NAVAL POSTGRADUATE SCHOOL

MONTEREY, CALIFORNIA

THESIS

APPLYING ARTIFICIAL INTELLIGENCE TO IDENTIFY CYBER SPOOFING ATTACKS AGAINST THE GLOBAL POSITIONING SYSTEM

by

Rohan Kennedy

September 2021

Thesis Advisor:	Bonnie W. Johnson
Co-Advisor:	James Baker (MCTSSA)
Second Reader:	Ying Zhao

THIS PAGE INTENTIONALLY LEFT BLANK

REPORT DOCUMENTATION PAGE

Form Approved OMB No. 0704-0188

1. AGENCY USE ONLY
(Leave blank)

2. REPORT DATE
September 2021

3. REPORT TYPE AND DATES COVERED
Master's thesis

4. TITLE AND SUBTITLE
APPLYING ARTIFICIAL INTELLIGENCE TO IDENTIFY CYBER SPOOFING ATTACKS AGAINST THE GLOBAL POSITIONING SYSTEM

5. FUNDING NUMBERS

6. AUTHOR(S) Rohan Kennedy

7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)
Naval Postgraduate School
Monterey, CA 93943-5000

8. PERFORMING ORGANIZATION REPORT NUMBER

9. SPONSORING / MONITORING AGENCY NAME(S) AND ADDRESS(ES)
N/A

10. SPONSORING / MONITORING AGENCY REPORT NUMBER

11. SUPPLEMENTARY NOTES The views expressed in this thesis are those of the author and do not reflect the official policy or position of the Department of Defense or the U.S. Government.

12a. DISTRIBUTION / AVAILABILITY STATEMENT
Approved for public release. Distribution is unlimited.

12b. DISTRIBUTION CODE
A

13. ABSTRACT (maximum 200 words)

Interference on the Global Positioning System (GPS) infrastructure poses a threat to the nation's security and economy as systems become more dependent on the technology. The pervasiveness of GPS interference methods such as jamming and spoofing present multiple opportunities for adversaries to infiltrate and inject false data on systems as diverse as military, banking, shipping, ecommerce, transportation and other critical economic sectors. The study of GPS spoofing detection methods requires innovative and novel schemes to meet the challenge posed. With the increasing processing power of computer systems, artificial intelligence methods have become a prime candidate for application to the detection and reporting of these cyber threats. This thesis studied the application of machine learning and data analytics to identify false data injection attempts on military GPS. The study combined live and simulated GPS message traffic data to train and test machine learning algorithms to identify the threats. Applying both unsupervised and supervised learning methods to the dataset helped advance the study of the GPS spoofing problem and proved to be effective tools to monitor GPS traffic while serving as another layer of security to the GPS infrastructure.

14. SUBJECT TERMS
GPS, machine learning, data analytics, artificial intelligence, assured PNT

15. NUMBER OF PAGES
135

16. PRICE CODE

17. SECURITY CLASSIFICATION OF REPORT	**18. SECURITY CLASSIFICATION OF THIS PAGE**	**19. SECURITY CLASSIFICATION OF ABSTRACT**	**20. LIMITATION OF ABSTRACT**
Unclassified	Unclassified	Unclassified	UU

THIS PAGE INTENTIONALLY LEFT BLANK

APPLYING ARTIFICIAL INTELLIGENCE TO IDENTIFY CYBER SPOOFING ATTACKS AGAINST THE GLOBAL POSITIONING SYSTEM

Rohan Kennedy
Civilian, Department of the Navy
BS, Marquette University, 1997
MS, Marquette University, 2006

Submitted in partial fulfillment of the
requirements for the degree of

MASTER OF SCIENCE IN SYSTEMS ENGINEERING MANAGEMENT

from the

**NAVAL POSTGRADUATE SCHOOL
September 2021**

Approved by: Bonnie W. Johnson
Advisor

James Baker
Co-Advisor

Ying Zhao
Second Reader

Oleg A. Yakimenko
Chair, Department of Systems Engineering

THIS PAGE INTENTIONALLY LEFT BLANK

ABSTRACT

Interference on the Global Positioning System (GPS) infrastructure poses a threat to the nation's security and economy as systems become more dependent on the technology. The pervasiveness of GPS interference methods such as jamming and spoofing present multiple opportunities for adversaries to infiltrate and inject false data on systems as diverse as military, banking, shipping, ecommerce, transportation and other critical economic sectors. The study of GPS-spoofing–detection methods requires innovative and novel schemes to meet the challenge posed. With the increasing processing power of computer systems, artificial intelligence methods have become a prime candidate for application to the detection and reporting of these cyber threats. This thesis studied the application of machine learning and data analytics to identify false data injection attempts on military GPS. The study combined live and simulated GPS message traffic data to train and test machine learning algorithms to identify the threats. Applying both unsupervised and supervised learning methods to the dataset helped advance the study of the GPS spoofing problem and proved to be effective tools to monitor GPS traffic while serving as another layer of security to the GPS infrastructure.

THIS PAGE INTENTIONALLY LEFT BLANK

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

THIS PAGE INTENTIONALLY LEFT BLANK

LIST OF ACRONYMS AND ABBREVIATIONS

AFATDS	advanced field artillery tactical data system
AFB	Air Force Base
AI	artificial intelligence
ANN	artificial neural networks
ARL	Army Research Laboratory
AS	anti-spoofing
AUC	area under curve
BP	belief propagation
CA	classification accuracy
C/A	coarse/acquisition
C/No	carrier-to-noise ratio
CART	Classification and Regression Trees
CDMA	Code Division Multiple Access
COMSEC	communications security
CONOPS	concept of operations
COTS	commercial-off-the-shelf
CS	control segment
DA	data analytics
DAGR	Defense Advanced GPS Receiver
DAR	data at rest
DIT	data in transit
DOD	Department of Defense
DOP	dilution of precision
DoS	denial of service
DoT	Department of Transportation
DSSS	direct sequence spread spectrum
EHE	estimated horizontal error
EPE	estimated position error
EVE	estimated vertical error
EW	electronic warfare

FCC	Federal Communications Commission
FDI	false data injection
FOM	figure of merit
FoS	Family of Systems
FPR	false positive rate
GCCS-TCO	Global Combat and Control System Tactical Combat Operations
GDOP	geometric dilution of precision
GGA	global positioning system fix data
GLONASS	Globalnaya Navigazionnaya Sputnikovaya Sistema, or Global Navigation Satellite System
GNSS	Global Navigation Satellite System
GPS	Global Positioning System
GSA	GNSS DOP and active satellites
GSV	GNSS satellites in view
HDOP	horizontal dilution of precision
IA	information assurance
ICD	Interface Control Document
ISR	intelligence, surveillance, and reconnaissance
JBC-P	Joint Battle Command Platform
J/S	jammer-to-signal ratio
JTCW	Joint Tactical Common Operational Picture (COP) Workstation
km	kilometers
KNN	k-nearest neighbor
kph	kilometers per hour
MEO	medium earth orbit
MGUE	Military GPS User Equipment
ML	machine learning
NAL	Notice of Apparent Liability for Forfeiture
NavIC	Navigation with Indian Constellation
NCCIC	National Cybersecurity & Communications Integration Center
NIST	National Institute of Standards and Technology
NMEA	National Marine Electronics Association

NN	neural network
NTP	Network Time Protocol
P-code	precise ranging code
PCA	principal component analysis
P(Y)-code	precise ranging code or encrypted precise ranging code
PDOP	position dilution of precision
PNT	positioning, navigation, and timing
PPS	precise positioning system
PRN	pseudo random noise
PVT	position, velocity, and time
QZSS	Quasi-Zenith Satellite System
RF	Radio Frequency
RMC	Recommended Minimum Specific GNSS Data
RNN	recurrent neural network
ROC	receiver operating characteristic curve
SA	selected availability
SDR	Software Defined Radio
SNR	signal to noise ratio
SPS	standard positioning service
SV	satellites in view
SV	space vehicle
SVM	support vector machines
SWaP	size weight, and power
T&E	test and evaluation
THS	Target Handoff System
TDOP	time dilution of precision
TFOM	time figure of merit
TOA	Time Of Arrival
TPR	true positive rate
UAV	unmanned aerial vehicle
USNO	United States Naval Observatory
UT	University of Texas at Austin

UTC	Coordinated Universal Time
VDOP	vertical dilution of precision
Vel	velocity
VMF	Variable Message Format
Y-code	encrypted precise ranging code

