IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received 20 September 2022, accepted 7 October 2022, date of publication 13 October 2022, date of current version 19 October 2022.

Digital Object Identifier 10.1109/ACCESS.2022.3214217

RESEARCH ARTICLE

Multi-Label Classification for AIS Data Anomaly Detection Using Wavelet Transform

MARTA SZARMACH[1] **AND IRENEUSZ CZARNOWSKI**[2], **(Senior Member, IEEE)**
[1]Department of Electrical Engineering, Gdynia Maritime University, 81-225 Gdynia, Poland
[2]Department of Information Systems, Gdynia Maritime University, 81-225 Gdynia, Poland

Corresponding author: Marta Szarmach (m.szarmach@we.umg.edu.pl)

This work was supported by the Program of the Minister of Science and Higher Education through the name "Regional Excellence Initiative" in 2019-2022 under Project 006/RID/2018/19, the amount of financing PLN 11 870 000.

ABSTRACT Thanks to the Automatic Identification System (AIS), ships and other maritime equipment are able to communicate with each other, for example, by sending information about their position. This solution allows for early collision detection when two or more ships are on a collision course. In the newer version of AIS, a satellite infrastructure is used to extend the communication range. Unfortunately, satellite AIS deals with so-called packet collision effect: since there is a problem with synchronizing AIS data coming from multiple terrestrial areas, a single satellite may receive several AIS messages at the same time and be unable to correctly process them, causing the data to get lost or garbled. In this article, a machine learning based framework for detecting the incorrect AIS data is presented. In this approach, after the first stage (clustering), a dedicated anomaly detection algorithm searches for damaged AIS messages and conducts multi-label classification (with Random Forest and wavelet transform) to decide which fields of such message requires further correction. The results of measuring the effectiveness of the proposed approach using real AIS data are presented.

INDEX TERMS AIS data analysis, anomaly detection, multi-label classification, wavelet transform.

I. INTRODUCTION

AIS (Automatic Identification System) is a telecommunication system that allows maritime equipment (transponders on ships, shore-based stations, etc) to send and receive information about vessels in a given area and their movement [11]. The dynamic information provided by AIS is ship's position, speed, course, and so on, while the static information includes, for instance, vessel's identification number, MMSI (Maritime Mobile Service Identity) [11], [12]. The exchange of such information can result in early collision detection that may greatly improve the overall maritime security.

At first, AIS, existing as a so-called terrestrial segment, allowed for communication in a range of view (ship-to-ship or around coastal zones), using two VHF (Very High Frequency) frequencies 161.975 MHz and 162.025 MHz with a 25-kHz bandwidth [8]. However, the main restriction of the terrestrial segment was its range, limited to 74 km (40 nautical miles). To overcome this limitation, a satellite AIS segment (SAT-AIS) was introduced. In SAT-AIS, dedicated satellites (for example, the AAUSAT3 in the Low Earth Orbit [33]) mediate the communication between many small terrestrial AIS regions (called cells), highly increasing the range of the AIS system. Nonetheless, SAT-AIS struggles againts its own problems, coming from the fact that the transmission scheme in each cell is sychronized (using Self Organized Time Division Multiple Access technology) within such cell, but not necessarily between them. When a satellite receives an AIS packet from two or more cells at the same time, it is unable to properly process them and a problem of packet collision occurs [33]. Packet collision leads to the transmitted AIS data being incomplete (missing) or incorrect. Therefore, the maritime security may be harmfully influenced, for example, the lack of correct data may cause two ships to collide.

The associate editor coordinating the review of this manuscript and approving it for publication was Huiyan Zhang.

Therefore, there is a need to keep AIS data clean and stable, in the sense of their completeness. In other words, it is necessary to reconstuct messages that are lost or damaged due to a packet collision.

The AIS data can be analysed in relation to the machine learning theory and practices of implementation of tools belonging to this family. Based on this area of knowledge, the damaged data can be observed as anomalies. In other words, an anomaly detection results in an identification of damage in AIS data packets: we consider 'anomaly' as an AIS message that consist of at least one field which value got distorted during the transmission. From the definition point of view, anomalies are data points that have different nature from normal and typical character for considered phenomena. Anomaly detection is called unsupervised learning and is based on an assumption that the features of data anomalies are significantly different from those of normal instances. Among data scientists the process is also called outlier detection and machine learning domain offers a several different approaches for outlier detection (like, for example, density-based [5], classification-based using Decision Trees [6] or neural networks [14]).

The machine learning deals also with a problem of prediction of quantitative and qualitative values. It is one of the typical task of machine learning tools (that also can be observed by view of data mining [2]). The problem of missing AIS data can be related to the problem of prediction of the incomplete data [17], lost due to packet collision.

The aim of the research work reported in this paper was to evaluate the performance of the approach to anomaly detection, based on wavelet transform and multi-label classification, for identification of damaged fields in an AIS message. The paper presents the results of the evaluation together with their dicussion. Thus, the question addressed in this paper concerns possiblities of detection of damaged messages and their fields using wavelet transform and multi-label classification, however, with a need of clustering before the detection.

The mentioned techniques and tasks are integrated under a dedicated framework for AIS data reconstruction including the following stages: clustering, anomaly detection and prediction. In this paper, the last one stage is not considered in details and will be the attention of future research.

The contributions of the paper are as follows:

- A dedicated framework for AIS data reconstruction is shown.
- A new approach for anomaly detection in AIS messages (identification of damaged fields) based on wavelet transform and multi-label classification is proposed.

This paper is organised as follows. The next section contains a literature review. The problem formulation is included in Section III. A framework for AIS data reconstruction is presented in Section IV, where also a general concept of the proposed approach is included, together with details on a searching process of damaged AIS messages and damaged fields in AIS messages. A discussion of the experimental results is given in Section V. The final section contains our conclusions and directions for future research.

II. RELATED WORK

The topic of detecting anomalies in AIS data is still widely discussed in the literature. There are many ways in which researchers around the globe try to deal with this problem. Machine learning techniques are present in many of the proposed frameworks.

Most of the published works focus on analysing AIS data to define the trend of ships' trajectories (in other words, their natural behaviour). In such approach, what is considered as an anomaly is a part of trajectory that do not follow the trend. For example, in [34], researchers trained a bayesian recurrent neural network to detect the anomalous vessel behaviour. The statistical approach using gaussian process was proposed in [16]. The method proposed in [15] uses DBSCAN (Density-Based Spatial Clustering of Applications With Noise) algorithm to cluster similar trajectories to define the trend and in [18] the distance between points is measured to detect anomalies. Other approach involving clustering is presented in [32]: here, the convolutional neural network with auto-encoder is used.

One popular way of modelling the vessels' trajectory is to define specific trajectory points called waypoints. They are points in a given area where ships tend to turn, speed up, etc. Trajectory is described as a graph; its vertices are the waypoints and its edges are the stages of the actual ship trajectory. Works that propose methods of describing vessels' movement utilizing waypoints are, for example, [7] (using genetic algorithm) or [31].

Another set of works deals with the problem of predicting the ships' trajectories. In might be considered as a next step of the experiment, after analyzing the trend of those trajectories. The usage of neural network in many kinds is quite common in papers covering this topic. In [35], backpropagation neural network was utilized, in [13] — recurrent neural network, while in [19] — convolutional U-net neural network. The algorithms presented in mentioned articles are fed with AIS data, however, in the literature we can find also other methods for missing data imputation, such as presented in [24] based on Decision Trees and fuzzy clustering.

In the currently proposed solutions to AIS data reconstruction, there is a need for inventing fast and reliable algorithms that do not need a lot of data (thus do not require long observation time to create a trend of maritime traffic) to detect anomalies. An anomaly does not often mean a route that does not fit to the defined trend — sometimes the source of the problem lies within the damaged data transmitted via AIS. An approach that analyses each AIS message and manages to identify the incorrect values in that message fields is research challenge to improve the overall quality of AIS data. Our approach presented in this article focuses on this problem.

