

**Name:** Rouben Azad

**Rank and service or organization:** LT Navy

**Estimated departure date**1**:** 04/04/2020

**Dashboard link:** https://python.nps.edu/Thesis/Dashboard?idb=MTQ4Nzcx&tid=32771&doctype=

**File-sharing site:** https//github.com/rlazad/Thesis or https://nps01.sharepoint.com/sites/tpo2020/rlazad/SitePages/Home.aspx

**Title:** The Impact of Machine Learning on AIS Data

**Advisor:** Dr. John V. Monaco

**Co-advisor:** Dr. Geoffrey Xie

**Second reader:** n/a

**Distribution statement**2**:** Click here to enter.

**If this is a *co*-authored paper, are you earning *different* degrees?**

**Abstract (not to exceed 1,500 characters, including spaces)3:**

|  |
| --- |
| Automated identification system data is utilized commonly in the open sea and in ports. It is also easily “spoofed” -- meaning, it can be altered before transmission. Some vessels use spoofing to hide their identity or ship loadout. Security risks are on the rise due to frequent use of spoofing data from sea-going vessels, especially in this technological age where data altering is common. This research will focus on building and training a model through machine learning that will be used to make predictions and decision support. The goal is to construct a ship-to-ship network through the AIS database which is comprised of; ship coordinates, dimensions, identification and characteristics. The algorithm will be used to detect changes and abnormal activities over a given period of time. Most vessels out at sea don’t have fraud detection that can prevent against spoofing. This algorithm we develop will learn and detect changes via the data it receives, which could potentially aid in collision avoidance. By gathering information on high-risk vessels, it enables watch standers to make better decisions at sea. |

1 If a team, earliest team member’s.

2 If thesis is restricted, it should be shared only via SharePoint or Box, encrypted email, or hand delivery.

3 To see character count, including spaces, highlight the abstract, click Word’s *Review* tab, then the *Word Count* button.

TABLE OF CONTENTS

I. INTRODUCTION 1

A. Thesis Objective 2

B. Benefits of this study 2

C. Methodology 2

D. Thesis Structure 3

II. Literature review 4

III. Data management 5

appendix. Optional 6

List of References 8

initial distribution list 11

THIS PAGE INTENTIONALLY LEFT BLANK

LIST OF FIGURES

Figure 1. Basic Formatting Steps 1

Figure 2. Heading Levels and Their Associated Styles. Adapted from Hawks (2015). 4

Figure 3. A Basic Figure 6

Figure 4. A Figure with a Citation after the Figure Title in APA and Chicago-Author Date Style. Source: Smith (2017). 7

Figure 5. A Figure Cited Using a Footnote in Chicago Notes and Bibliography Style. 8

Figure 6. Placement of Optional Secondary Caption in Figure Title 9

Figure 7. Variation—Multi-line Figure Title, with First Sentence Only in List of Figures. Adapted from Doe (2017). 10

Figure 8. Variation—Figure Title above Figure 11

THIS PAGE INTENTIONALLY LEFT BLANK

LIST OF TABLES

Table 1. Styles to Use and Element Placement for Figures and Tables. Source: [5]. 13

THIS PAGE INTENTIONALLY LEFT BLANK

LIST OF ACRONYMS AND ABBREVIATIONS

AIS Automatic Identification System

SOG Speed Over Ground

COG Course Over Ground

ML Machine Learning

NPS Naval Post Graduate School

MMSI Maritime Mobile Service Identity

DIM Dimension

ROT Rate of Turn

PIM Position of Intended Movement

CPA Closest Point of Approach

GPS Global Positioning System

NOAA National Oceanic and Atmospheric Administration

[list of classifier acronyms to follow]

THIS PAGE INTENTIONALLY LEFT BLANK

EXECUTIVE SUMMARY

This is NOT your abstract. Your abstract must be entered into your thesis dashboard in Python, and it will output as part of your First 8 Pages file.