EXECUTIVE SUMMARY

Can artificial intelligence (AI) methods detect spoofing on the military Global Positioning System (GPS) infrastructure? Using AI and machine learning (ML) tools, we demonstrated the successful detection of spoofing on the Defense Advanced GPS Receiver (DAGR). Using systems engineering principles, we conducted an analysis of the problem space to include a literature review to identify the state of the art in AI. The result of this exploration revealed novel solutions applied to solve this problem. During the early stages, we considered various system designs before settling on a system that incorporated both live and simulated GPS message traffic. We integrated model-based systems engineering (MBSE) principles to the design concept to map the system levels and interactions. Humphreys et al. (2008) defines GPS spoofing threat by three techniques classed as simplistic, intermediate, and sophisticated attacks. A simplistic attack builds on the concept of using a commercial GPS signal simulator, amplifier, and antenna to broadcast signals towards target GPS receiver. An intermediate spoofing attack applies receiver-based spoofers to generate the spoofing signal towards the target receiver's antenna. A sophisticated spoofing attack is the most complex of the three methods, has the capacity to vary both the carrier and code phase outputs transmitted by each antenna while controlling the relative code/carrier phases among the transmit antennas (Humphreys et al. 2008). Because successful GPS spoofing attacks impact time, frequency, and space domains, the developed system, at a minimum, must consider these parameters. The design concept employed requirements of identifying nonobvious and nontrivial relationships in the dataset.

The system design followed a two-pronged approach; 1) develop a hardware system to inject a spoofing signal on the GPS infrastructure and 2) develop a software application to detect the injection of spoofing. The hardware system consisted of a GNSS simulator used to create spoofing scenarios, a radio frequency (RF) splitter to facilitate the input of both live and simulated message traffic, a DAGR and various data collection tools. The system operations followed the simplistic spoofing attack technique to execute overt spoofing attacks. A feature of overt spoofing is the jam-then-spoof strategy. Chapman

(2017, 1) describes overt spoofing attack as one where “the counterfeit GPS signals are simply broadcast at a significantly higher power level than the authentic satellite signals.” In overt spoofing the adversary increases the power of the spoofing signal to overpower the legitimate GPS signal feed. We successfully applied overt spoofing techniques to the engineered system and collected the data for analysis. The dataset formed the basis of the AI development tools and consisted of both National Marine Electronics Association 0183 (NMEA 0183) and Interface Control Document-GPS 153 (ICD GPS153) message traffic. While the NMEA 0183 standard defines GPS message for commercial use, the ICD 153 standard is used in the design and implementation of messages used on military platforms. We used messages from both the NMEA 0183 and the ICD 153 message standards in this study.

Applying data reduction tools such as principal component analysis (PCA) on the dataset revealed the correlation of the parameters resulting in approximately 94% of the variance of dataset. The first principal component PC1 explains the variance. The study of AI tools identified applicability of both unsupervised and supervised learning tools. Unsupervised learning is effective in identifying characteristics within a dataset, while supervised learning methods are applied to datasets that have a known target. Using clustering methods such as *k*-means, we clearly identified the clusters formed by the application of spoofing on the signal. Clustering is effective as a visual tool. The unsupervised learning models effectively identified clusters formed by the spoofing scenarios. The onset of spoofing on the data structure was revealed in a separate cluster from the clusters formed before and after the application of the spoofing signal. We identified peculiarities and previously unidentified correlations within the data parameters that proved enlightening to the study.

Using data mining and data analytics tools, we again processed the dataset to apply a labeled parameter and trained a supervised model to classify spoofing. We processed the dataset and examined the results using several supervised learning models. We executed the models on the labelled dataset with 85% of the data used for training and 15% reserved for testing in one while using cross validation. Applying cross-validation to the model eliminates the need for a validation split of the dataset. The results of the random forest and logistic regression models show a 100% true positive rate on both the training set and the test set further proving that the AI model can detect spoofing on the GPS user infrastructure.

We evaluated the effectiveness of supervised learning models using a suite of performance measures that are generally applied to ML, data science, and statistics problems. The training of the models presented excellent results with perfect recall and precision for all models. Recall is an important metric used in assessing the effectives of a tool in the detection of malicious activity such as spoofing attempts on the DAGR. The results of this study revealed that given the appropriate tools and access, an adversary can effectively spoof a military GPS device. The tools we developed and demonstrated throughout the thesis show that AI methods can detect spoofing attacks on the military GPS infrastructure.

References

Humphreys, Todd E, Brent M Ledvina, Mark L Psiaki, Brady W O'Hanlon, and Paul M Kintner Jr. 2008. *Assessing the Spoofing Threat: Development of a Portable GPS Civilain Spoofer.* Conference, Savanna: ION GNSS Conference.

Chapman, Adam. 2017. GPS Spoofing. ECE Senior Capstone Project, Medford: Tufts University Department of Electrical and Computer Engineering.

THIS PAGE INTENTIONALLY LEFT BLANK

ACKNOWLEDGMENTS

“I can do all things through Christ which strengtheneth me.” – Philippians 4:13 KJV

I would like to express my gratitude and appreciation to my advisors, Dr. Bonnie Johnson, Dr. James Baker, and Dr. Ying Zhao for providing me with the guidance to ensure success in this endeavor. I want to also express my gratitude to the Marine Corps Systems Command (MCSC) and my Command, Marine Corps Tactical Systems Support Activity (MCTSSA) for supporting me through this process.

Special thanks to my dear wife, Wairimu, and my daughters, Ebony and Wambui, for their patience and understanding in accommodating my absences and allowing me to focus on this work.

THIS PAGE INTENTIONALLY LEFT BLANK

I. INTRODUCTION

This chapter presents an overview of the Global Positioning System (GPS) and the rationale for the study being undertaken in this thesis. In addition, the outline for the rest of the work is given to provide a foundation for understanding the methods applied to achieve the results.

A. BACKGROUND AND PROBLEM STATEMENT

GPS technology is ubiquitous in current positioning and navigation applications covering both civilian and military systems. In an article in *IEEE Spectrum*, the authors report that in 2011, Iran captured a classified drone belonging to the CIA (Psiaki, Humphreys, and Stauffer 2016). It is believed that the Iranians were able to interfere with the drone's GPS to make it land in Iran rather than the planned destination of Afghanistan (Psiaki, Humphreys, and Stauffer 2016). The U.S. Department of Transportation (DoT) tracks and documents GPS issues on its public-facing web portal. Recent searches of the portal report that there have been "multiple instances of significant GPS interference reported worldwide in the maritime domain, resulting in lost or inaccurate GPS signals affecting bridge navigation, GPS-based timing, and communications equipment" (U.S. DoT 2020) In 2013, researchers from the University of Texas demonstrated that a GPS receiver could be tricked by broadcasting counterfeit GPS signals to send a set of false coordinates to an unsuspecting user (Newberry 2014). In another report, the Iranian military captured two U.S. Navy patrol boats with 10 U.S. servicemen onboard in Iranian waters (Psiaki, Humphreys and Stauffer 2016). These are just a few instances of high-profile GPS interference that have been reported. With the dependence on GPS systems for navigation, the conclusion is that navigation systems are vulnerable to GPS interference. This thesis explores the question: Can an end user detect and identify a spoofing attempt on their receiver?

The GPS was developed by the U.S. Department of Defense (DOD) and serves marine, airborne, and terrestrial users. Multiple sources associate the words "position," "navigation," and "timing" with GPS. A synthesized definition of GPS is "a satellite navigation system capable of operating in all-weather conditions, providing users with positioning, navigation, and timing (PNT) services" (GPS 2014). The GPS system is composed of three components:

> 1. The space segment, which consists of a constellation of satellites that transmits radio signals to users.
> 2. The ground control segment, which consists of a global network of ground facilities that track, monitor transmissions, perform analyses, and send commands and data to the satellite constellation.
> 3. The user segment, which consists of both vehicle borne and handheld portable devices. (GPS 2014)

Navigation is not the only function provided by GPS; "communications networks, banking systems, financial markets, and power grids all depend on GPS for precise time synchronization" (GPS 2014). Figure 1 is a depiction of the components of the GPS architecture. The user segment shown in Figure 1 was the focus of this thesis. Although the study was geared towards spoofing of military systems, instances of commercial user systems were also explored and evaluated. These instances are documented in later sections of the study.