III. PROBLEM FORMULATION

The problem of reconstructing missing or incorrect AIS data can be described as follows. Let us denote the ship's trajectory as a set of the following vectors (1):

$$T_i^{t_m} = [x_{i1}^{t_m}, x_{i2}^{t_m}, x_{i3}^{t_m}, \dots, x_{iN}^{t_m}], \tag{1}$$

where:

- i — ship's identifier,
- $T_i^{t_m}$ — ith ship's trajectory point observed in time (observation step) t_m, where $m = 1, \dots, M$ and:
  - M – the number of received AIS messages,
  - N — the feature number of AIS message,
  - $x_{in}^{t_m}$ — nth feature of an AIS message from ship i received in time t_m, where $n = 1, \dots, N$.

Thus, a trajectory of ship ith can be described as follows:

$$T_i = \{T_i^{t_1}, T_i^{t_2}, T_i^{t_3}, \dots, T_i^{t_M}\}. \tag{2}$$

When the problem of packet collision exists, some parts of the AIS data received from ith ship can be damaged, which means that there are messages (i.e. $T_i^{t_m}$) or their fields (i.e. elements $x_{in}^{t_m}$ of a vector $T_i^{t_m}$), that are incorrect, which can be expressed as:

$$\exists_{t_m} T_i^{t_m}, \text{ that is missing/incorrect} \tag{3}$$

or

$$\exists_{n:n=1\dots N}\ x_{in}^{t_m}\text{ , that is missing/incorrect.} \tag{4}$$

The problem of reconstructing missing or incorrect AIS data can be defined as a detection of incorrect (damaged) messages or prediction of the missing elements of messaged (i.e. their fields), and their correction, to the goal expressed as:

$$\forall T_i(\nexists T_i^{t_m} \vee \nexists x_{in}^{t_m})\text{ , that are not missing/incorrect.} \tag{5}$$

IV. PROPOSED APPROACH FOR AIS DATA RECONSTRUCTION

A. GENERAL ALGORITHM

The described AIS data reconstruction approach is based on machine learning techniques. The main framework can be described as Algorithm 1.

Algorithm 1 Framework of AIS Data Reconstruction Using Machine Learning

```
Require: $D$ — dataset; $K$ (optional, according to the chosen clustering algorithm) — number of clusters (equal to the number of individual vessels appearing in a dataset)
begin
    $D_c = D$
    Map messages from $D$ into $K$ clusters (1 cluster = messages from 1 ship) using a selected clustering algorithm.
    Let $D_1, \dots, D_K$ denote the obtained clusters and let $D = D_1 \cup D_2 \cup \dots \cup D_K$.
    for all $i = 1, \dots, K$ do
        Search for anomalies (potentially damaged messages) in $D_i$.
        Let $\hat{D}_i$ denote cluster with detected anomalies.
        Predict the correct values of damaged fields for $\tilde{D}_i$.
        Update $D_c$ using $\tilde{D}_i$.
    end for
    return $D_c$ — corrected dataset.
end
```

Based on the above code, the following 3 stages are identified.

1) CLUSTERING STAGE

The first stage in AIS data reconstruction is clustering. Clustering is an act of dividing data into groups, where datapoints gathered in one group are more similar (according to a chosed distance metric) to each other than to datapoints from other groups [27]. Data collected in an AIS dataset comes from multiple vessels, however, to effectively analyse the trajectories of each ship, only data (AIS messages) related to that ship should be considered. Only then the data that require correction can be found or the actual reconstruction can be made. That is why the clustering stage is used. The dataset is mapped into smaller groups such that each group (ideally) consists of datapoints describing one and only one particular vessel. Clustering algorithms such as k-means [21], [28], DBSCAN [9] or others can be used here. However, relying on our previous research of the performance of clustering algorithms in AIS data analysis [22], [23], we decided to use DBSCAN in our framework.

Another method of distinguishing individual vessels' trajectories would be sorting the dataset according to ships' identifiers (MMSI, Mobile Maritime Service Identity), but since MMSI is transmitted in AIS messages along with other information, the field carrying that information could possibly be damaged and contain erroneous value. The use of clustering algorithm should be able to find the similarity between datapoints of one ship, despite such errors.

2) ANOMALY DETECTION STAGE

The second stage of AIS data reconstruction in this proposed framework is called the anomaly detection stage. Anomaly detection can be explained as searching for the data that somehow do not match the rest (in terms of their probability distribution, etc). In the case of AIS data, abnormal datapoints very possibly come from damaged AIS messages. Finding such outliers is, in other words, finding AIS messages that need further correction.

Not only identifying the whole damaged AIS message is important, but also the index of a certain field (or fields) which value is corrupted has to be defined, as well as the group that the damaged datapoint should belong to (if after the clustering stage it is not assigned to any other existing group).

Moreover, it is advisable here not to make a long observation of a typical maritime movement in a given area to define the trend and to treat every trajectory points that do not fit

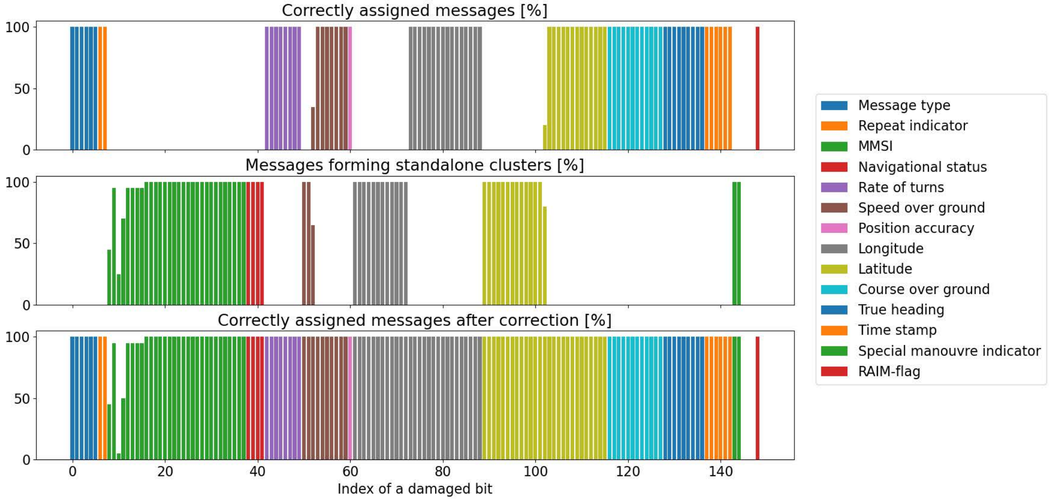

FIGURE 1. Impact of a position of a damaged bit in an AIS messaged on the clustering quality, showing which damaged bits create a standalone cluster.

to that trend as outliers. We should try to analyse the real movement of a particular ship instead.

3) PREDICTION STAGE

After the damaged AIS messages and their incorrect fields are identified, it is finally known which points must be considered and processed during the AIS message reconstruction, that is carried out in the prediction stage. The aim of this stage is to predict the real values of the damaged AIS messages fields, based on the information gained from the gathered data from specific group (related to one particular ship) and update the entire dataset, so that it will no longer contain incorrect data. Naturally, after the datapoint is labeled as an 'anomaly' in the previous stage, it is not used in prediction stage.

Ideally, the reconstruction process should allow all AIS messages to retrieve their real form in a fast and adequate way — when it comes to maritime safety, waiting for a long time for reconstruction, while relying on false data may lead to two ships colliding, is unacceptable.

In this paper, the anomaly detection stage is mainly discussed and analysed.

B. SEARCHING FOR DAMAGED AIS MESSAGES

1) MOTIVATION: STANDALONE CLUSTERS ANALYSIS

While searching for outlying AIS messages that are likely damaged and need correction, it might be useful to identify messages (datapoints) that form a standalone cluster (consisting of only one datapoint). The motivation behind this logic is that if the datapoint was considered by the clustering algorithm as not being similar to any other existing group (not resembling any ship trajectory), then that point might contain incorrect values.

The validity of this assumption was pre-examined in a simple experiment. Each bit of randomly chosen AIS message was artificially damaged and the whole dataset was clustered. The results show that in some cases (as presented in 2. part of Fig. 1) such corrupted messages indeed forms an additional 1-element cluster. By finding such clusters, the AIS messages that need to be reconstructed can be detected.

Naturally, there is also a possibility that a 1-element cluster contains message from a ship that manages to send only one position report during the observation. An easy way to enable such messages not to be considered an outliers is checking the number of fields that need correction (described later) and if more than a half of the fields seems broken, that message should probably form an independent cluster.