An executive summary is required or recommended for the departments listed in our Abstracts versus Executive Summaries guidance [here](https://my.nps.edu/web/thesisprocessing/resources), under the “Style and Grammar” section. Please consult your advisors for their preferences regarding length and style.

The executive summary should be able to stand apart from the thesis. Therefore, if you include figures or tables in your Executive Summary, do not apply the figure or table styles to the titles. Instead, use **Normal** style and manually format the title to match the titles in the body of the thesis. This will keep these titles out of the lists of figures and tables, and allow the figure and table numbering to start at “1” in the thesis body, as required.

If you include *parenthetical* citations (APA style, for example) in the Executive Summary (but weaving your sources directly into your sentences is preferable), include a separate reference list at the bottom of the last page of the summary. Use the same citation/reference format as in the body of the thesis.

**References**

If you are using IEEE, Chicago Author-Date, or APA citation style, include a separate reference list for the executive summary.

.

THIS PAGE INTENTIONALLY LEFT BLANK

ACKNOWLEDGMENTS

I would like to thank my future wife, advisors and cohort for their support in this new field I’ve experienced in my educational years at NPS. To further strengthen my and motivate me in this process, thank you. (more to add)

THIS PAGE INTENTIONALLY LEFT BLANK

# INTRODUCTION

An Automated Identification System (AIS) [2] is the primary source of identification for surface vessels on the high seas. Almost all ships are equipped with them, and they are used frequently by other vessels to aid in collision avoidance. The information that AIS provides, such as, unique identification, course, position, speed, tonnage, and ship dimensions, is especially useful in and around ports. Port controls use this data to direct traffic and maintain channel control. Maritime authorize and asses this data, and physically verify the vessels for security, or otherwise.

Although this data is relied on heavily by all factors listed above, and more, it is easily “spoofed” -- meaning, it can be altered before transmission. Some vessels use spoofing to hide their identity or ship loadout. Security risks are on the rise due to frequent use of spoofing data from sea-going vessels, especially in this technological age where data altering is common. [1]

Artificial intelligence (AI) and machine learning have been a new facet of discussion. This fairly new technique can have a powerful impact on information sharing across vessels because it has the ability to predict and prevent security risks at sea.

Our main steps in building a machine learning algorithm are; preparing the data, choosing our model, training, evaluating, tuning and prediction. This will be discussed further in the methodology section. [6]

AIS data will be used to train a model to make predictions and decisions without explicit programming. The goal is to find outliers in this data by the machine’s use of pattern recognition, classification, and regression algorithms, which will point us to our anomalies.

This machine learning algorithm will parse the AIS data to detect and categorize anomalies which will help maritime authorities gain deeper knowledge on risky vessels.

Most vessels out at sea don’t have fraud detection that can prevent against spoofing. Automated anomaly detection could potentially aid in collision avoidance. By gathering information on high-risk vessels, it also enables watch standers to make better decisions at sea.

## Thesis Objective

This research will demonstrate how a machine learning algorithm can effectively identify a wide variety of anomalies in vessel dimensions, identification, ship to ship networking, and movement behaviors in an Automated Identification System (AIS) database. By analyzing this data through machine learning, we can show a whole new field of possibility for maritime security by supplying them with additional tools to enable better security procedures out at sea and ports. Machine learning is now being used across all fields of technology and the United States Department of Defense can benefit greatly by this implementation. Sea-going vessels are notorious for using spoofing technologies to alter AIS data. This thesis will involve development of an algorithm that can spot these anomalies and categorize vessels as seen fit.

This thesis will show how a machine learning algorithm can be used to detect abnormal activities extracted from this AIS data. This algorithm will be able to categorize vessels by their maneuvers, behaviors, and network connections by using pattern recognition, observing anomalies, and finding outliers.

## Benefits of this study

If our hypothesis is correct, this work will help maritime authorities gain deeper knowledge on high-risk vessels and help eliminate misinformation and illegal activities at sea. This will also equip watch standers with an improved situational awareness aided by the use of this algorithm, which will enhance the information received by AIS for better decision making and judgment control.