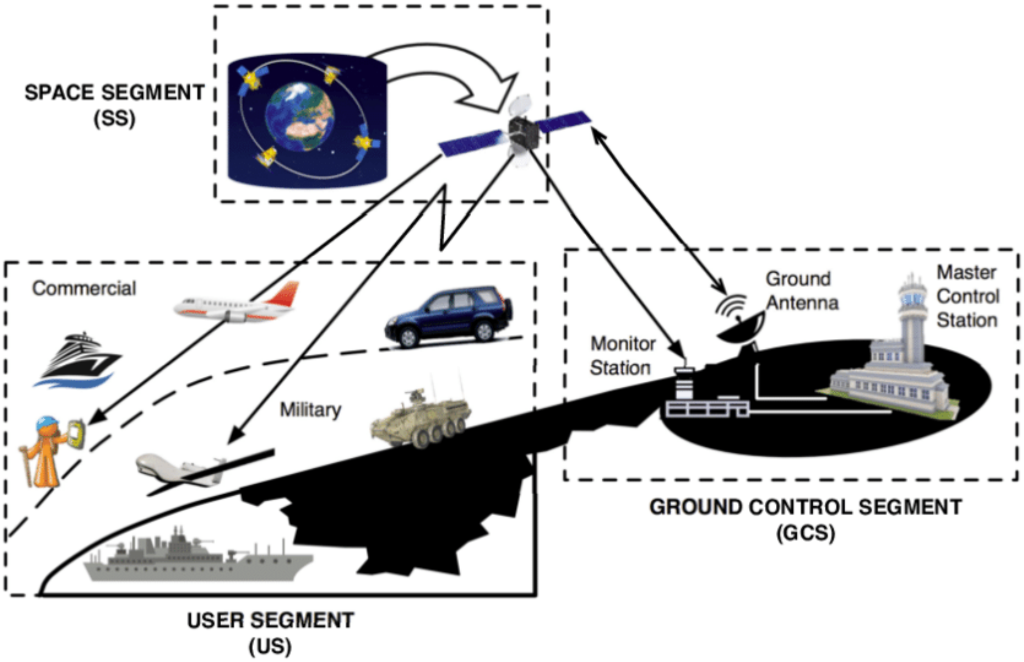


Figure 1. GPS System Components. Source: Ferre (2018).

As a military asset, GPS is used in conducting land, sea, and airborne navigation in areas such as search and rescue operations, vehicle location systems, friendly forces tracking, and aerial refueling operations. A 2012 white paper from NovAtel explains that U.S. defense officials acknowledge that GPS was a force multiplier in both Operations Desert Storm and Desert Shield (NovAtel 2012). According to the paper, military vehicles were able to navigate using GPS and not rely on landmarks as points of reference. Former Defense Secretary James N. Mattis, in a 2018 speech, stated that "Success does not go to the country that develops a new technology first, but rather, to the one that better integrates it and more swiftly adapts its way of fighting" (Garamone 2018). GPS is one such application of technology that has been developed and adapted to the benefit to the U.S. military forces. Global Navigation Satellite Systems (GNSS) provides GPS service to a global audience. The U.S.-operated GPS is the first of the international operational GNSSs. Russia has the *Globalnaya Navigazionnaya Sputnikovaya Sistema* (GLONASS), China has the BeiDou Navigation Satellite System, and the European Union has the Galileo system (GPS 2014a). While the four GNSSs have a global reach, both Japan and India operate regional GPS systems. Japan operates the Quasi-Zenith Satellite System (QZSS), and India operates the Navigation with Indian Constellation (NavIC) system (Kaplan and Hegarty 2017). GPS is the first operational GNSS with worldwide availability.

With all the benefits of GPS, the technology is not without weaknesses and vulnerabilities. There are many types of threats, both natural and manmade, that can interfere with a GNSS receiver's functionality. Threats can be manifested as intentional or unintentional. Table 1 shows a list of threats and their impacts on the GPS infrastructure.

Table 1. Types of Threats to GPS. Source: ICAO (2019, 6–7).

Threat Source	Threat Type	Description	Impact on the User
Solar Storms	Unintentional	Electromagnetic interference from solar flares and other solar activity “drowns out” the satellite signals in space.	Loss of signal, or range errors affecting the accuracy of the location or timing information.
Jamming	Intentional	Locally generated RF interference is used to “drown out” satellite signals.	Loss of signal (if the jammer is blocking out all satellite signals) or range errors affecting the accuracy of the location or timing information.
Spoofing	Intentional	Fake satellite signals are broadcast to the device to fool it into believing it is somewhere else, or at a different point in time.	False location and time readings, with potentially severe impacts on automated and autonomous devices and devices that rely on precise GNSS timing.
RF Interference	Unintentional	Noise from nearby RF transmitters (inside or outside the device) obscures the satellite signals.	Loss of signal (if the transmitter is blocking out all satellite signals) or range errors affecting the accuracy of the location reading (if the receiver is at the edge of the transmitter’s range).
Signal Reflection	Unintentional	Reflection due objects such as buildings.	GNSS signals can reflect off relatively due to distant objects, such as buildings, which would cause gross errors in position accuracy if the receiver falsely locks onto the reflected signal instead of the direct signal.
User Error	Unintentional	Users over-rely on the GNSS data they are presented with, ignoring evidence from other systems or what they can see.	Can lead to poor decision-making in a range of scenarios.

Jamming and spoofing attacks on GPS operations are on the rise (NovAtel 2012; Huber 2018). Military GPS systems must be robust and contain features that can withstand such attacks and remain operational despite the interference. This point is expanded upon in later chapters of the study. GPS signals are transmitted using low power thereby presenting a vulnerability to adversarial interference. The GPS signal strength reaching the surface of the Earth is about -130 dBm (Bonebrake and O'Neil 2014). As a result, the GPS technology is susceptible to disruptive interference methods. GPS jamming is described as the intentional or unintentional interference of the signal that prevents a legitimate signal from being received. GPS interference can also originate from unintentionally produced RF waveforms that essentially raise the effective noise floor in the receiver processing, resulting in the degradation or denial of a receiver’s ability to operate (National Cybersecurity & Communications Integration Center 2016). Figure 2 is a depiction of a jammer prosecuting a target. The radio signal emanating from the jammer needs to be at the same frequency as the user equipment in the vehicle. The impact of the jammer is that the vehicle is unable to determine its true position, as the jamming inhibits the user equipment’s ability to lock onto the signal being broadcast by the satellite.

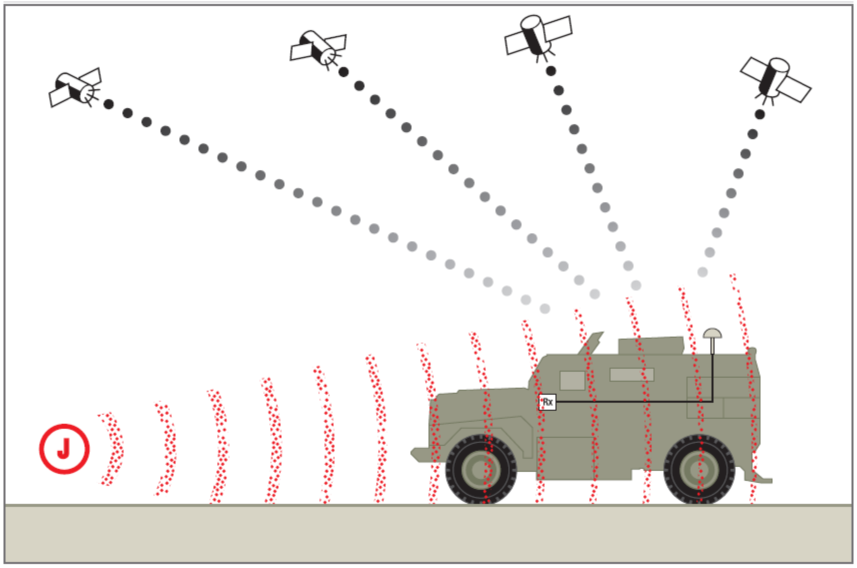


Figure 2. GPS Jammer. Source: NovAtel (2012).

GPS spoofing is caused by false radio frequency (RF) waveforms that are characterized as valid signals with the aim of interfering with a legitimate GPS signal. GPS spoofing is a malicious attack that can degrade, disrupt, or deceive the end user. In a spoofing scenario, false satellite signals are broadcast to the device to misrepresent its location and time. Spoofing "requires simulating the GPS accurately and capturing the user's receiver away from the true signal to steer it off course" (Cole 2015). The graphic in Figure 3 is an illustration of a spoofing attack. The actual position being reported, and the real position are shown to be some distance apart; subsequent chapters of this study expand on this action.

