Data-centric AI practice in maritime: securing trusted data quality via a computer vision-based framework

Ke Wang[1], Ong Qi Hao Tristan[2], Xiaocai Zhang[1]⋆, Xiuju Fu[1], and Zheng Qin[1]

***Abstract*—The advancement of data-driven Artificial Intelligence (AI) applications becomes increasingly important in optimizing maritime operations. Establishing trustworthy decision-making processes hinges upon precise diagnosis of data quality—a fundamental prerequisite. However, prevalent statistical-based methodologies encounter inherent challenges, such as requiring precise threshold settings, overlooking contemporary insights from recent data, etc. To address these challenges, this study proposes a vision-inspired framework for the classification of Automatic Identification System (AIS) data quality issues, particularly within highly imbalanced datasets. Four typical data quality issues are included in this study, namely *Zig-zag Value*, *Identity Theft*, *Temporal Missing*, and *Abnormal Constant Value*. The overall framework includes a graphical transformation process to represent the spatial information of the trajectories, a data augmentation process to mitigate the class imbalance issue, and a deep learning model for image classification. Extensive experiments show that 1)The proposed method could achieve an impressive 99.29% accuracy and 99.27% F1 score in AIS data quality issue classification; 2)The ConvNeXtV2 model, an enhanced convolutional neural network, demonstrated its superiority in this application, overtaking other state-of-the-art models by 2.14% in accuracy, 2.35% in F1 score, and 3.40% in MCC; 3) The MixUp-based data augmentation method outperformed other imbalance learning strategies such as CutOut, Focal Loss, WeightedLoss, etc. As one of the first few practices on data-centric AI in the maritime sector, this study promises to notably reinforce maritime data reliability, fostering enhanced decision-making processes industry-wide.**

***Index Terms*—data-centric AI, image classification, AIS data quality, imbalance learning, data augmentation**

I. Introduction

The maritime industry increasingly relies on data-driven Artificial Intelligence (AI) to optimize operation efficiency and safety measures. However, the foundational infrastructure of this digitalization, data, commonly encounters quality issues stemming from sensor failures and human errors [1]. Studies indicate that a substantial 92% of AI practitioners face data issues, impacting the reliability of AI applications [2], [3]. Early detection of data quality issues could mitigate such risks and prevent unnecessary costs. A wide range of data quality issues have been identified within Automatic Identification System (AIS) data, a fundamental data type in the maritime sector [1]. For example, *Temporal Missing* arises with large time gaps between successive data points; *Identity Theft* represents the scenario where more than one ship shares the same MMSI at the same time; *Zig-zag Value* characterizes the fluctuations of attributes such as latitude, longitudes, etc; *Abnormal Constant Value* denotes a stagnant value that anomalously repeats a previous value. To detect these data quality issues, a common practice is to apply statistical-based approaches, as exemplified by [4] [5] [6]. These approaches, e.g., the 3-sigma principle, moving average method, are easy to implement whereas require dedicated parameter setting. Besides, they often disregard insights from the data itself as they impose statistical hypotheses on the dataset. Recent pioneering research has explored AIS data analysis using data-driven visual methods. For instance, [7] introduced a deep learning approach aimed at categorizing typical vessel activities from AIS data. [8] employed a CNN-based framework to differentiate vessel loitering behaviours based on the shape of trajectories, yielding encouraging outcomes. Embracing visual representations of trajectories, together with a data-driven method, holds the potential to unveil concealed data patterns [9]. To this end, this study delves into the potential of vision-based techniques to diagnose AIS data quality issues, via a "representation-augmentation-classification" framework. Leveraging an expert-labelled dataset, the proposed framework aims to identify the unique patterns among five types of AIS data: *Temporal Missing*, *Identity Theft*, *Zig-zag Value*, *Abnormal Constant Value*, and *High Data Quality*. Comprehensive ablation tests were conducted to test the efficacy of the proposed approach. The key contributions are outlined below:

- This study introduces a novel vision-inspired framework designed to effectively distinguish data containing four unique AIS quality issues from high-quality data.
- This study adopts a data-driven approach to diagnosing data quality issues coupled with an effective imbalance-learning strategy.
- This study conducts a comprehensive analysis of framework components, providing crucial insights for similar applications.

II. Methodology

Fig. 1 outlines the entire workflow of the proposed approach, aiming to classify five distinct data types: *Temporal*

*Corresponding author: Xiaocai Zhang.