## Methodology

This thesis will involve working with programming languages such as python, with additional modeling interfaces for analyzing and assessing data. The tools that will be utilized for this work are IPython for its interactive shell commands that support data visualities and graphical user interface (GUI) toolkit; and Python’s Pandas and Matplotlib, which are both great tools for data analysis and modeling. Matplotlib will be primarily used for plotting libraries in its 2D plotting interface.

The first step in the process is data preparation, by loading our data into a suitable place. Then begin parsing the data using an established algorithm with a few modifications to support the changing data structures. After completion of the parsing and organization, a model will be constructed that suits the data. Afterword’s we will train our model with the given data, and incrementally improve its ability to make better predictions. Once the training is complete evaluation beings to test the model’s accuracy. From the evaluations we can decide to further improve our training by tuning our parameters. And lastly the prediction step, we use our model to predict the correct answers. [6]

This machine-learning algorithm will be constructed using the previously stated methods. There will be multiple levels of AI to be established in this structure, beginning with data analyzation using 2D models with Matplotlib; machine-learning acquisition and testing; data categorization; and ultimately, anomaly detection of real-world historic AIS data.

The second step will be analyzing large amounts of data over a long period of time using the first step. Afterwards, the algorithm will be put through validation to narrow down errors and complete the program.

Finally, a result of the stress tests will be documented and improved upon. In turn, this program will be used to analyze data in real time to provide a quick response to maritime security.

## Thesis Structure

In the follow-on chapters we will cover Literature Review in Chapter II, Data Management in Chapter III, Modeling and Analysis in Chapter IV and Summary and Future Work in Chapter V. In Chapter II we will cover, all related fields of research that have been done at NPS regarding AIS, anomaly detection and vessel network distribution. The focus on these topics are particularly helpful in our research and could help improve and further the study in this subject. In Chapter II we will concentrate on data management and manipulation to extract the desired information necessary to expand our hypothesis. This will be done through various methods that involves transformation and matrix building to allow for data arrangement. We will also define different tools and techniques used to model our data and portray vessel interaction, trajectories and ship networking. In Chapter IV we showcase the different methods applied and their findings via graphical and analytical methods to further enhance the usage of the machine learning algorithm. In Chapter V we conclude our finding and elaborate on the successes and failures of our modeling techniques. We further pose more solutions that could be established to advance our objective through additional research.

# Literature review

Recent events in naval history suggest the use of AIS data to be more meaningful. On 21 Aug 2017 USS John S. McCain collided with a commercial vessel near Singapore in a highly congested traffic separation scheme,[3] due to errors on behave both vessels and watch standers. There are other shipping related collision that occur on occasion. AIS data and machine learning tools could be used to limit these collisions by understanding networking behaviors of vessels and their trajectories. Various research has been done on this topic; vessel trajectory prediction, anomaly detection by use of machine learning classifiers such as Neural networks, random forest, and more that are described in this chapter. Using machine learning we can further assess the influence of AIS by characterizing each vessel and using different classifiers to output desired results. In this chapter we will discuss the studies and methods implemented in previous research.

Ristic, La Scala, Morelande, and Gordon (2008) have built an anomaly detection algorithm from historic AIS data and historic motion pattern data by using adaptive kernel density estimation and Gaussian sum tracking filter.

Mou, Tak, Ligterngen (2010) have combined AIS data with radar data to increase the effectiveness of trajectory projection and prediction. By using an algorithm that implements the CPA in a real time scenario, along with AIS data to predict vessel trajectories.

Morris and Trivedi (2011) use trajectory clustering and hidden Markov models to construct routes and predict vessel maneuvers.

Pallottta, Vespe, Bryan (2013) have constructed a rule-based system to capture anomalous activities. These rules involve; COS, SOG, DIM, and area-based navigation as attributes in their machine learning model. Their main goal was to acquire data on vessels engages in illegal activities.