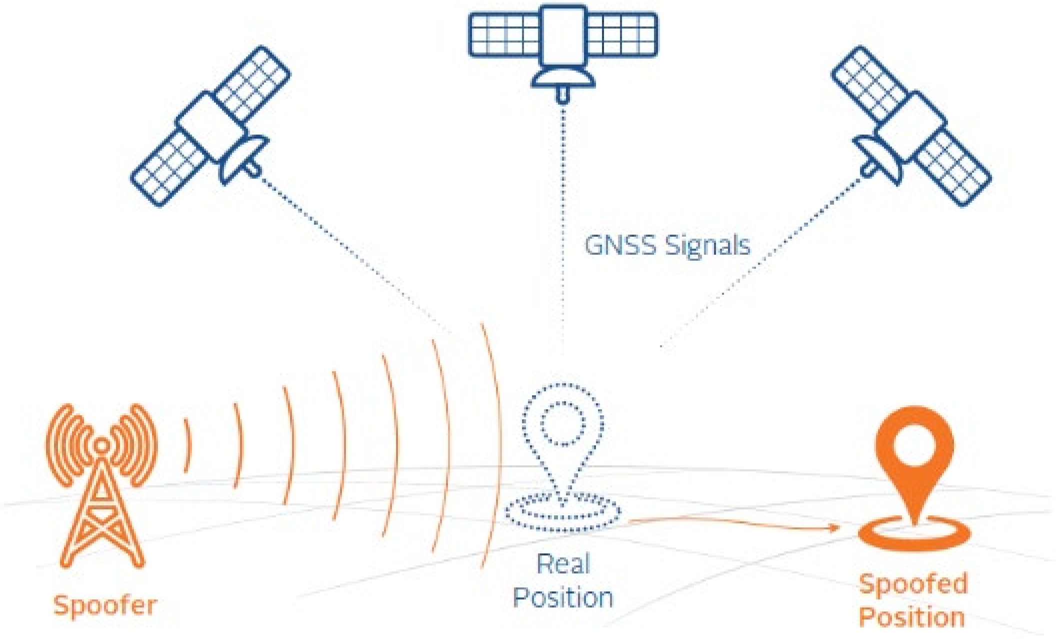


Figure 3. GPS Spoofing: Source: Hohman (2020).

A GPS spoofing attack is more difficult to detect and has the potential to be significantly more menacing than jamming because the user device is most likely oblivious to the threat (Jafarnia-Jahromi et al. 2012). Humphreys et al. amplifies this concern stating that "spoofing is more sinister than intentional jamming because the targeted receiver cannot detect a spoofing attack and so cannot warn users that its navigation solution is untrustworthy" (Humphreys et al. 2008, 1). These revelations about spoofing became motivators for undertaking this study as we believe that there are tools within our reach to address this vulnerability. Research shows that in the past, state-sponsored actors were the main players in the GNSS spoofing arena; however, spoofing is no longer in the domain of just nation actors. With the advances in software-defined radio (SDR) technology, GNSS spoofing can be executed effectively with cheap SDR using open-source software downloaded from the internet. Measurement spoofing and data spoofing are the two main classes of GPS spoofing (National Cybersecurity & Communications Integration Center 2016). Measurement spoofing injects enough interference upon the RF waveforms with the result that the target receiver produces incorrect measurements of critical measurement parameters such as time-of-arrival and frequency-of-arrival. Data spoofing, on the other hand, pushes false digital data to the target receiver. The receiver uses this false data in the processing of signals and the calculation of PNT (National Cybersecurity & Communications Integration Center 2016). Either of these spoofing activities results in a disruption of the GPS signal.

This thesis studied how artificial intelligence (AI) can improve the protection of the GPS against cyber-attacks by developing and analyzing a software system that incorporates machine learning (ML) and data analytics (DA) to identify GPS spoofing attempts. The developed ML algorithm identifies GPS spoofing attempts and alerts an end user of said spoofing attempt.

B. MOTIVATION FOR THESIS

The United States Armed Forces are the leaders in military space technologies and are dependent on satellite communications for the prosecution of armed conflicts. This technological advantage is not without risk. This risk has been made public, even reaching the halls of Congress. In a March 2017 "congressional hearing on space threats and the implications for homeland security, Rep. Donald Payne Jr. pointed to an example in his district of the dangers of GPS disruption" (Swarts 2017). He outlined the story of a truck driver using an illegal GPS jamming device in a company issued vehicle to deceive his employers on his location. This GPS jamming device presented interference and disrupted the satellite-based tracking system at Newark Liberty International Airport. The GPS tracking system at airports serves a crucial role in the safety of air traffic both in the air and on the ground. GPS tracking is a vital tool used by air traffic controllers allowing them to communicate with the on-board systems on an aircraft. “The potential disruption and harm that such an attack could do to critical infrastructure, in particular maritime and aviation systems, are particularly troubling” (Swarts 2017). The Federal Communications Commission (FCC) issued a Notice of Apparent Liability for Forfeiture (NAL) to the driver imposing a fine of $31,875 for the resulting disruption (FCC 2013). If an enemy learns to spoof false GPS data into our GPS reliant weapons, they could be turned against us, or our allies (Rooker 2008).

In an article in National Defense in 2010, the opening sentence was, “When it comes to battlefield intelligence, it’s far better to have too much than too little” (Magnuson 2010). Those words were attributed to then Undersecretary of Defense for Intelligence James R. Clapper. The article covers conversations with senior members of the U.S. DOD, the intelligence community, and top military personnel discussing sensors and data. Lt. Gen. David A. Deptula, Air Force deputy chief of staff for intelligence, surveillance and reconnaissance was quoted as saying “We’re going to find ourselves in the not too distant future swimming in sensors and drowning in data” (Magnuson 2010). The DOD incorporates sensors on every platform in their arsenal, from ground sensors to airborne systems with each node collecting data, thereby lending credence to his statement. On the local level, our command collects data while participating in test events, whether in our onsite labs or on field test events. This thesis leveraged the stream of test and operational data being collected through our test and field operations. This operational data served as the source used to implement the AI/ML system. This system produced from this study will serve as a redundant channel to the existing systems used to inform operators of any spoofing attempts on the end user equipment. Developing this asset using existing technologies enhances the decision-making process and removes some of the uncertainty with which an operator might be faced. AI is a tool to help facilitate data analytics by providing a means to analyze the data within a system and aid the operator’s decisions. The developed system presents notification to an operator when actions need to be taken based on the data received from the environment.

C. THESIS OBJECTIVE

Given the GPS vulnerabilities and many reported instances of GPS jamming and spoofing attempts, a means of detecting these events and alerting the end user is required. Evidence shows that users are not aware of jamming or spoofing events until the events have occurred. Knowing the danger that an anomalous activity poses, it is imperative that a system be put in place to enhance the current concept of operations (CONOPS). This thesis studied how AI can improve the protection of military GPS against cyber-attacks by developing and analyzing a software system that incorporates AI, ML, and DA tools to identify GPS spoofing attempts.

D. RESEARCH METHOD

This thesis applied a systems engineering analysis approach to conduct a problem space needs analysis, explore the use of AI and ML as a possible design solution, and to conduct a proof-of-concept to demonstrate the utility of the proposed solution to the original problem space. The problem space needs analysis consisted of researching military GPS systems, their data structures, and their system vulnerabilities. The problem space analysis included understanding the parameters that constitute the GPS data and identifying parameters that are susceptible to interference. The next step explored AI and ML as a design solution to address the problem space. A literature search identified model selection as one of the most significant and challenging issues in ML (Fujimaki et al. 2009). Toward this end, it was imperative that the area of ML was explored thoroughly to maximize the benefits of implementing the appropriate tools. The process of researching design solutions included the study of other programs that have successfully applied ML to anomaly detection, to adapt principles that have proven to be effective in achieving the objectives of this study. The final step in the thesis research was the demonstration of the application of ML to the GPS spoofing vulnerability problem. The study used live operational and simulated GPS data to train and test a variety of ML approaches (including anomaly detection, supervised learning, and unsupervised learning) to detect and classify GPS spoofing activities. This demonstration served as a proof-of-concept to explore the use of AI and ML methods for providing layers of security for GPS and for quantifying the levels of PNT assuredness.

E. EXPECTED BENEFIT OF THESIS RESEARCH

An end user of a military GPS system will not know immediately that a spoofing attack is occurring on his or her receiver. One consequence of a spoofing attack is the reporting of false position and time information. False data can have severe impacts on an operation, on automated and autonomous systems, and on systems that rely on precise GNSS timing. Using both operational and simulated data for training and testing, the vision is to create an intelligent system that can detect and react to GPS spoofing attempts while alerting the operators of said actions. This thesis increases the military's understanding of how AI can support GPS spoofing detection and reporting.