2) USAGE OF *k* NEAREST NEIGHBOURS ALGORITHM IN PROPER GROUP ASSIGNMENT

As mentioned before, identyfing the damaged AIS messages (if they form standalone clusters) is not enough, there is also a need to find the ships that those messages originated from (in other words, to find clusters that those messages should be assigned to). Without that knowledge, it would be impossible to decide which datapoints to use while reconstructing such messages. A classification aspect of an algorithm called *k* nearest neighbors (*k*-NN) can be used here.

General form of the *k* nearest neighbors is described as Algorithm 2. The main idea behind this algorithm is that objects that share common traits should be somehow similar to each other (in mathematic words, the distance measured between two points with a common label is expected to be smaller than be distance between two points of different classes). That is why to classify a single example it is possible to find a few points from the training set closest to that example and check which class those points mostly belong to.

The same logic can be used in deciding which ship a damaged message in a 1-element cluster originates from. The distance between this datapoint and other points reflecting messages from that particular ship is believed to be

Algorithm 2 *K* Nearest Neighbours Algorithm in Classification Tasks [1]

```
Require: $X$ — training dataset;
         $y$ —training dataset labels;
         $D$ — data to classify;
         $k$ — number of nearest points to rely on.
begin
    Compute the distance between $D$ and examples in $X$.
    Find $k$ lowest distances and check which labels from $y$
    they are referring to.
    return most frequent label from selected $k$ labels.
end
```

smaller (they are more similar to each other) than between this datapoint and points (messages) from any other vessel. Hence, in our approach for AIS data reconstruction, if during the clustering stage a standalone cluster j is found, a k-NN classifier is trained (its input training data X are observed examples clustered in normal, non-standalone clusters and the training labels y are the indices of the clusters that those points are assigned to) and a classification of the content of that standalone cluster is proceeded. The predicted label should correspond to the index of a group i that the damaged AIS message possibly should belong to.

To sum up, the classification described here is a part of multi-class classification problem: the task is to assign a damaged AIS message (T_j^1, belonging to cluster j that consists of only one message, but originating from ship i) to a group T_i (together with other messages from ship i), where i can be one natural number from 1 to K and K is the number of ships observed in a dataset:

$$T_j^1 \subset T_i : i \in \{1, 2, \dots K\} \tag{6}$$

A single AIS message can come from only one of many vessels, that is why multi-class classification is used here.

The third part of Fig. 1 shows the confirmation of the usage of k-NN algorithm in damaged AIS messages classification — 96.54% of artificially damaged datapoints (where, as mentioned before, 1 particular bit was corrupted) was correctly classified to belong together with other datapoints from the desired vessel. The k value (number of closest datapoints to find) was set arbitrarily to 5.

C. DETECTION OF A DAMAGED FIELDS IN AIS MESSAGES

When the possibly damaged AIS messages are identified, next important analysis must be carried out with aim to search for damaged fields in those AIS messages.

1) USAGE OF A WAVELET TRANSFORM

One way to accomplish this is with the usage of a wavelet transform. Wavelet transform [4] is a transformation that measures how much an input signal $f(t)$ is similar to another signal called wavelet $\psi(t)$. Wavelet transform can be described as [4]:

$$F_\psi(a, b) = \frac{1}{\sqrt{a}} \int_{-\infty}^{\infty} f(t) \cdot \psi\left(\frac{t-b}{a}\right) \mathrm{d}t \tag{7}$$

where:

- $F_\psi(a, b)$ is a wavelet transform of a singal $f(t)$ based on the wavelet $\psi(t)$, where t denotes time,
- a is a scale parameter,
- b is a time-shift parameter.

A wavelet $\psi(t)$ is a mathematic function that meets the following criterium [4]:

$$\int_{-\infty}^{\infty} \frac{|\hat{\psi}(\omega)|^2}{\omega} \mathrm{d}\omega < \infty \tag{8}$$

where $\hat{\psi}(\omega)$ is a Fourier transform of a wavelet $\psi(t)$.

Fig. 2 shows the intuition behind the usage of the wavelet transform in damaged AIS message fields detection. The first chart presents the waveform of one field — longitude of one vessel — with one value (marked in red) being incorrect. It can be noticed that the nature of the damage somehow resembles the nature of a wavelet (in the form of a sudden value change). The numeric indication of this,,similarity'' can be computed as the result of a wavelet transform, thus 2 such transforms were calculated:

- for normalized waveform $\Delta\hat{w}_{in}$ describing the differences between consecutive reported values of a feature n (longitude in this case) of the analyzed ship i (with the maximum value in the waveform being the normalizing scale):

$$\begin{aligned} \Delta\hat{w}_{in} &= [x_{in}^{t_2} - x_{in}^{t_1}, \quad x_{in}^{t_3} - x_{in}^{t_2}, \quad \dots \quad x_{in}^{t_M} - x_{in}^{t_{M-1}}] \\ \Delta\hat{w}_{in} &= \frac{\Delta\hat{w}_{in}}{\max(\Delta\hat{w}_{in})} \end{aligned} \tag{9}$$

- the same waveform without the value from the damaged message $\hat{x}_{jn}^{t_m}$ (normalized with the same scale):

$$\begin{aligned} \Delta w_{in} &= [x_{in}^{t_2} - x_{in}^{t_1}, \quad \dots \quad x_{in}^{t_M} - x_{in}^{t_{M-1}}]^{x_{in}^{t_m} \neq \hat{x}_{jn}^{t_m}} \\ \Delta w_{in} &= \frac{\Delta w_{in}}{\max(\Delta\hat{w}_{jn})} \end{aligned} \tag{10}$$

The results of a wavelet transform $\hat{W}_\psi^{in}$ and W_ψ^{in} (with $a = 1$ and the use of a Morlet wavelet) for both such inputs are visualized in the second chart of Fig. 2. It is clearly noticable that the transform $\hat{W}_\psi^{in}$ of the corrupted waveform differs from the original one around the place of the damage. This can be detected by computing the relative difference of maximum values of both transforms:

$$\Delta\psi_{in} = \frac{|\max(\hat{W}_\psi^{in}(a=1, b)) - \max(W_\psi^{in}(a=1, b))|}{\max(\hat{W}_\psi^{in}(a=1, b))} \tag{11}$$

The closer the difference to 1, the more likely that waveform truly consists of a wavelet-like damage. Example values for each field are shown in the bottom part of Fig. 2 — for fields that were not corrupted, the relative difference

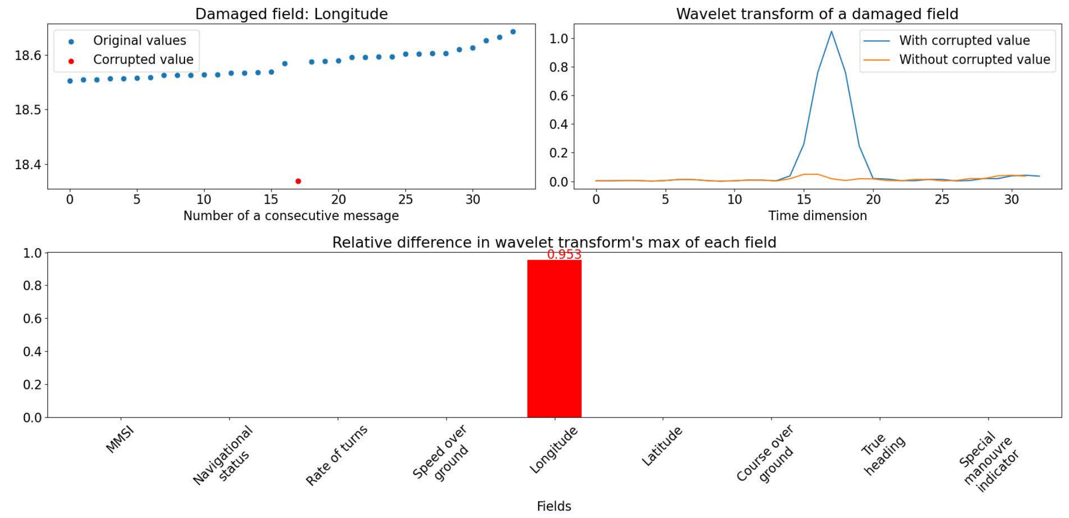

FIGURE 2. Wavelet transform in damaged AIS message fields detection.

in maximum wavelet transforms is around 0 (indicating,,no damage''), while only for the artificially damaged fields it is high (0.953, indicating,,damage'').