Bay (2017) uses clustering on AIS data to measure weather effects and identify navigation routes. This data is based in Port Fourchon, LA and it examines the northern Gulf of Mexico’s oil structure and function as prediction model for vessel trajectories, which could prove useful in navigating through degraded weather.

Young (2017) uses random forest classifier to predict vessel locations at sea by analyzing historic vessel trajectories from AIS data.

Cull (2018) uses historic AIS data based in South China Sea to build a gradient boosted machine to asses the validity of certain vessels operating in that region. She uses two different models, the first is to capture the dependent nature of AIS signals in a historic database, the second to reduce the dependency between the signals and characterize the patterns of vessels in the region.

Liraz (2018) uses a recurrent neural network model on AIS data to predict vessel behavior and detect anomalies.

These researches all seek to provide a goal of a more efficient and safer navigation on the high seas and waterways. Machine learning models are being improved on each year, this gives researchers leverage over this unique tool, that could help secure sea routes of piracy, establish maritime security, aid in dense traffic navigation and an overall safer transit on the ocean.

# Data management

In collaborating with the coast guard, we’ve accumulated AIS data for 3 different ports in the United Sates: San Diego, San Francisco and Norfolk Virginia to start our initial testing and analysis. The data is comprised of 21 columns that defines each vessel with various details such as latitude and longitude, speed, heading etc., we will further elaborate on these in future sections. All 3 data sets are historic, and they rang in a year period from January 2018 through Jan 2019. The AIS data shows all vessels that transited the area with the given Latitude/Longitude roughly covering the size of the harbor.

Each dataset is broken up by moths totaling 12 sheets. Initially python is used to concatenate the data into a data frame for further analysis and feature selection. A graphics tool is used to plot the data without any manipulation to observe the normal state. Furthermore, other various plots are displayed to show the effects of the feature selection.

# appendix. Optional

Appendix titles are also styled as **Heading 1**, minus a roman numeral—backspace to remove the roman numeral. Then, type “Appendix,” two spaces, a letter, and a title: “APPENDIX A. DATA.” *However,* *if you have only one appendix, do not add the letter “A.”*

If you apply Heading 2 style in your appendices, and the lettering does not begin with “A,” right click on the first Heading 2 of the appendix, and choose “Restart at 1.”

THIS PAGE INTENTIONALLY LEFT BLANK

# List of References

Apply **Reference List** style to your list of references to re-create or retain the formatting of this page. Remove manual line spaces that you have entered between entries, as the style comes with the proper spacing.

A *bibliography* is uncommon in NPS theses; a list of references is the standard. A bibliography differs from a reference list in that it also includes sources you consulted, but did not cite.

All in-text citations must have a matching entry in the List of References, with few exceptions; consult your citation style guide.

Use an established citation style such as Chicago, APA, AMS, etc. Made-up or hybrid styles will not be accepted. You are required to use a [department-required or advisor-approved](https://libguides.nps.edu/citation) citation style. Guides to the most-common citation styles used at NPS are available here: [Citation guides.](https://libguides.nps.edu/citation)

Again, if you use reference-list generating software, such as RefWorks, ensure that you fill in all fields completely and accurately when creating your citation list. **You must edit references for punctuation and formatting** **after importing them**.

To edit most lists, you must remove the field code. Do this by highlighting all entries and pressing **Shift** + **Control** + **F9** *or* **Control** + **6** at the same time. In Word’s citation manager, this is achieved by clicking on the list and choosing “convert to static text.”

