(f)	Feature_curv

$$\Delta f = f_i - f_{i-1} \qquad (feature_diff)$$

$$f' = \frac{f_i - f_{i-1}}{P_i - P_{i-1} + \varepsilon} \qquad (feature_grad)$$

$$\kappa = \frac{f''}{(1 + (f')^2)^{3/2} + \varepsilon} \qquad (feature_curv)$$

$$where \quad f'' = \frac{f_i' - f_{i-1}'}{\Delta P}$$

Table 1. Feature Information of IF model. (Blue text applied only to the PIF2 model)

- The accuracy of the Standard Isolation Forest (SIF) without preprocessing was compared to the Processed Isolation Forest (PIF) with preprocessing.
- The preprocessing steps included threshold filtering, addition of derived features, and removal of the initial surface segment to enhance model reliability.
- To identify features that contribute most to anomaly detection, permutation importance was calculated by randomly shuffling feature values and evaluating the resulting change in prediction performance.

3 Results

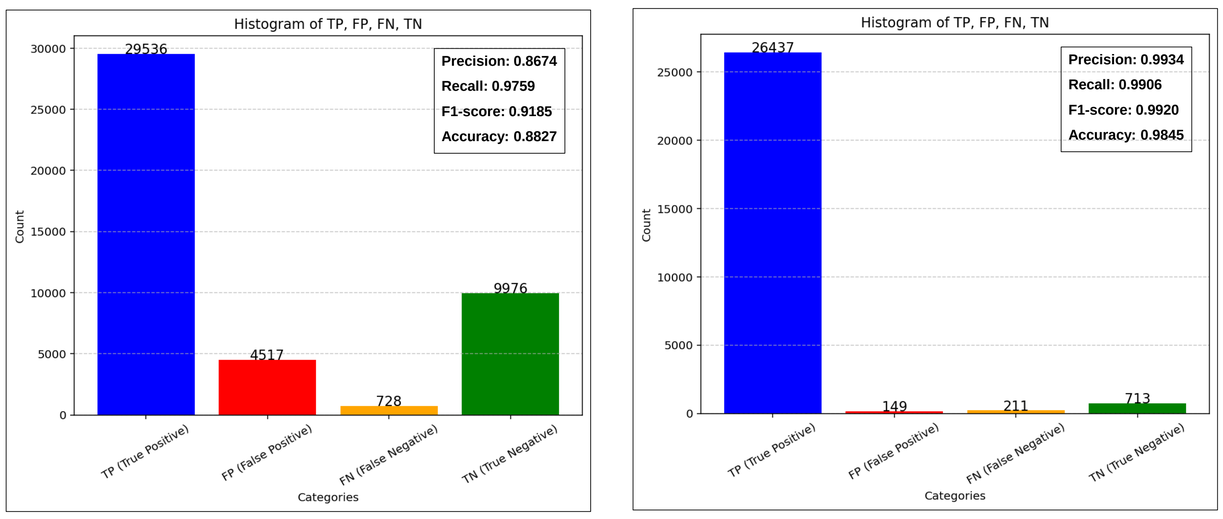


Figure 3-a, 3-b. Result of SIF(Left) and PIF1(Right) Model (accuracy histogram)

	Precision	Recall	F1-score	Accuracy
SIF	0.8674	0.9759	0.9185	0.8827
PIF1	0.9944	0.9921	0.9932	0.9869
PIF2	0.9934	0.9906	0.9920	0.9845

Table2. Accuracy Comparison for Model Result

- A comparison of anomaly detection performance showed that the PIF model outperformed the standard SIF model. While SIF detected approximately 90% of anomalies, PIF achieved a detection rate of around 98%.