2) MEASURING THE DIFFERENCE OF STANDARD DEVIATION

Another way of determining if a single value does not match the others may utilize the standard deviation σ of values form a given waveform. Standard deviation indicates to what extent values of a given set can differ from their mean. An appearance of an outlying value (possibly distant from the mean) may cause the standard deviation to reach much higher values.

Thus, we can calculate another relative difference, i.e. between the standard deviation $\sigma_{\hat{w}_{in}}$ of values from waveform $\hat{w}_{in}$ (all values of a given field in an analysed group) and standard deviation $\sigma_{w_{in}}$ of values from w_{in} (without the value from a standalone cluster j), as follows:

$$\begin{aligned}\hat{w}_{in} &= [x_{in}^{t_1}, \quad x_{in}^{t_2}, \quad \ldots \quad x_{in}^{t_M}]\\ w_{in} &= [x_{in}^{t_1}, \quad x_{in}^{t_2}, \quad \ldots \quad x_{in}^{t_M}]^{x_{in}^{t_m} \neq \hat{x}_{jn}^{t_m}}\\ \Delta\sigma_{in} &= \frac{|\sigma_{\hat{w}_{in}} - \sigma_{w_{in}}|}{\sigma_{\hat{w}_{in}}}\end{aligned} \tag{12}$$

Again, the closer the difference is to 1, the more likely the analysed value is damaged and needs to be corrected.

3) MUTLI-LABEL FIELD CLASSIFICATION USING RANDOM FOREST

Let $[\Delta\psi_{in}, \Delta\sigma_{in}]$ denote a 2-element vector describing the impact of adding the field's value from a standalone cluster to other values from a given group (related to the particular ship). As mentioned before, the higher the values in this vector, the more likely the field they were calculated for is corrupted. However, there is a need to answer one question: which vectors exactly indicate damaged fields and which correct ones?

In this article, we introduce two ways of classifying fields:

- by setting the constant threshold (say 0.5) — if both values $\Delta\psi_{in}$ and $\Delta\sigma_{in}$ are higher than the threshold, the field is considered damaged,
- by using a well known binary classifier, such as Random Forest — the prediction would be made upon the $[\Delta\psi_{in}, \Delta\sigma_{in}]$ vectors and the predicted value 0 indicates,,no damage in a field'' while 1 —,,damage''.

Random Forest [10] is an ensemble method belonging to supervised machine learning. It is described as Algorithm 3. It uses a given number of decision trees (which make predictions regardning each example using pretrained *if...else* rules) to classify examples independently and then the final prediction is (when it comes to classification task) the most frequently predicted label. Random Forest is relatively fast (compared to, for instance, neural network approaches [26]), accurate (compared to logistic regression [3]) and less prone to overfitting (compared to a single Decision Tree [10]), hence, we find it useful in our framework.

As a result of using this kind of field classifier, we would like to receive the list of all fields in a given message that seem incorrect. We cannot know a prori how many such fields will be — it can be none, one, two, etc — hence, the classifiction problem described here should be considered as a multi-label classification (each label being the number of AIS message field to reconstruct), in which there can be many labels assigned to a single observation point. This can be done by creating multiple Random Forest binary classifiers — each for one AIS message field. By iterating over every Random Forest classifier that decides whether add this class' (field) label to the analyzed record or not, in the end, we get the desired list of damaged fields.

Algorithm 3 Random Forest Algorithm in Binary Classification Tasks [10]

```
Training:
Require: $X$ — training dataset;
         $y$ —training dataset labels;
         $max\_depth$ — maximum number of levels of a single tree in a forest;
         $n\_estimators$ — number of trees in a forest.
begin
  for $i = 1, 2, \ldots n\_estimators$ do
    while $max\_depth$ not exceeded do
      On a subset of $X$ find a feature $n$ and its threshold that most accurately divides $X$ to reflect $y$.
      Split the tree into 2 parts using the calculated threshold.
    end while
  end for
  return $model$ — trained Random Forest.
end
Prediciton:
Require: $D$ — dataset to classify;
         $model$ — trained Random Forest.
begin
  for $i = 1, 2, \ldots n\_estimators$ do
    Follow the trained if instructions:
if field value > trained threshold
until the last tree level is reached.
    Get the label on the last tree level.
  end for
  return most frequent label among all $n\_estimators$ trees.
end
```

Equations (9)-(12) are constructed in a way that for the correct AIS messages they would produce low-valued $[\Delta\psi_{in}, \Delta\sigma_{in}]$ vectors, so the Random Forest classifier would classify them as non-damaged and no label would be assigned to them, indicating that there is no field that requires correction.

To conclude, the classification described here is a part of multi-label classification problem: the task here is to assign a damaged AIS message T_j^i to groups C_n (which can be many), where n, indicating the number of AIS message field that is damaged, is a natural number from 1 to N and N is the number of features in a dataset:

$$T_j^1 \in C_n : n \in \{1, 2, \ldots N\} \tag{13}$$

A single incorrect AIS message can be damaged in many fields, that is why multi-label (and not multi-class) classification is used here.

The entire anomaly detection algorithm (based on searching for the standalone clusters in AIS data) is presented as Algorithm 4.

To sum up the, the flowchart of the entire presented AIS data reconstruction algorithm is shown in Fig. 3. We can see how the data is passed to the clustering stage to distinguish single trajectories, analysed to find damaged messages and fields, and then reconstructed, with the use of the presented algorithm and proper hypeparameters (such as Random Forest's *max_depth*) optimized to obtain the best performance and reduce model overfitting.

Algorithm 4 Anomaly Detection in AIS Data Based on Standalone Clusters Search

```
Require: $D_1, D_2, \ldots D_K$ — subsets of the AIS dataset (groups obtained during the clustering phase); $N$ — number of fields in AIS messages type 1-3.
begin
  for $j = 1, 2, \ldots K$ do
    if $|D_j| == 1$ then
      Add index of $T_j^1$ to $idx\_list$.
      Run $k$-NN algorithm to find group $i$ ($i \neq j$) — the one that the message $T_j^1$ should be assigned to.
      Add $i$ to $i\_list$.
      for $n = 1, 2, \ldots N$ do
        Compute the wavelet transform $\hat{W}_{\psi}^{in}$ of normalized sequence of differences between the consecutive values of field $n$ in group $i$.
        Compute the wavelet transform $W_{\psi}^{in}$ of the same sequence, excluding the value from the potentially damaged message $T_i^1$.
        Compute $\Delta\psi_{in}$, the relative difference between the maximum values of both wavelet transforms.
        Compute $\Delta\sigma_{in}$, the relative difference between the standard deviation of values of field $n$ from group $i$ and those values excluding the one from the potentially damaged message.
        Run Random Forest to classify vector $[\Delta\psi_{in}, \Delta\sigma_{in}]$.
        if classification result == ‘field $n$ is damaged’ then
          Add $n$ to $n\_list$.
        end if
      end for
    end if
  end for
  return $idx\_list$ — list of indices of AIS messages that require correction, $i\_list$ — list of indices of groups that those messages should be assigned to, $n\_list$ — list of indices of AIS message fields that require correction.
end
```

V. COMPUTATIONAL EXPERIMENT

A. OVERVIEW

The computational experiment for the anomaly detection stage has been carried out. The main aim of the experiment was to examine the effectiveness of the proposed method (relying on standalone clusters analysis, k-NN algorithm in