[1]Ke Wang, Xiaocai Zhang, Xiuju Fu and Zheng Qin are with the Institute of High Performance Computing (IHPC), Agency for Science, Technology and Research (A*STAR), 1 Fusionopolis Way, #16-16 Connexis, Singapore 138632, Republic of Singapore wang_ke@ihpc.a-star.edu.sg; zhang_xiaocai@ihpc.a-star.edu.sgg

[2]Ong Qi Hao Tristan is with the School of Computing, Singapore Polytechnic, 500 Dover Road, Singapore 139651, Republic of Singapore

979-8-3503-5409-6/24/$31.00 ©2024 IEEE
DOI 10.1109/CAI59869.2024.00082

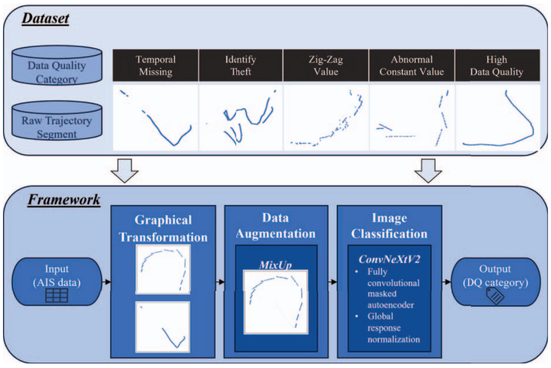

Fig. 1. The overall workflow of the "representation-augmentation-classification" framework designed for AIS data quality categorization.

Missing, *Identity Theft*, *Zig-zag Value*, *Abnormal Constant Value*, and *High Data Quality*. The methodology unfolds in three key stages: Firstly, AIS data is transformed into a graphical format to visually encapsulate its geospatial attributes; Subsequently, the MixUp technique is used to augment samples within minority classes to mitigate the class imbalance issue; Finally, ConvNeXtV2 model, recognized as the state-of-the-art CNN-based model for image classification and segmentation, is applied to categorize images into different types. Detailed components are introduced below:

A. Graphical transformation of AIS data

The AIS data undergoes a graphical transformation process. Through this process, every data record is symbolized as a point within an image, and all the points construct the complete trajectory. In such a way, the tabular data is structured and visualized in a graphic format, as depicted in Fig. 1. This is commonly applied in the AIS data mining process [8], [10], [11].

B. Data augmentation via MixUp technique

To rectify class imbalance issues involved in this application, a specialized data augmentation approach is implemented: oversampling through the MixUp technique. This method involves the synthetic generation of new samples to enrich the dataset's underrepresented classes, i.e., *Identity Theft* and *Zig-zag Value* in this context. Specifically, minority classes are augmented by the creation of a weighted combination of random image pairs. Given two images and their ground truth labels: (x_i^p, y_i^p),(x_j^q, y_j^q), a new synthetic sample is generated as:

$$\hat{x} = \lambda x_i^p + (1 - \lambda) x_j^q, \tag{1}$$

$$\hat{y} = \lambda y_i^p + (1 - \lambda) y_j^q. \tag{2}$$

Where, x, y denote the image and label of a sample, respectively. i, j refers to the index of the samples. p, q signify the respective classes to which the sample belongs. In this study, p, q are kept the same to generate new samples within these minority categories.

C. Image classification via ConvNeXtV2 model

At the final stage, the ConvNeXtV2 model is employed to categorize images into distinct data types. ConvNeXtV2 model is a pure convolutional model inspired by the design of Vision Transformer and a successor of ConvNeXt [12]. It integrates the fully convolutional masked autoencoder (FCMAE) and global response normalization (GRN) techniques to the original ConvNeXt model. Notably, ConvNeXtV2 has showcased its effectiveness in a spectrum of tasks, including ImageNet classification, COCO detection, ADE20K segmentation, etc [13].

III. EXPERIMENTS

A. Dataset

The efficacy of the proposed approach was evaluated using a real-world AIS dataset in Singapore, which was validated by domain experts. Raw vessel trajectory data was employed for the analysis to preserve intricate details and prevent loss of information. An 80%-20% ratio is used for the train-test split. Considering the class imbalance issue within the dataset, notably prevalent in the *Identity Theft* and *Zig-zag Value* categories, the MixUp-based data augmentation technique is applied to the two categories separately. The details of the dataset are illustrated in TABLE I.