F. THESIS OVERVIEW

Chapter I provided an introduction, a description of the problem statement, an explanation of the motivation for this research, the thesis statement, and the research method for this study.

Chapter II provides a summary of the literature surveyed in developing the methods used in this thesis. Chapter III describes the conceptual GPS-spoofing–detection AI system and the prototype that was developed for experimentation. Chapter III contains a description of the prototype demonstration and ML methods used for experimentation. Chapter IV provides the overall results of the study and ML demonstration of the prototype GPS-spoofing–detection system. Chapter V concludes the thesis with a summary of the research and recommendations for future work.

II. LITERATURE REVIEW

This chapter provides a summary of the literature and sources that were used to increase the level of understanding on the study of GPS spoofing detection strategies. The literature review was focused in three areas: understanding GPS characteristics and capabilities, reviewing documented studies related to GPS spoofing, and exploring AI and ML methods being used in practice.

The subject of AI is an evolving science that has adapted to the changing state of the art in computing. With that in mind, this thesis focused on reviewing literature from the last 15 years to form the basis of this research. The literature search explored AI and ML methods currently available and being used in both commercial and military applications. The literature review identified AI methods that show potential as solutions to address the GPS spoofing problem. The literature review provided a knowledge framework to support the thesis research by informing the needs analysis, identifying methods for the development of a design solution, and providing a basis for the ML proof-of-concept analysis.

A. GPS DESCRIPTION SYSTEM AND SYSTEM VULNERABILITIES

GPS is a U.S.-owned satellite-based navigation system made up of 30+ navigation satellites circling the Earth and providing users with positioning, navigation, and timing (PNT) services. GPS has become a modern-day convenience used to support navigation between locations. The GPS system has three components: space systems, ground stations, and user equipment. The space systems, in the form of satellite constellations, consist of 24 to 32 satellites. The satellites are organized in a configuration occupying six orbital planes with each plane oriented at angle of inclination of 55- degrees. The satellites travel in medium earth orbit (MEO) and are positioned 12,532 miles (approximately 20,200 km) above the Earth. Each configuration is equipped with four satellites per plane as shown in Figure 4, where each satellite completes one orbit in one-half of a sidereal day. *Merriam-Webster* defines sidereal as "of, relating to, or expressed in relation to stars or constellations." Expanding on this definition, a sideral day is a measure of the rotation of the Earth relative to the stars and is approximately 23 hours 56 minutes and 4.1 seconds (Crockett 2012). The length of a sidereal day contrasts with a solar day, which is a measure of the 24 hours that it takes the earth to spin once on its axis while revolving around the sun. This configuration, combined with the global reach of GPS, makes navigation and timing available to users at any location on Earth. These users have persistent access to at least four satellites on demand.

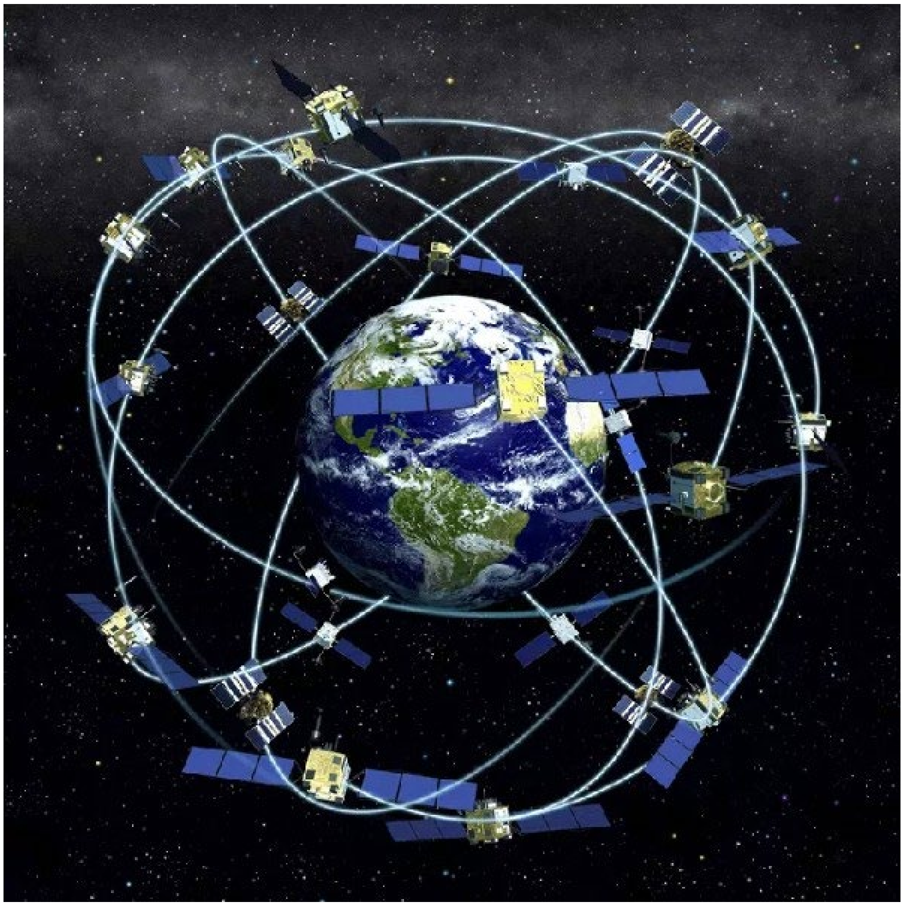


Figure 4. The Orbits of GPS Satellites. Source: Howell (2018).

Communication between a GPS receiver and the satellites is based on one-way transmission from the satellites to the receiver using direct sequence spread spectrum (DSSS) modulation. DSSS works by multiplying the original satellite signal with a pseudo random noise spreading code. This technique reduces signal interference and produces a continuous scrambled signal. The GPS system uses a one-way time-of-arrival (TOA) ranging technique in its operations (Kaplan and Hegarty 2017). TOA works on the principle that the distance between the satellite and the GPS receiver can be determined by using the propagation time and the speed of light. Transmission between the satellites and the user equipment depends on accurate synchronization of the clocks on both systems. The ranging codes and navigation data are broadcast using a code division multiple access (CDMA) technique on two frequencies in the L Band, L1 (1,575.42 MHz) and L2 (1,227.6 MHz) (Kaplan and Hegarty 2006). CDMA allows multiple users to receive service on the same carrier frequency. GPS's use of the L-band was crucial to the developmental concept of the system. GPS was designed to be able to operate in all types of environmental conditions and was required to operate using frequencies below 2 GHz (Ogaja 2011). The L-band occupies the 1–2 GHz frequency range with wavelengths between 15–30 cm. The L-band signals allow GPS to operate in many weather conditions such as snow, rain, and fog while also allowing GPS signals to penetrate some vegetation.

The GPS control segment (CS) consists of a global network of ground facilities as shown in Figure 5. The master control station is located at Schriever Air Force Base (AFB) in Colorado, with an alternate master control station at Vandenberg AFB in California (GPS.gov 2018a). The CS, unlike the user segment, provides two-way transmission between the ground station and the satellites allowing for the tracking and maintenance of the satellites in space. Other functions of the CS include monitoring the health and signal integrity of satellites and maintenance of their orbital configuration. According to Kaplan and Hegarty, other functions assigned to the CS include updating the ephemerides and other critical parameters used to determine user position, velocity, and time (PVT) (Kaplan and Hegarty 2017). *Merriam Webster* defines ephemeris, singular of ephemerides, as "a tabular statement of the assigned places of a celestial body for regular intervals" (Merriam-Webster. 2021). For GPS, ephemerides represent the position of the satellites relative to time and includes information on week number, satellite accuracy and health, age of data, and orbital parameters (Kaplan and Hegarty 2006).