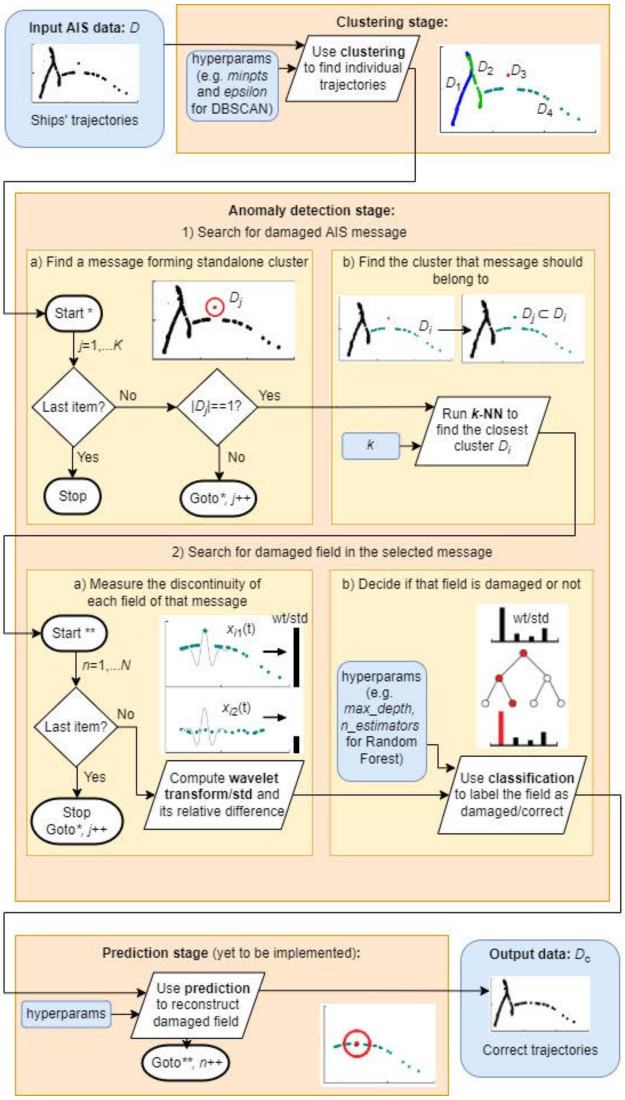
Input AIS data: D
Ships' trajectories
Clustering stage:
hyperparams (e.g. minpts and epsilon for DBSCAN)
Use clustering to find individual trajectories
D1
D2
D3
D4
Anomaly detection stage:
1) Search for damaged AIS message
a) Find a message forming standalone cluster
Start *
j=1,...K
Last item?
No
Yes
Stop
|Dj|==1?
Yes
No
Goto*, j++
Dj
b) Find the cluster that message should belong to
Di
Dj ⊂ Di
k
Run k-NN to find the closest cluster Di
2) Search for damaged field in the selected message
a) Measure the discontinuity of each field of that message
Start **
n=1,...N
Last item?
No
Yes
Stop Goto*, j++
xi1(t)
xi2(t)
wt/std
Compute wavelet transform/std and its relative difference
b) Decide if that field is damaged or not
hyperparams (e.g. max_depth, n_estimators for Random Forest)
Use classification to label the field as damaged/correct
Prediction stage (yet to be implemented):
hyperparams
Use prediction to reconstruct damaged field
Goto**, n++
Output data: Dc
Correct trajectories

FIGURE 3. Flowchart of the presented AIS data reconstruction algorithm.

correct group classification, wavelet transform and Random Forest in damaged fields identification) on detecting:

- damaged AIS messages in a given dataset,
- the ship that those messages originated from,
- damaged fields in a selected message.

The algorithms have been implemented using Python programming language, utilizing mostly Scikit-learn library [25]. The calculations have been run on Visual Studio Code.

B. DATA

In AIS, messages of 27 different types are transmitted [12]. However, only 3 of them carry information regarding ships' trajectories (their position, speed, course, etc): messages of type 1, 2 and 3. They are called position reports and consist of 168 bits in total. The structure of a position report message is described in Table 1.

In this experiment, data from a real, operational AIS was used. Recorded AIS messages of types 1-3 were divided into 3 datasets:

TABLE 1. Structure of AIS messages type 1-3 [12].

Field	Bits	Format
Message ID	1-6	*Unsigned integer*
Repeat indicator	7-8	*Unsigned integer*
User ID (MMSI)	9-38	*Unsigned integer*
Navigational status	39-42	*Enumarated (unsigned integer)*
Rate of turns	43-50	*Signed integer with scale*
Speed over ground	51-60	*Unsigned integer with scale*
Position accuracy	61-62	*Boolean*
Longitude	62-89	*Signed integer with scale*
Latitude	90-116	*Signed integer with scale*
Course over ground	117-128	*Unsigned integer with scale*
True heading	129-137	*Unsigned integer*
Time stamp	138-143	*Unsigned integer*
Special manoeuvre indicator	144-145	*Enumarated (unsigned integer)*
Other (spare, radio status)	148-168	-

1) 805 messages from 22 vessels from the area of Gulf of Gdańsk,
2) 19 999 messages from 524 vessels from the area of Gibraltar,
3) 19 999 messages from 387 vessels from the area of Baltic Sea.

All ships' trajectories from those 3 datasets are visualised in Fig. 4, but those datasets were further divided into training (50% trajectories), validation (25%) and test (25%) sets.

Among all position report fields, 8 of them were selected to form the input for the machine learning based AIS data reconstruction algorithm. Those chosen features are set as follows:

- longitude,
- latitude,
- navigational status,
- speed over ground,
- course over ground,
- true heading,
- rate of turns (not used in clustering phase),
- special manoeuvre indicator,
- MMSI (*Maritime Mobile Service Identity*).

Information transmitted in other position report fields (namely Message ID, repeat indicator, position accuracy and timestamp) is not necessarily related to ships' trajectories, but mostly to the message itself — therefore, we do not find them useful in reconstructing the most meaningful trajectory-based data.

For the clustering phase, the data was additionally preprocessed. Some features, namely identifiers, such as MMSI, navigational status and special manoeuvre indicator, were one-hot encoded (converted to binary vectors with the value of 1 placed in only one position) to avoid machine learning algorithms from considering their values as the intensity of some physical quantities. Datasets were also standarized (values of each feature were rescaled to fit into a probability distribution with the mean of 0 and variance of 1 by subtracting their means and dividing by standard deviation).

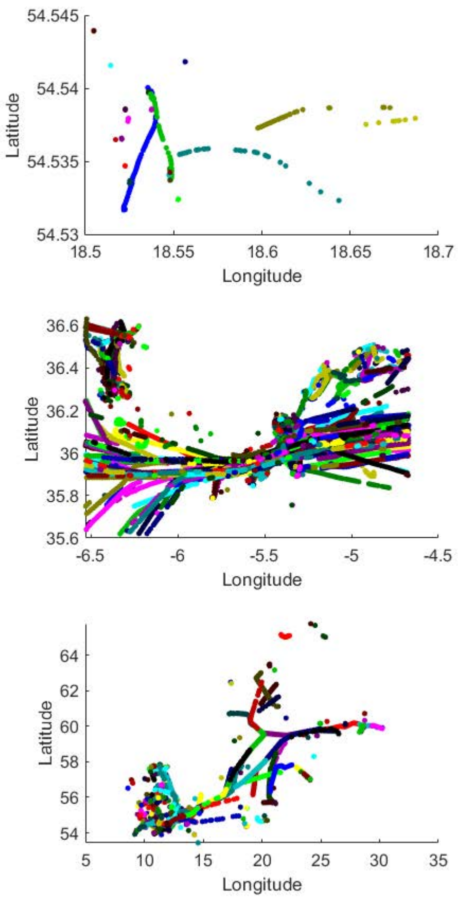

FIGURE 4. Visualisation of trajectories (each marked with a different colour) from all 3 datasets.

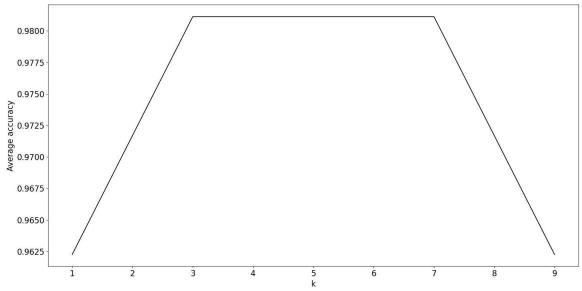

FIGURE 5. Searching for the optimal value of *k*.

C. TRAINING CLASSIFIERS

The initial stage of AIS data reconstruction was training dedicated classifiers to decide which group should the damaged standalone message be assigned to and whether given AIS message field is damaged or not. As mentioned before, in this experiment, *k*-NN and Random Forest classifiers were used in this task.

1) OPTIMIZING *k*-NN

The hyperparameter that has to be tuned in *k*-NN model is *k* — the number of closest points that the prediction relies on. In this simple experiment, 5 different values of *k* was examined to find the optimal one: 1, 3, 5, 7, and 9.