TABLE I
DESCRIPTION OF DATASET

Category	Number of Samples		
	Original training set	Augmented training set	Testing set
ConstantValue	422	422	105
TemporalMisisng	553	553	137
Identity Theft	24	346	6
Zig-zag Value	281	561	70
High Data Quality	1537	1537	384

B. Evaluation metrics

The proposed approach is evaluated against a wide range of metrics , akin to those utilized in prior studies [14], including Area Under Receiver Operating Characteristic Curve (AUCROC), Matthews Correlation Coefficient (MCC), accuracy (ACC), and F1 Score.

The details are outlined below:

- Area Under Receiver Operating Characteristic Curve (AUCROC). It assesses a model's capability to differentiate between classes by plotting the true positive rate (TPR) against the false positive rate (FPR) at various threshold settings [15]. The TPR and FPR are computed using the following equations:

$$TPR = \frac{TP}{TP + FN}, \tag{3}$$

$$FPR = \frac{FP}{FP + TN}, \tag{4}$$

- Matthews Correlation Coefficient (MCC). This singular metric encapsulates information from the confusion matrix, providing an evaluation of the quality of multiclass classification. It considers true and false positives and negatives, calculated as:

$$MCC = \frac{TP \times TN - FP \times FN}{\sqrt{(TP+FP)(TP+FN)(TN+FP)(TN+FN)}}, \tag{5}$$

- Accuracy (ACC). It is a fundamental metric, that offers a straightforward assessment of the correctness of predicted results. It quantifies the ratio of correctly predicted instances to the total number of instances evaluated:

$$acc = \frac{TP+TN}{TP+TN+FP+FN}, \tag{6}$$

- F1 score. It is a comprehensive metric amalgamating a model's precision and recall, providing a consolidated assessment of its overall performance.

$$pre = TP/(TP+FP), \tag{7}$$

$$recall = TP/(TP+FN), \tag{8}$$

$$F1 = 2 * (pre * recall)/(pre + recall), \tag{9}$$

Where, TP, FP, FN, and TN represent true positive, false positive, false negative, and true negative, respectively.

IV. RESULT ANALYSIS

A. Overall performance of the proposed framework

The performance of the proposed approach on multi-class AIS data quality classification is impressive, showcasing high metrics across various evaluation criteria. Specifically, the model achieved high scores: 99.98% for AUCROC, 98.87% for MCC, 99.29% for ACC, and 99.27% for F1 Score. Fig 3 shows the model's efficacy in distinguishing classes such as *Abnormal Constant Value*, *Temporal Missing*, and *Zig-zag Value* from the rest. However, the performance regarding the *Identity Theft* class appears less desirable, potentially attributed to inadequate instances within the training dataset. Nonetheless, the overall performance remains satisfactory, marked by minimal misclassifications.

B. Ablation test 1: impact of imbalance-learning strategies

The evaluation of imbalance-learning strategies is detailed in TABLE II. This investigation includes data-level, algorithm-level, and hybrid strategies. At the data level, MixUp and CutOut techniques are applied specifically to the minority classes, namely *Identity Theft* and *Zig-zag Value*, for sample augmentation. Algorithm-level strategies involve applying Focal Loss and Weighted CrossEntropy in model training process. These approaches are further combined to evaluate their collective impact. Among all the settings explored, the MixUp-based oversampling strategy, used in this study, outperformed others, by a margin of 0.10% in ROCAUC, 0.57% in accuracy, 0.66% in F1 score, and 0.91% in MCC. Additionally, the combination of Focal Loss and CutOut strategy displayed competitive performance, presenting itself as a viable alternative.