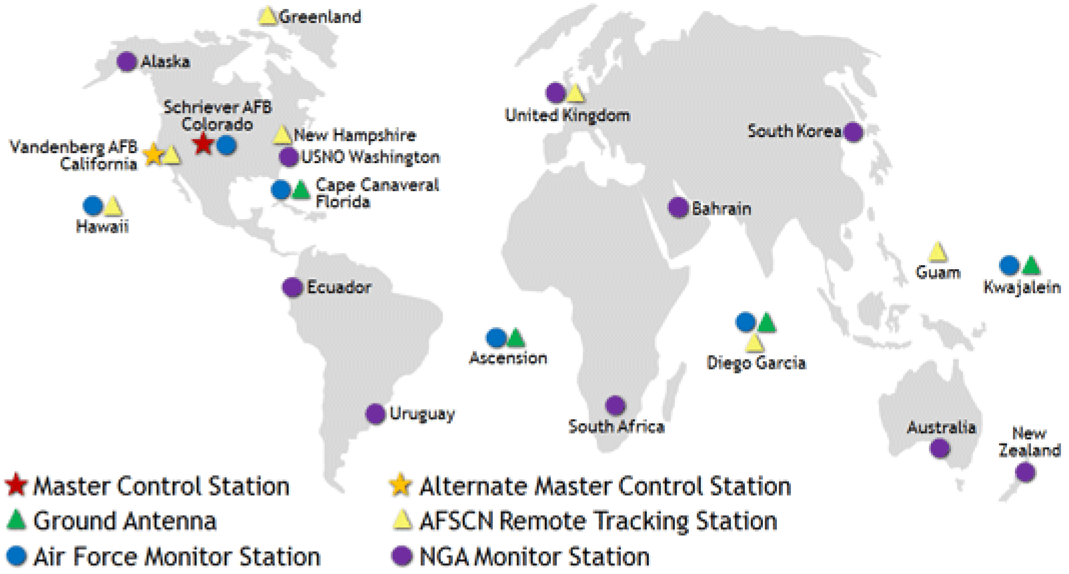


Figure 5. Control Segment. Source: GPS.gov (2018).

The user segment is represented by a GPS receiver. A GPS receiver unit takes many forms and can be as basic as wearables used for recreation and sport or as advanced as the equipment providing PNT capabilities to autonomous vehicle navigation. A GPS receiver contains an antenna for receiving signals, a receiver/processor unit for converting RF signals to a navigation solution, and a control/display unit, for displaying navigation to the user. The user receiver equipment performs navigation, timing, and other related GPS functions. An example of a GPS receiver used for military applications is the Defense Advanced GPS Receiver (DAGR). A DAGR is a military GPS receiver that can be used as a handheld device or as part of a vehicle-mounted system. A properly configured and fielded DAGR provides resilience to both jamming and spoofing thereby providing more security and reliability in the field. Figure 6 shows a DAGR being used in the handheld configuration.

Figure 6. Defense Advanced GPS Receiver.
Source: BAE Systems (2020).

GPS time is based on atomic clocks and is related to Coordinated Universal Time (UTC). The United States Naval Observatory (USNO) maintains the DOD reference for time and time interval (National Institute of Standards and Technology [NIST] 2019). To coordinate time between the satellites and receivers, “the time on each satellite is derived by steering the on-board atomic clocks to the time scale at the GPS Master Control Station, which is monitored and compared to UTC” (NIST 2019). The National Institute of Standards and Technology (NIST) explanation of GPS time further states that “since GPS time does not adjust for leap seconds, it is ahead of UTC by the integer number of leap seconds that have occurred since January 6, 1980, plus or minus a small number of nanoseconds” (NIST 2019). The time offset from UTC is understood to be contained in the GPS broadcast message that usually is applied automatically by GPS receivers (NIST 2019). Tsui, (2000, 9) states that “In GPS, the position of the satellite is known from the ephemeris data transmitted by the satellite.” It is documented that the U.S.-operated GPS is the world’s first global navigation satellite systems (GNSS). GPS was developed as a military asset but now provides service to both civilian and military users. While the civilian service is freely available to all users on a continuous, worldwide basis, availability of the military service is restricted to the U.S. services, allied armed forces, and approved Government agencies. GPS provides two levels of service: the standard positioning service (SPS) and the precise positioning service (PPS). SPS "was originally designed to provide civil users with a less accurate positioning capability than PPS, through a feature known as Selective Availability (SA)" (Office of the Department of Defense 2020). PPS is reserved for U.S. authorized military and select government agency users and provides a higher level of accuracy than SPS. Because PPS is used for military applications, access is controlled through cryptography. Anti-spoofing (AS) and SA are two measures used for PPS. AS is used to defeat deception jamming. Kaplan and Hegarty (2006, 4) describes deception jamming as "a technique that can be used to deceive an unsuspecting receiver by replicating the satellite ranging codes, navigation data signal(s), and carrier frequency Doppler effects."

Each satellite transmits on the L1 and L2 frequencies with ranging codes which differ by satellites. Ranging codes are also used by the system in computing PVT solution. Selection of the ranging codes is based on the criteria of having low cross-correlation properties relative to other satellites. GPS system is structured such that "each satellite generates a short code, the coarse/acquisition or C/A code and a long code, the precision or P(Y) code" (Kaplan and Hegarty 2006, 3). These two codes essentially distinguish the civilian and military use of GPS. Secure GPS and assured PNT refers to military-encrypted GPS that uses military P(Y) code and M-Code (Cole 2015). The encrypted P(Y) signal provides accuracy to within centimeters of the target, while the C/A signal produces an accuracy of about 5 meters (GPS.gov 2020). Y-code, the encrypted P-code used in military applications, is a pseudo random code that operates at ten times the frequency of that used in civilian applications. M-code is used in the L1 and L2 GPS bands, encrypts receiver signals, and provides the capability to detect and reject false GPS signals (Barker, et al. 2005). While Y-code and M-code adds robustness to the system while offering higher levels of cyber resiliency to military systems; they do not eliminate all risks to assured PNT. Figure 7 presents a graphic illustration of both the GPS P(Y) Code and the GPS M-Code on the L1 frequency band along with the other worldwide GNSS on said L1 frequency. The figure shows that the M-Code occupies a wider portion of the L1 spectrum than the P(Y) code furthering the robustness and resiliency of the system to interference.

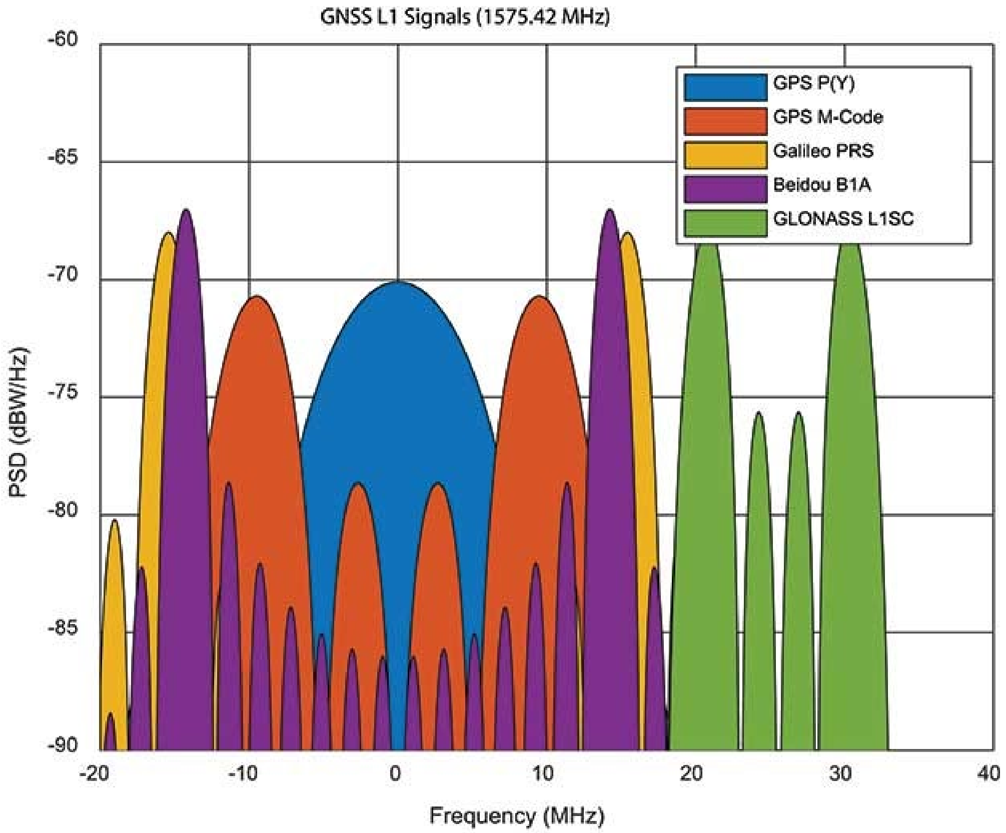
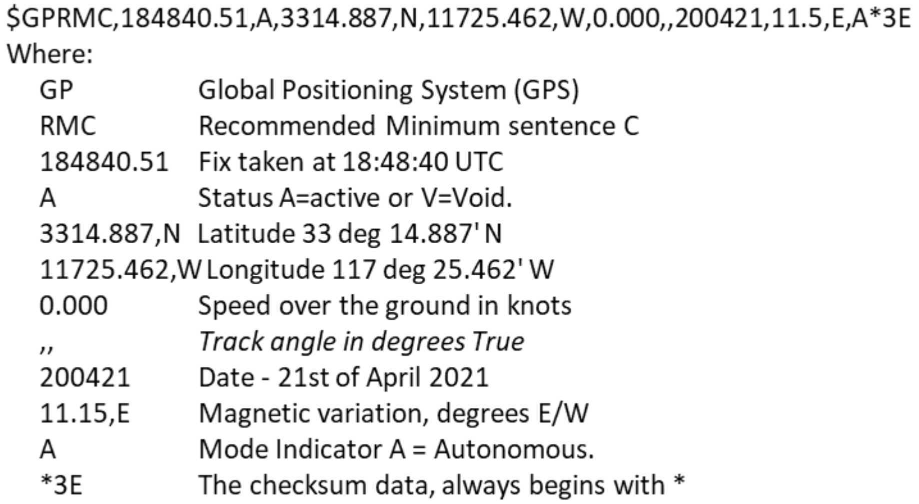