Here are a few example entries:

Hawks, Mathew A. “Graph-Theoretic Statistical Methods for Detecting and Localizing Distributional Change in Multivariate Data,” Ph.D. diss., Naval Postgraduate School, Monterey, CA, 2015. (**Chicago N-B style**)

Naval Postgraduate School. (2017). Thesis\_template\_times [Word template]. Monterey, CA: Naval Postgraduate School. Retrieved from https://my.nps.edu  
/documents/105790666/106471216/Thesis\_Template\_Times.docx (**APA style**)

[1] B. Orend, *Morality of War*, 2nd ed. Tonawanda, NY: Broadview Press, 2013. (**IEEE style**)

Kyle Wilhoit and Marco Balduzzi, “Vulnerabilities Discovered in Global Vessel Tracking Systems.” TrendLabs Security Intelligence Blog, 8 Aug. 2014, blog.trendmicro.com/trendlabs-security-intelligence/vulnerabilities-discovered-in-global-vessel-tracking-systems/.

“What Is the Automatic Identification System (AIS)?” *MarineTraffic Help*, Marine Traffic, help.marinetraffic.com/hc/en-us/articles/204581828-What-is-the-Automatic-Identification-System-AIS-.

Natalija Jolić, Mato Šimić, and Sanja Brnadić-Zoranić. “Strategy of Implementing Automated Identification System -ITS in Port Systems.” *Promet (Zagreb)* 15.5 (2003): 333–337. Web.

Marco Balduzzi, Alessandro Pasta, and Kyle Wilhoit, “A Security Evaluation of AIS Automated Identiﬁcation System.” *A Security Evaluation of AIS Automated Identiﬁcation System*, 2 Jan. 2015, www.iseclab.org/people/embyte/papers/ais\_acsac14.pdf.

Lynch, Robert. “EMC ANALYSIS OF UNIVERSAL AUTOMATIC IDENTIFICATION AND PUBLIC CORRESPONDENCE SYSTEMS IN THE MARITIME VHF BAND .” *Transition*, Feb. 2004, transition.fcc.gov/omd/dataquality/peer-reviews/ais/source-1.pdf.

“IEC TECHNICAL COMMITTEE 80: MARITIME NAVIGATION AND RADIOCOMMUNICATION EQUIPMENT AND SYSTEMS.” INTERNATIONAL ELECTROTECHNICAL COMMISSION IEC. *Making Electrotechnology Work for You*, 2011, www.iec.ch/about/brochures/pdf/technology/maritime\_nav\_rc\_2.pdf.

Yufeng, and Yufeng G. “The 7 Steps of Machine Learning.” *Towards Data Science*, 31 Aug. 2017, towardsdatascience.com/the-7-steps-of-machine-learning-2877d7e5548e.

Kaluza, Pablo, et al. “The Complex Network of Global Cargo Ship Movements.” *Royal Society Publishing,* 19 Jan. 2010, royalsocietypublishing.org/doi/pdf/10.1098/rsif.2009.0495.

Cull, Kimberly M. “Classifying Vessels Operating in the South China Sea by Origin with the Automatic Identification System.” *Calhoun*, Naval Postgraduate School, Mar. 2018, calhoun.nps.edu/bitstream/handle/10945/58289/18Mar\_Cull\_Kimberly.pdf.

Liraz, Shay Paz. “SHIPS’ TRAJECTORIES PREDICTION USING RECURRENT NEURAL NETWORKS BASED ON AIS DATA.” *Calhoun*, Naval Postgraduate School, Sept. 2018, calhoun.nps.edu/bitstream/handle/10945/60431/18Sep\_Liraz\_Shay\_Paz.pdf.

Hintze, Jhon R. “An Analysis of Vessel Waypoint Behavior through Data Clustering.” *Calhoun,* Naval Postgraduate School, Sept. 2017, calhoun.nps.edu/bitstream/handle/10945/56135/17Sep\_Hintze\_John.pdf.

[3]https://www.navytimes.com/news/your-navy/2019/08/06/new-report-blames-big-navy-for-fatal-collision/

THIS PAGE INTENTIONALLY LEFT BLANK

# initial distribution list

1. Defense Technical Information Center

Ft. Belvoir, Virginia

2. Dudley Knox Library

Naval Postgraduate School

Monterey, California