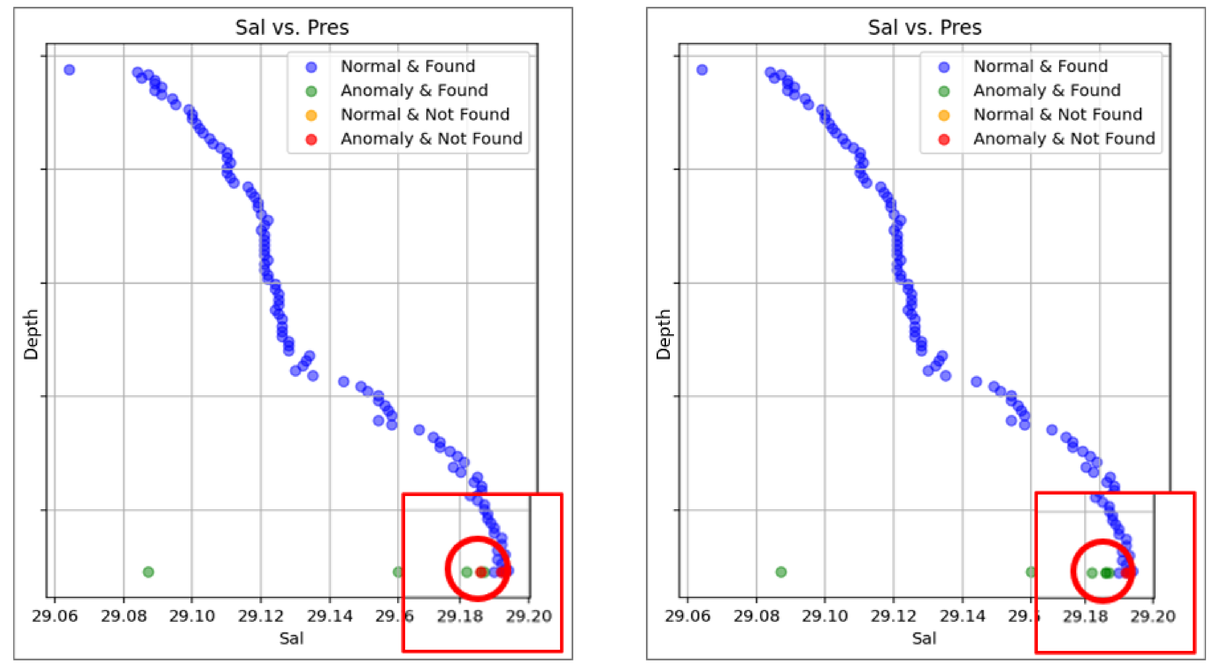


Figure 4-a, 4-b. Results of SIF(Left) and PIF1(Right) Model (Example Profile)

- In an example profile comparison, the SIF model failed to detect bottom-layer anomalies, whereas the PIF model successfully identified them.

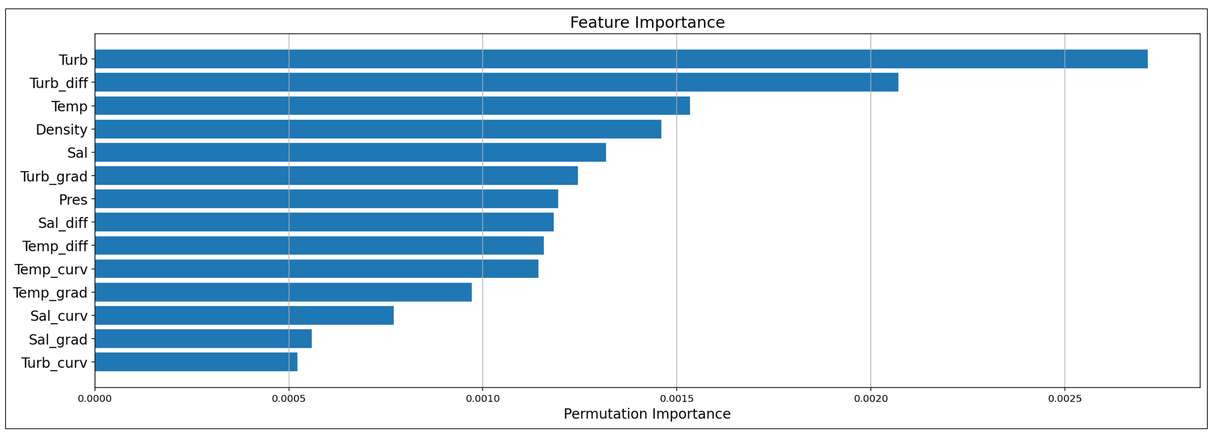


Figure 5. Permutation Importance of Feature

- An evaluation of feature importance revealed that primary variables such as temperature, salinity, and turbidity had the greatest influence on anomaly detection, while curvature-based features had minimal impact.

4 Conclusions

- The proposed model demonstrated high accuracy and efficiency in detecting anomalies from CTD vertical profiles collected in coastal waters, achieving an anomaly detection rate of approximately 98%.
- Primary variables such as temperature, salinity, and turbidity had the greatest influence on performance, while derived features like curvature showed limited impact.
- The model effectively detected frequent surface and bottom anomalies typical of coastal observations, ensuring high detection accuracy.

이 논문은 2024년도 해양수산부 재원으로 해양수산과학기술진흥원의 지원을 받아 수행된 연구임 (RS-2022-KS221548, 경기인천씨그랜트). / 이 논문은 2024년도 정부(과학기술정보통신부)의 재원으로 정보통신기획평가원의 지원을 받아 수행된 연구임 (NO.RS-2022-00155915, 인공지능융합혁신인재양성사업(인하대학교)).