Figure 7. GNSS Encrypted Signals. Source: Jones (2019).

GPS jamming and spoofing are two methods used to interfere with GPS signals. GPS spoofing is an attack in which valid waveforms with invalid data are transmitted to trigger a GPS receiver to produce valid messages with invalid PNT information. GPS spoofing aims to interfere with a legitimate signal by broadcasting false satellite information to the user device to misrepresent its location and time. Using a network analogy, a GPS receiver looks like an open port to an adversary. The receiver is always in a state of readiness, listening for, and receiving GPS signals, making it vulnerable to spoofing attacks. In the past, spoofing attacks were carried out by state actors, but with changes in technology and the low-cost of the required electronic components, non-state actors can easily acquire the equipment needed to interfere with GPS signals. GNSS spoofing can be executed with commercially available hardware and open-source software, easily putting this capability in the hands of the general population (Korolov 2019). Cuntz et al. (2012) reiterate that GPS spoofing requires a more concerted effort than jamming, while the risk and consequences of not detecting such attacks are extremely high.

The GPS signal presented at the antenna port of a user device contains information relating to PNT along with other information broadcast by the satellite. GPS signals to the user equipment travels on a one-way path from the satellite to the user equipment; there is no information exchange from the user equipment that is transmitted to the satellites. There are a specific set of messages and corresponding data fields that are used in the DAGR in providing a valid solution to the user. Within these messages there are data items that when interfered with, impacts the GPS solution presented to the end user. The National Marine Electronics Association (NMEA) 0183 Interface Standard "defines the electrical signal requirements, data transmission protocol and time, to support the one-way serial data transmission used for GPS" (National Marine Electronics Association 2012, xx). Typically, NMEA 0183 standard messages range from 11 to a maximum of 79 characters in length. Messages are generally transmitted at a maximum rate of one per second, according to the standard. All compliant GPS devices conforms with the NMEA 0183 standard data format supported by GPS manufacturers. Messages that are impacted by GPS interference include the global positioning system fix data (GGA), the GNSS satellites in view (GSV), GNSS dilution of precision (DOP) and active satellites (GSA), and the time of day (ZLZ). The GGA message represents the time, position and fix related data for a GPS receiver. The GSV message presents the number of satellites (SV) in view, satellite ID numbers, elevation, azimuth, and SNR value. The NMEA 0183 standard dictate that the GSV sentence includes a maximum of four satellites per transmission. The GSA message contains "GNSS receiver operating mode information, the satellites used in the navigation solution reported by the GGA sentence, and the DOP values" (NMEA 2012, 94). The standard requires that the ZLZ message contains the time of day in hours-minutes-seconds, both with respect to (UTC) and the local time zone. The NMEA 0183 standard describes the Recommended Minimum Specific GNSS Data (RMC) message to include the "time, date, position, course, and speed data provided by a GNSS navigation receiver" (NMEA 2012, 113) as shown in Figure 8. The standard dictates that an RMC message "is always accompanied by RMB when a destination waypoint is active" (NMEA 2012, 113). The NMEA 0183 standard dictates that messages start with the *$* character and that each data field is separated by a comma as illustrated in the following data sentence structure.



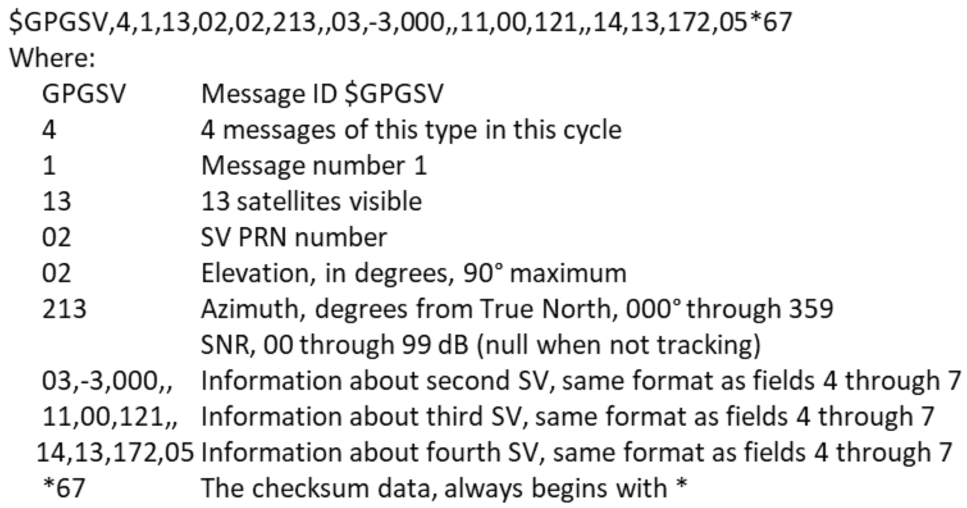
$GPRMC,184840.51,A,3314.887,N,11725.462,W,0.000,,200421,11.5,E,A*3E

Where:

GP	Global Positioning System (GPS)
RMC	Recommended Minimum sentence C
184840.51	Fix taken at 18:48:40 UTC
A	Status A=active or V=Void.
3314.887,N	Latitude 33 deg 14.887' N
11725.462,W	Longitude 117 deg 25.462' W
0.000	Speed over the ground in knots
,,	*Track angle in degrees True*
200421	Date - 21st of April 2021
11.15,E	Magnetic variation, degrees E/W
A	Mode Indicator A = Autonomous.
*3E	The checksum data, always begins with *

Figure 8. GPS Message Structure: Source: NMEA (2012).

RMC and RMB are the recommended minimum data to be provided by a GNSS receiver as stated in the NMEA 0183 standard. Further, all data fields must be accounted for in the message. In the event of a null field, there will be no data between the commas representing said entity. The NMEA 0183 standard describes the RMC sentence as a periodic message containing position, velocity, and time parameters that is transmitted at no more than two second intervals. The GSA GPS DOP and GSV message in Figure 9 uses a similar structure as the RMC message. The GSV message string contains information on the number of satellites in view, the PRN, elevation, azimuth, and SNR value for each satellite in view (NMEA 2012).



$GPGSV,4,1,13,02,02,213,,03,-3,000,,11,00,121,,14,13,172,05*67

Where:

GPGSV	Message ID $GPGSV
4	4 messages of this type in this cycle
1	Message number 1
13	13 satellites visible
02	SV PRN number
02	Elevation, in degrees, 90° maximum
213	Azimuth, degrees from True North, 000° through 359
	SNR, 00 through 99 dB (null when not tracking)
03,-3,000,,	Information about second SV, same format as fields 4 through 7
11,00,121,,	Information about third SV, same format as fields 4 through 7
14,13,172,05	Information about fourth SV, same format as fields 4 through 7
*67	The checksum data, always begins with *

Figure 9. NMEA 0183 GSV Message: Source: NMEA (2012)

The GNSS DOP and GSA message contains information such as the receiver's operating mode, the satellites used in the navigation solution, and DOP values. Fields 3 through 14 of the GSA represents the ID of the satellites in view while fields 15–17 represent the various DOP elements. The information in the GSA represents the active satellites. Field 15 represents PDOP (position dilution of precision) and describes the error produced by the relative position of the GPS satellites. Field 16 represents horizontal dilution of precision (HDOP) in meters, while field 17 represents the vertical dilution of precision (VDOP) in meters. The DOP parameters are factors in the source of GPS error. The metrics that capture the error caused by the relative position of the GPS satellites are geometric dilution of precision (GDOP) or position dilution of precision (PDOP) (GIS Geography 2020). Even though we are using a DAGR in a static land-based construct in this experiment, both the horizontal and vertical accuracy and error factors are being factored in the analysis. Error and accuracy will be impacted by spoofing attempts. Additionally, geometry, atmospheric conditions, and nearby objects can impact the quality of a valid GPS solution.