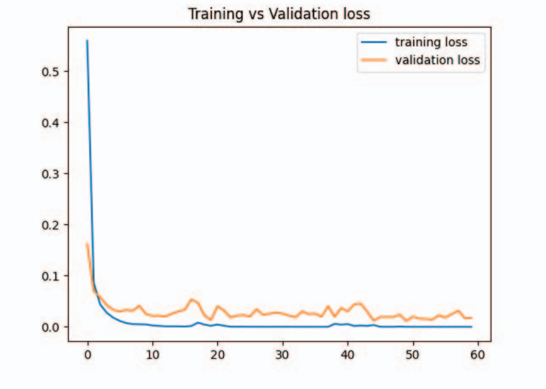

Fig. 2. Plotting of the loss in model training and validation process

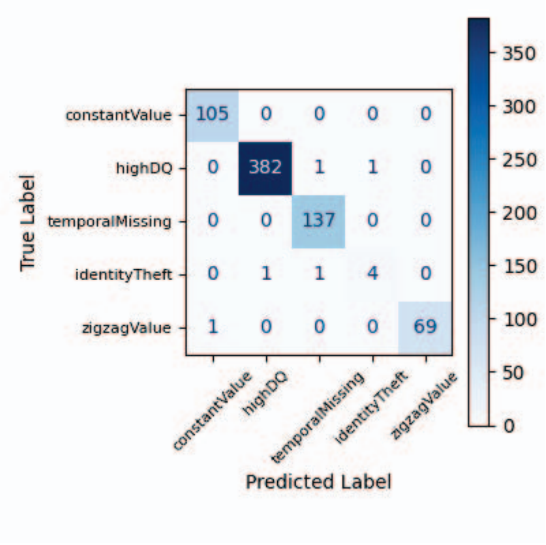

Fig. 3. Confusion matrix for AIS data quality classification

C. Ablation test 2: impact of deep learning models

An ablation test was executed to assess the impact of diverse classification models, and the comparative analysis of performance metrics is presented in TABLE III. ConvNeXtV2 emerges as the leading model, surpassing other models by up to 1.52% in AUCROC, 2.14% in accuracy, 2.35% in F1 score, and 3.40% in MCC. Following ConvNeXtV2, other convolution models like ResNet50, GhostNet100, EfficientNet, and Xception exhibit varying performance levels. Notably, SWINV2, a transformer-based model, secured the second-highest performance. Conversely, MobileNetV2 surfaces as the least effective model for this application, recording an AUCROC of 0.9846, an accuracy of 0.9715, an F1 score of 0.9692, and an MCC of 95.47%.

TABLE II
COMPARISON OF PERFORMANCE UNDER DIFFERENT IMBALANCE-LEARNING STRATEGIES.

Strategy	Loss function	Over-sampling	Model performance metrics			
			ROCAUC	ACC	F1	MCC
Data level	Cross entropy	MixUp	0.9998	0.9929	0.9927	0.9887
		CutOut	0.9999	0.9886	0.9861	0.9819
Algorithm level	Weighted CrossEntropy	No	0.9989	0.9872	0.9866	0.9796
		No	0.9993	0.9872	0.9847	0.9796
Hybrid	Weighted CrossEntropy	MixUp	0.9996	0.9872	0.9870	0.9797
		CutOut	0.9997	0.9915	0.9909	0.9864
	FocalLoss	MixUp	0.9992	0.9886	0.9874	0.9819
		CutOut	0.9998	0.9915	0.9900	0.9864

TABLE III
COMPARISON OF PERFORMANCE UNDER DIFFERENT DEEP LEARNING MODELS.

Model	Model performance metrics			
	ROCAUC	ACC	F1	MCC
MobileNetV2	0.9846(8)	0.9715(8)	0.9692(8)	0.9547(8)
ViTtiny	0.9965(5)	0.9744(7)	0.9734(7)	0.9593(7)
Xception41	0.9988(3)	0.9815(5)	0.9803(6)	0.9706(5)
EfficientNetlite0	0.9862(7)	0.9815(5)	0.9804(5)	0.9706(5)
GhostNet100	0.9990(2)	0.9829(4)	0.9818(4)	0.9728(4)
ResNet50	0.9983(4)	0.9872(3)	0.9847(3)	0.9796(3)
SWINV2	0.9896(6)	0.9886(2)	0.9861(2)	0.9819(2)
ConvNetxV2	(0.9998(1))	(0.9929(1))	(0.9927(1))	(0.9887(1))

V. CONCLUSION AND FUTURE WORK

To secure trusted data quality in the maritime digitalization process, this study proposed a vision-inspired framework tailored specifically for classifying AIS data quality issues in highly imbalanced datasets. Rigorous experiments show that the proposed framework demonstrated a remarkable 99.29% accuracy and 99.27% F1 score. Notably, the ConvNeXtV2 model emerged as the standout performer, outstripping other state-of-the-art models with advancements of 2.14% in accuracy, 2.35% in F1 score, and 3.40% in MCC, showcasing its superiority in this domain. Furthermore, the MixUp-based data augmentation strategy exhibited clear advantages over alternative imbalance learning methods, establishing itself as an effective solution for addressing class imbalance challenges in this context. Further studies will be conducted on the following aspects. Firstly, more representative data quality issues will be included to provide a more comprehensive diagnosis of AIS data quality. Secondly, the integration of temporal information in various forms, such as audio, could enhance the understanding of spatial-temporal correlations within AIS data. Finally, considering the dataset's collection in the Singapore Port, crucial endeavours will focus on testing the method's efficacy across varied maritime scenarios. In essence, this study serves as one of the few data-centric AI practices in the maritime sector, offering insights into fortifying maritime data reliability and data management practices.