For each *k* value, the *k*-NN classifier was trained several times on a slightly different dataset (each time a randomly chosen AIS message was artificially corrupted, i.e. one of its bit was swapped, then the whole dataset was standarized and the classifier was trained on this data X excluding the damaged message, while the number of group that each point was assigned to after the clustering phase was the label y). Then the model tried to predict the right label (group) for the previously chosen message and the accuracy of such classification was stored. The mean accuracy for each *k* value is presented in Fig. 5.

According to the mentioned figure, the optimal *k* values in this case are 3, 5 or 7. The value of 5 was chosen to be used in the upcoming experiments as a hypothetical local maximum of the presented curve.

2) OPTIMIZING RANDOM FOREST

We built 9 independent Random Forest models, each destined to classify different AIS message field, and therefore, 9 training and validating sets were created. Each set was built (based on all messages from the 1. available dataset with AIS messages from Gulf of Gdańsk) by artificially damaging the value of the desired field (for such message a label of '1' indicating 'damaged' was added) or leaving the message undamaged (with a label '0' indicating 'not damaged') and calculating the 2-element vectors with relative differences in wavelet transform and standard deviation. Eventually the dataset for each field was divided into training (75% of all generated entries) and validation set (25%).

Validation sets were used to tune the Random Forest model hyperparameters. We decided to tune 2 of all its possible hyperparametres [25]:

- *max_depth* — the maximum number of levels (branches) in each decision tree in a forest: the higher this value, the more complex decisions can be made, but the model is more likely to overfit to the training set,
- *n_estimators* — the number of individual decision trees forming the entire forest.

The optimal values were checked in range from 2 to 100 (for *max_depth*) or 200 (for *n_estimators*). For each value, the classification on both training and validation set was run and its performance for both sets was calculated: we computed the binary accuracy, i.e. the percentage of examples where the classifier made a correct prediction, for all 9 Random Forests corresponding to each examined field and we took the mean of all those accuracies (please note that later, while measuring the performance of the actual prediction, we use different metric, such as recall and precision, described in Quality Metrics section). The *n_estimators* value resulting in the highest such mean binary accuracy on the validation set was considered optimal. According to the obtained results,

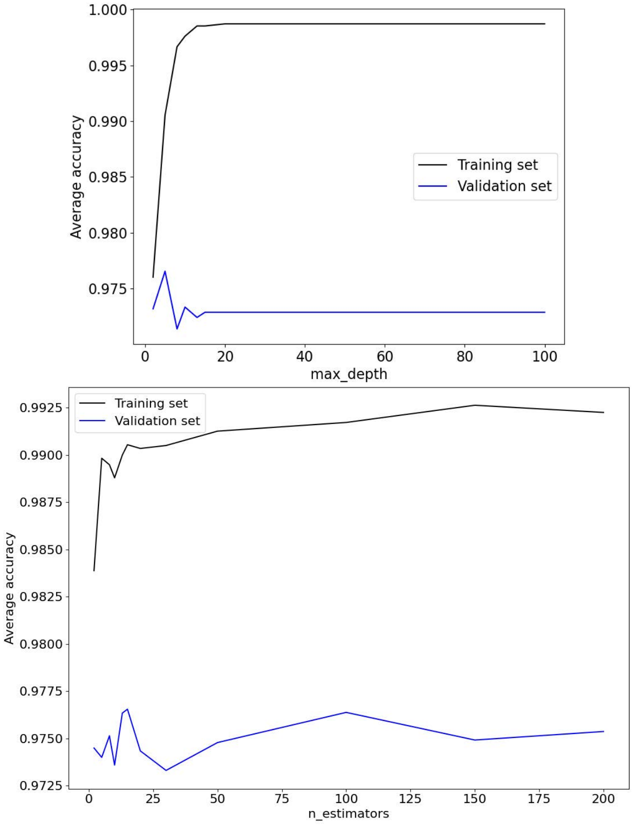

FIGURE 6. **Searching for the optimal value of *max_depth* and *n_estimators*.**

the *max_depth* value was set to 5, and *n_estimators* to 15, as shown in Fig. 6.

After the final training of the Random Forest model, the average binary accuracy on the training set was 98.96%.

D. ANOMALY DETECTION RESULTS

Finally, the effectiveness of anomaly detection stage using proposed algorithm has been evaluated.

1) QUALITY METRICS

At first, the quality of determining the correct clusters for messages from 1-element clusters has been examined. In this case, a simple accuracy metric was used, in other words, we measured the percentage of messages correctly classified to their original groups.

Then, the correctness of finding damaged fields in AIS messages has been evaluated. As mentioned before, for a single AIS message, the act of detecting its damaged fields can be considered as a multi-label classification. That being said, to conduct a proper evaluation of such classification, special quality metrics had to be used, including [25]:

- recall — percentage of detected damaged fields (*predicted*) among all truly damaged fields (*real*):

$$\text{recall} = \frac{|\text{real} \cap \text{predicted}|}{|\text{real}|} \tag{14}$$

Regarding the analysed issue of finding damaged AIS message fields that need corrction, recall should be considered as the most important quality measure, since it is mostly desired to find all damaged fields, even if some fields classified as damaged are actually correct.

- precision — percentage of detected truly damaged fields among those classified as damaged:

$$\text{precision} = \frac{|\text{real} \cap \text{predicted}|}{|\text{predicted}|} \tag{15}$$

- F1 score — harmonic mean of precision and recall:

$$\text{F1} = 2 \cdot \frac{\text{recall} \cdot \text{precision}}{\text{recall} + \text{precision}} \tag{16}$$

- Jaccard score — the size of an intersection between truly damaged fields and fields classified as damaged, divided by the size of an union of those two sets:

$$\text{Jaccard} = \frac{|\text{real} \cap \text{predicted}|}{|\text{real} \cup \text{predicted}|} \tag{17}$$

- Hamming score — Hamming distace between two vectors (with the length equal the number of all fields): one containing the labels of classification (damaged/correct), the other containing information whether the field is truly damaged or not.

Each metric indicates a higher quality of multi-label classification, the closer its value is to 100%.

2) RESULTS

The numerical results of AIS data anomaly detection are presented in Tab. 2. The course of the experiment was as follows: 100 AIS messages (from each dataset's test set) was randomly selected and 1 or 2 of their bits (also randomly selected from the examined fields) were artificially damaged (swapped, i.e. changed from 0 to 1 or from 1 to 0) to mimic the effect of a noise being present in a radio channel where the messages are transmitted, causing some of bits to be misinterpreted by the receiver. Next, the proposed anomaly detection algorithm was run. Depending on the classification results and labels of truly damaged fields, the classification quality metrics described earlier were computed.

The results are promising — while using Random Forest, the value of most important metric, recall, did not drop below 75.5% (3. dataset with 524 vessels, 2 damaged bits), mainly varying from 80% to even 100%.

Model with a static threshold has proven to obtain lower recall (71.5% - 99%), however, other metrics (such as precision) were slightly higher there. Apparently this model classifies less fields as damaged, but when it does, most of those fields truly require correction.

The performance of a single Decision Tree (with a *max_depth* hyperparameter also set to 5) was also examined. We can see that the Random Forest model outperforms Decision Tree in most cases: only in the 3rd dataset the recall obtained by Decision Tree was slightly higher.

Naturally, our models work better when the number or corrupted bits was lower and on smaller datasets (that is why

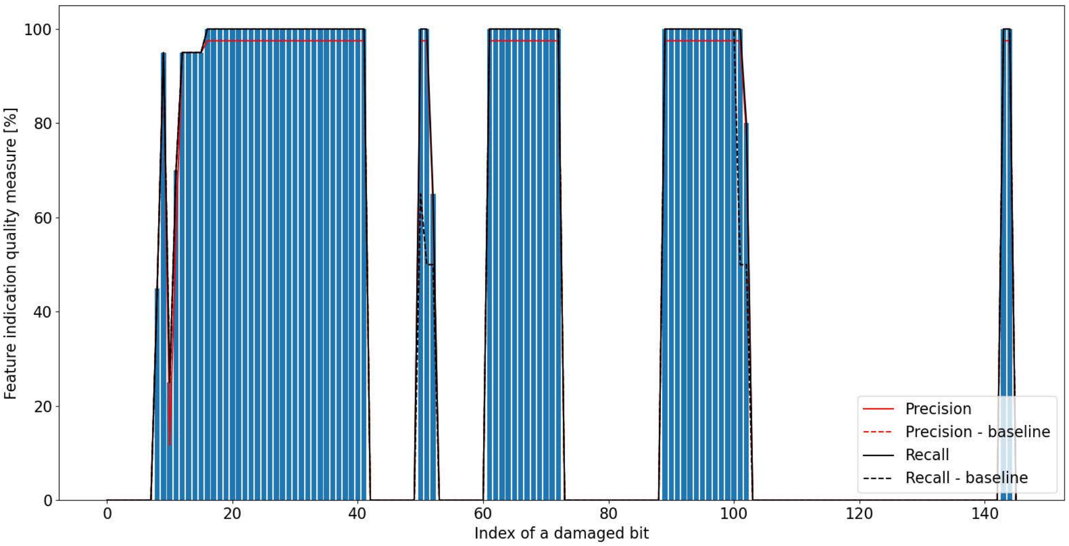

FIGURE 7. Impact of the position of a damaged bit on the quality of detecting the damaged AIS message fields.

TABLE 2. Results of detecting the damaged AIS message fields.

-	Metric	1. dataset (22 vessels)	2. dataset (387 vessels)	3. dataset (524 vessels)
1 damaged bit, Random Forest	Clust. acc.	96.00%	59.00%	72.00%
	Recall	100.00%	100.00%	99.00%
	Precision	93.33%	70.83%	82.50%
	F1	95.33%	78.33%	87.33%
	Jaccard	93.33%	70.83%	82.50%
	Hamming	98.86%	93.93%	97.00%
1 damaged bit, Decision Tree	Clust. acc.	97.00%	62.00%	74.00%
	Recall	99.00%	100.00%	100.00%
	Precision	86.83%	68.17%	76.00%
	F1	90.50%	76.67%	83.17%
	Jaccard	86.83%	68.17%	76.00%
	Hamming	97.79%	93.64%	95.86%
1 damaged bit, model with a threshold	Clust. acc.	96.00%	59.00%	72.00%
	Recall	96.00%	99.00%	96.00%
	Precision	95.50%	72.00%	85.83%
	F1	95.67%	78.83%	88.83%
	Jaccard	95.50%	72.00%	85.83%
	Hamming	99.64%	94.21%	97.93%
2 damaged bits, Random Forest	Clust. acc.	99.00%	68.00%	72.00%
	Recall	81.00%	79.00%	75.50%
	Precision	98.50%	82.00%	86.33%
	F1	86.53%	75.37%	76.13%
	Jaccard	79.92%	64.08%	65.42%
	Hamming	97.00%	93.36%	94.00%
2 damaged bits, Decision Tree	Clust. acc.	100.00%	72.00%	76.00%
	Recall	81.00%	80.00%	76.50%
	Precision	95.17%	79.33%	84.17%
	F1	85.03%	75.00%	75.03%
	Jaccard	78.17%	63.33%	63.50%
	Hamming	96.57%	93.00%	93.64%
2 damaged bits, model with treshold	Clust. acc.	96.00%	68.00%	72.00%
	Recall	96.00%	76.50%	71.50%
	Precision	95.50%	83.17%	90.7%
	F1	95.67%	74.57%	76.40%
	Jaccard	95.50%	63.50%	66.42%
	Hamming	99.64%	93.36%	94.86%

and additional experiment on establishing the optimal batch size is required).

The impact of the position of a damaged bit in an AIS message on the quality of the damaged field classification was also examined. Again, each bit of 20 randomly chosen AIS messages (from test set of 1. dataset) was artificially corrupted (one by one), anomaly detection stage has been run and the quality metrics of multi-label classification (only recall and precision) have been calculated for damaging each bit, if the damage caused the message to form a standalone cluster.

According to the Fig. 7, in most cases the recall keeps very high value for both models: based on Random Forest (marked as a black solid line) or static threshold (dotted black line), higher than precision (red lines). By looking at the blue columns (representing the percentage of messages forming a standalone cluster after having the particular bit corrupted), we can see that most of the time, the lines follow the height of the columns, indicating that if the damaged message forms a standalone cluster, its damaged field is identified correctly. Only when the damage appeared in *speed over ground* field, the model with the treshold (called *baseline* here) struggled to detect some of the damaged bits.

E. COMPARISON AND REVIEW

The last step of the experiment should be the review of the proposed method — in other words, the comparison of the efficiency of our method and others found in the literature in detecting damaged AIS messages or their fields.

However, this task is more complicated than it seems. To the best of our knowledge, most of previously published works focused only on detecting anomalies in two fields among all existing in AIS messages of types 1-3: longitude and latitude, ignoring other important features of a ship trajectory such as its speed, course, MMSI identifier, etc (those fields were only used as features in the proposed models), like in paper [7]. However, method proposed in [18] manages to find anomalies also in speed and heading fields —

the researchers claim their model detected 1482 anomalies among 15261 trajectory points.

Other works define anomalies like a sequence of points (in other words, as a part of a ship trajectory) that do not follow the movement pattern rather than as single outlying points. In [34], the system of detecting anomalous vessels behaviour (parts of trajectories that are unnatural) works with 82.3% accuracy on Nanjing dataset. Another method (statistical approach using gaussian process, proposed in [16]), where the system was trained on data from North-East/South-West Channel, achieved 0.8678 AUC on detecting 22 abnormal vessel tracks.

By comparing the results of our algorithm of damaged AIS message fields detection based on a wavelet transform, which acheived 80%-100% recall, we find the results promising. Moreover, our approach enables finding anomalies not only in latitude and longitude part of vessels trajectories, but also in other data (like speed or course).

VI. CONCLUSION

In this paper a machine learning based approach for detecting damaged AIS messages and its parts is described. This can be considered as the second stage of reconstruction of AIS data that got corrupted due to the packet collision effect.

In the proposed approach it is assumed that AIS data that form standalone clusters (consisting of only one AIS message) after the clustering stage should be considered as damaged since they do not resemble any of other existing data. *k*-NN algorithm is used to decide which ship had sent the message that was labeled as an outlier and Random Forest classifier, supported by wavelet transform, predicts whether a given field in this outlying message has to be reconstructed. Each field has its own trained Random Forest classifier. By iterating over every Random Forest, the multi-label classification is conducted — the labels assigned to a damaged message indicate which of its fields require further correction (it is important that there might bo more than one label assigned, since several fields might contain false values). The recall (percentage of detected damaged fields among all truly damaged fields) of such multi-label field classification varies in most cases between 80% to 100%, which can be considered as a promising result.

In the near future, the research work will focus on finding damaged AIS messages in full-fledged clusters (containing 2 or more datapoints) and on defining how long the observation time should take for optimal anomaly detection (i.e. what should be the value of a single batch to analyse AIS data in order to get the best results in anomaly detection stage).

APPENDIX A
LIST OF ABBREVIATIONS

AIS — Automatic Identification System
DBSCAN — Density-Based Spatial Clustering of Applications With Noise
k-NN — *k* Nearest Neighbours
MMSI — Maritime Mobile Service Identity
SAT-AIS — Satellite Automatic Identification System
VHF — Very High Frequency

ACKNOWLEDGMENT

The authors specially thanks to Marcin Waraksa and Prof. Jakub Montewka from Gdynia Maritime University for sharing the raw data that they used in their experiment.