VI. ACKNOWLEDGEMENT

This research was funded by Programme MARS (Programme of Maritime AI Research in Singapore) with funding grant number SMI-2022-MTP-06 by Singapore Maritime Institute (SMI).

REFERENCES

[1] L. Zhao, G. Shi, and J. Yang, "Ship trajectories pre-processing based on ais data," *Journal of Navigation*, vol. 71, pp. 1210–1230, 9 2018.

[2] N. Sambasivan, S. Kapania, H. Highfill, D. Akrong, P. Paritosh, and L. M. Aroyo, ""everyone wants to do the model work, not the data work": Data cascades in high-stakes ai." ACM, 5 2021, pp. 1–15. [Online]. Available: https://dl.acm.org/doi/10.1145/3411764.3445518

[3] M. Priestley, F. O'donnell, and E. Simperl, "A survey of data quality requirements that matter in ml development pipelines," *Journal of Data and Information Quality*, vol. 15, pp. 1–39, 6 2023. [Online]. Available: https://dl.acm.org/doi/10.1145/3592616

[4] C. Iphar, C. Ray, and A. Napoli, "Data integrity assessment for maritime anomaly detection," *Expert Systems with Applications*, vol. 147, p. 113219, 6 2020, very important!1) many background info2) the architecture is adoptabale 3) sample of paper publication. [Online]. Available: https://linkinghub.elsevier.com/retrieve/pii/S0957417420300452

[5] M. Stróżyna, D. Filipiak, and K. Węcel, "Data quality assessment – a use case from the maritime domain," *Lecture Notes in Business Information Processing*, vol. 394, pp. 5–20, 2020, a simple case study. [Online]. Available: https://link-springer-com.libproxy1.nus.edu.sg/chapter/10.1007/978-3-030-61146-0_1

[6] C. Iphar, A. Napoli, and C. Ray, "A method for integrity assessment of information in a worldwide maritime localization system," 2016. [Online]. Available: https://hal-mines-paristech.archives-ouvertes.fr/hal-01421920

[7] I. Kontopoulos, A. Makris, K. Tserpes, and W. Kainz, "Geo-information a deep learning streaming methodology for trajectory classification," 2021. [Online]. Available: https://doi.org/10.3390/ijgi10040250

[8] "Loitering behavior detection and classification of vessel movements based on trajectory shape and convolutional neural networks," *Ocean Engineering*, vol. 258, 8 2022.

[9] "Visualization and visual analysis of vessel trajectory data: A survey," *Visual Informatics*, vol. 5, pp. 1–10, 2021. [Online]. Available: https://doi.org/10.1016/j.visinf.2021.10.002

[10] Y. Tavakoli, L. Peña-Castillo, and A. Soares, "A study on the geometric and kinematic descriptors of trajectories in the classification of ship types," *Sensors 2022, Vol. 22, Page 5588*, vol. 22, p. 5588, 7 2022. [Online]. Available: https://www.mdpi.com/1424-8220/22/15/5588/htmhttps://www.mdpi.com/1424-8220/22/15/5588

[11] H. Rong, A. P. Teixeira, and C. G. Soares, "Maritime traffic probabilistic prediction based on ship motion pattern extraction," *Reliability Engineering System Safety*, vol. 217, p. 108061, 1 2022.

[12] Z. Liu, H. Mao, C. Y. Wu, C. Feichtenhofer, T. Darrell, and S. Xie, "A convnet for the 2020s," vol. 2022-June. IEEE Computer Society, 2022, pp. 11 966–11 976.

[13] S. Woo, S. Debnath, R. Hu, X. Chen, Z. Liu, I. S. Kweon, S. Xie, and M. Ai, "Convnext v2: Co-designing and scaling convnets with masked autoencoders." [Online]. Available: https://github.com/facebookresearch/ConvNeXt-V2

[14] X. Zhang, Z. Xiao, X. Fu, X. Wei, T. Liu, R. Yan, Z. Qin, and J. Zhang, "A viewpoint adaptation ensemble contrastive learning framework for vessel type recognition with limited data," *Expert Systems with Applications*, vol. 238, p. 122191, 3 2024.

[15] S. Han, X. Hu, H. Huang, M. Jiang, and Y. Zhao, "Adbench: Anomaly detection benchmark," *SSRN Electronic Journal*, 2022. [Online]. Available: https://www.ssrn.com/abstract=4266498