REFERENCES

[1] N. S. Altman, "An introduction to kernel and nearest-neighbor nonparametric regression," *Amer. Statist.*, vol. 46, no. 3, pp. 175–185, Aug. 1992, doi: 10.2307/2685209.
[2] K. Bao, D. Shang, R. Wang, and R. Ma, "AIS big data framework for maritime safety supervision," in *Proc. ICRIS*, Sanya, China, Nov. 2020, pp. 150–153.
[3] R. Couronné, P. Probst, and A.-L. Boulesteix, "Random forest versus logistic regression: A large-scale benchmark experiment," *BMC Bioinf.*, vol. 19, no. 1, p. 270, Jul. 2018, doi: 10.1186/s12859-018-2264-5.
[4] L. Debnath, "Wavelet transform and their applications," *PINSA-A*, vol. 64, no. 6, pp. 685–713, Nov. 1998.
[5] D. Deng, "Research on anomaly detection method based on DBSCAN clustering algorithm," in *Proc. ISCTT*, Shenyang, China, Nov. 2020, pp. 439–442.
[6] W. Dian-Gang, D. Jin-Chen, H. Lin, and G. Yan, "Anomaly behavior detection based on ensemble decision tree in power distribution network," in *Proc. ICNISC*, Wuhan, China, Apr. 2018, pp. 312–316.
[7] A. Dobrkovic, M.-E. Iacob, and J. Van Hillegersberg, "Maritime pattern extraction from AIS data using a genetic algorithm," in *Proc. DSAA*, Montreal, QC, Canada, Oct. 2016, pp. 642–651.
[8] *Satellite—Automatic Identification System (SAT-AIS) Overview*. Accessed: Jan. 25, 2022. [Online]. Available: https://artes.esa.int/sat-ais/overview
[9] M. Ester, H.-P. Kriegel, J. Sander, and X. Xu, "A density-based algorithm for discovering clusters in large spatial databases with noise," in *Proc. KDD*, Portland, OR, USA, 1996, pp. 226–231.
[10] T. K. Ho, "The random subspace method for constructing decision forests," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 20, no. 8, pp. 832–844, Aug. 1998, doi: 10.1109/34.709601.
[11] *AIS Transponders*. Accessed: Jan. 25, 2022. [Online]. Available: https://www.imo.org/en/OurWork/Safety/Pages/AIS.aspx
[12] *Technical Characteristics for an Automatic Identification System Using Time Division Multiple Access in the VHF Maritime Mobile Frequency Band*, document Recommendation ITU-R M.1371-5, International Telecommunications Union, Feb. 2014.
[13] J. Jin, W. Zhou, and B. Jiang, "Maritime target trajectory prediction model based on the RNN network," in *Artificial Intelligence in China* (Lecture Notes in Electrical Engineering), vol. 572. Singapore: Springer, 2020, pp. 334–342.
[14] M. Karimi, A. Jahanshahi, A. Mazloumi, and H. Z. Sabzi, "Border gateway protocol anomaly detection using neural network," in *Proc. Big Data*, Los Angeles, CA, USA, Dec. 2019, pp. 6092–6094.
[15] I. Kontopoulos, I. Varlamis, and K. Tserpes, "Uncovering hidden concepts from AIS data: A network abstraction of maritime traffic for anomaly detection," in *Proc. MASTER*, Würzburg, Germany, 2020, pp. 6–20.
[16] K. Kowalska and L. Peel, "Maritime anomaly detection using Gaussian process active learning," in *Proc. FUSION*, Singapore, 2012, pp. 1164–1171.
[17] M. Liang, R. W. Liu, Q. Zhong, J. Liu, and J. Zhang, "Neural network-based automatic reconstruction of missing vessel trajectory data," in *Proc. ICBDA*, Suzhou, China, Mar. 2019, pp. 426–430.
[18] B. Lei and D. Mingchao, "A distance-based trajectory outlier detection method on maritime traffic data," in *Proc. ICCAR*, Auckland, New Zealand, Apr. 2018, pp. 340–343.
[19] S. Li, M. Liang, X. Wu, L. Zhao, and R. W. Liu, "AIS-based vessel trajectory reconstruction with U-Net convolutional networks," in *Proc. ICCCBDA*, Chengdu, China, 2020, pp. 157–161.
[20] T. Machado, R. Maia, P. Santos, and J. Ferreira, "Vessel trajectories outliers," in *Proc. ISAmI*, Toledo, Spain, 2018, pp. 247–255.
[21] J. MacQueen, "Some methods for classification and analysis of multivariate observations," in *Proc. 5th Berkeley Symp. Math. Statist. Prob.*, vol. 1, Berkeley, CA, USA, 1967, pp. 281–297.
[22] M. Mieczyńska and I. Czarnowski, "DBSCAN algorithm for AIS data reconstruction," *Proc. Comput. Sci.*, vol. 192, pp. 2512–2521, Oct. 2021.

[23] M. Mieczyńska and I. Czarnowski, "K-means clustering for SAT-AIS data analysis," *WMU J. Maritime Affairs*, vol. 20, no. 3, pp. 377–400, Sep. 2021.

[24] S. Nikfalazar, C.-H. Yeh, S. Bedingfield, and H. A. Khorshidi, "Missing data imputation using decision trees and fuzzy clustering with iterative learning," *Knowl. Inf. Syst.*, vol. 62, no. 6, pp. 2419–2437, Jun. 2020.

[25] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay, "Scikit-learn: Machine learning in Python," *J. Mach. Learn. Res.*, vol. 12, pp. 2825–2830, Nov. 2011.

[26] *Neural Networks vs. Random Forests—Does it Always Have to be Deep Learning?* Accessed: Sep. 1, 2022. [Online]. Available: https://blog.frankfurt-school.de/de/neural-networks-vs-random-forests-does-it-always-have-to-be-deep-learning/

[27] S.-S. Shai and B.-D. Shai, *Understanding Machine Learning: From Theory to Algorithms*. New York, NY, USA: Cambridge Univ. Press, 2014. [Online]. Available: https://www.cs.huji.ac.il/~shais/UnderstandingMachineLearning/understanding-machine-learning-theory-algorithms.pdf

[28] K. P. Sinaga and M.-S. Yang, "Unsupervised K-means clustering algorithm," *IEEE Access*, vol. 8, pp. 80716–80727, 2020.

[29] S. K. Singh and F. Heymann, "Machine learning-assisted anomaly detection in maritime navigation using AIS data," in *Proc. PLANS*, Portland, OR, USA, Apr. 2020, pp. 832–838.

[30] G. M. Swetha, K. Hemavathy, and S. Natarajan, *Overcome Message Collisions in Satellite Automatic ID Systems*. Accessed: Oct. 13, 2022. [Online]. Available: https://www.mwrf.com/technologies/systems/article/21849164/overcome-message-collisions-in-satellite-automatic-id-systems

[31] M. Vespe, I. Visentini, K. Bryan, and P. Braca, "Unsupervised learning of maritime traffic patterns for anomaly detection," in *Proc. DF TT*, London, U.K., 2012, p. 14.

[32] T. Wang, C. Ye, H. Zhou, M. Ou, and B. Cheng, "AIS ship trajectory clustering based on convolutional auto-encoder," in *Proc. IntelliSys*, vol. 1251, 2020, pp. 529–546.

[33] R. Wawrzaszek, M. Waraksa, M. Kalarus, G. Juchnikowski, and T. Gorski, "Detection and decoding of AIS navigation messages by a low earth orbit satellite," in *Aerospace Robotics III* (GeoPlanet: Earth and Planetary Sciences). Cham, Switzerland: Springer, 2019, pp. 45–62.

[34] Z. Xia and S. Gao, "Analysis of vessel anomalous behavior based on Bayesian recurrent neural network," in *Proc. ICCCBDA*, Chengdu, China, Apr. 2020, pp. 393–397.

[35] Z. Zhang, G. Ni, and Y. Xu, "Trajectory prediction based on AIS and BP neural network," in *Proc. ITAIC*, Chongqing, China, Dec. 2020, pp. 601–605.

MARTA SZARMACH is currently pursuing the degree with the Faculty of Electrical Engineering, Białystok University of Technology. Since 2018, she has been employed as an Assistant at the Department of Maritime Telecommunications, Gdynia Maritime University. Her research interest includes problems of using artificial intelligence methods in telecommunications.

IRENEUSZ CZARNOWSKI (Senior Member, IEEE) received the Ph.D. degree in the field of computer science from the Poznan University of Technology, Poland, and the Ph.D. degree in the field of computer science from the Wroclaw University of Science and Technology, Poland. He is currently a Professor of computer science with the Department of Information Systems, Gdynia Maritime University, Poland. He has authored or coauthored more than 100 papers in international journals and conference proceedings. His research interests include artificial intelligence, multi-agent systems, decision support systems, machine learning, artificial neural networks, data mining, data reduction, instance selection, and data preprocessing based on population-based algorithms. From 2016 to 2020, he was the Vice Rector of Research and International Cooperation with Gdynia Maritime University. He is associated with IEEE, since 2016, and is the Chair of the Polish Chapter of IEEE Systems, Man, and Cybernetics.