Military GPS receivers such as the DAGR are designed for compliance with NMEA 0183 standard and therefore implement the messages discussed. In addition to the NMEA 0183 standard, the DAGR implements Interface Specification IS-GPS-153 Interface

Control Document (ICD). IS-GPS-153 messages are also referred to as ICD-GPS-153 messages. The web portal GPS.gov describes ICDs as "the formal means of establishing, defining, and controlling interfaces and for documenting detailed interface design information for the GPS program" (GPS.gov 2021a). The interface protocol and data message formats for the DOD GPS equipment are documented in the ICD-GPS-153. The DAGR used in this study provides an ICD-GPS-153 compliant port and will be discussed in subsequent chapters as part of the experiment set up and data collection scheme.

There are two classes of GPS spoofing: measurement spoofing and data spoofing. The classes can be further expanded to capture the spoofing generation methods. According to Jafarnia-Jahromi et al. (2012) spoofing generation can be divided into three documented categories: GPS signal simulator, receiver-based spoofers, and sophisticated receiver-based spoofers. In the GPS signal simulator category, the output of the GPS signal simulator is mixed with an RF signal to be presented as authentic GPS signals as illustrated in Figure 10. The scenario captured in the figure shows a missile launched from a ship in coastal waters. The spoofer can lock on to the GPS coordinates from which the operator increases the signal strength to overpower the legitimate signals resulting in a change of target location for the missile. The second category, a receiver-based spoofer, is a more advanced type of spoofer than a GPS simulator spoofer. The receiver-based spoofers operate by concatenating a GPS receiver with a spoofing transmitter. For a receiver-based spoofer to be effective, a line-of-sight position to the system needs to be established, thereby making this a more difficult method to execute. The sophisticated receiver-based spoofers are the most complex and the most effective of the three spoofing categories described in the literature (Jafarnia-Jahromi et al. 2012). These broad categories can be defined as synchronous and asynchronous attacks.

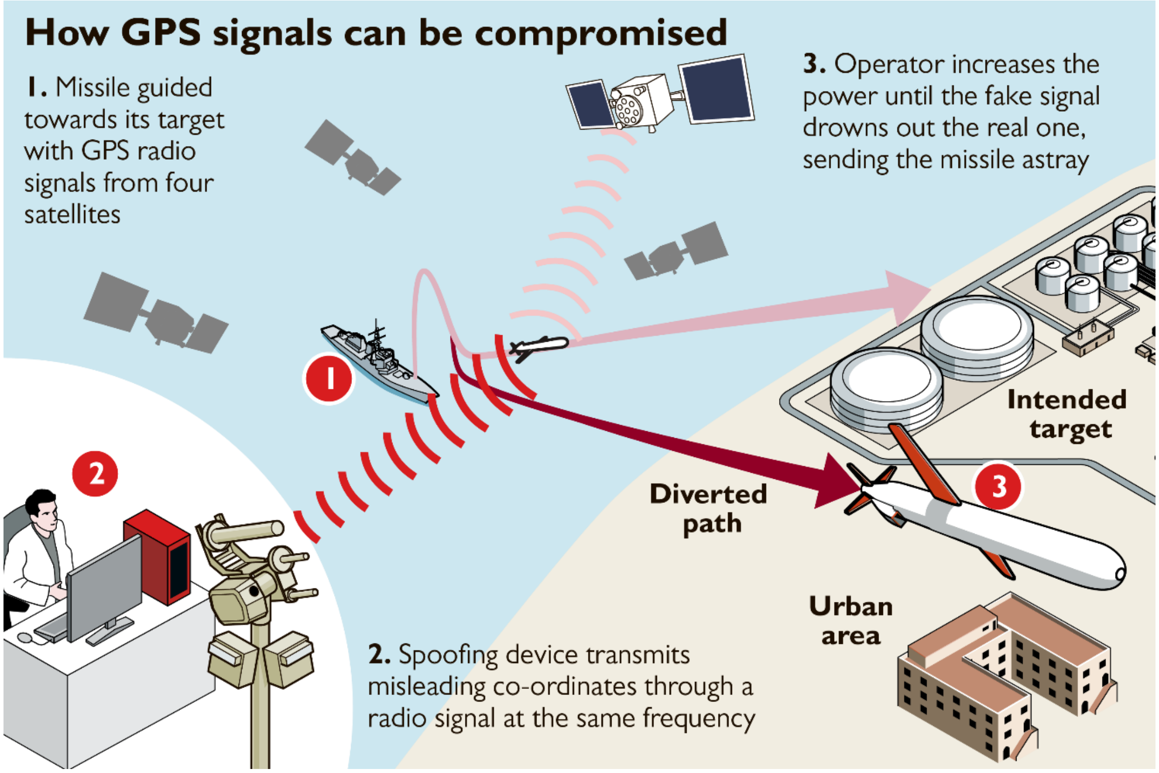

How GPS signals can be compromised
1. Missile guided towards its target with GPS radio signals from four satellites
3. Operator increases the power until the fake signal drowns out the real one, sending the missile astray
1
2
3
Intended target
Diverted path
Urban area
2. Spoofing device transmits misleading co-ordinates through a radio signal at the same frequency

Figure 10. GPS Spoofing: Source: Moody (2018).

This study used asynchronous attack methods to present a spoofing attack on the DAGR. An asynchronous attack attempts to overpower the legitimate GPS signal by increasing the power of the spoofing signal.

B. DOCUMENTED STUDIES RELATED TO GPS SPOOFING

Both jamming and spoofing attacks on GPS are in the realm of electronic warfare (EW) in the category of cyber-attacks. Jamming of GPS results in denial-of-service attack (DoS) inhibiting a user from accessing the required PNT information. According to the Cybersecurity and Infrastructure Security Agency (CISA), “A DoS attack occurs when legitimate users are unable to access information systems, devices, or other network resources due to the actions of a malicious cyber threat actor.” “A GPS jamming cyber-attack blocks GPS signals from getting to a receiver, whereas GPS spoofing allows hackers to interfere with navigation systems without operators realizing it” (McAfee 2020). GPS spoofing attacks are more difficult to detect and require a higher level of technical expertise to execute. As GPS spoofing impacts the GPS PNT solution, potentially placing a user in a location that is different from their planned position, the threat to safety is amplified.

Much of the work on GPS vulnerabilities can be traced back to the President's Commission on Critical Infrastructure that was established in July 1996. This Commission under the auspices of the DoT, ordered "an assessment of the vulnerability of the transportation infrastructure relying on the use of GPS" (Volpe 2001). The focus of the report was the perceived vulnerability of the U.S. transportation infrastructure; however, one recommendation was that the DOD "vigorously support and protect the spectrum for GPS and its applications" (Volpe 2001). The Volpe report identified jamming and spoofing as potential attack methods. The report also identified the exploitation potential of GPS, "as GPS further penetrates into the civil infrastructure, it becomes a tempting target that could be exploited by individuals, groups or countries hostile to the United States" (Volpe 2001). Twenty years later, this finding resonates with the same impact as when the report was first released. Several studies about GPS jamming have been documented on the subject. GPS spoofing studies have not been reported at the same level as GPS jamming. Spoofing attacks are not necessarily fewer than jamming, but as stated previously, spoofing is more sinister and much harder to detect than a jamming episode.

In the wake of the incident and outrage surrounding the capture of the U.S. drone by Iran in 2011, the U.S. Department of Homeland Security took up the charge to advance the study of spoofing of U.S. military assets. Behind the scenes, the speculation surrounding this capture was the cause of much consternation and led to congressional investigations. Todd Humphreys, an assistant professor at the University of Texas at Austin (UT), is one of the many researchers who has been advancing the work of identifying GPS spoofing activities. In June 2012, Professor Humphreys and his research team from UT were invited to participate in an exercise at the White Sands Missile Range in New Mexico to assess whether they could successfully commandeer military systems by injecting false GPS data. One objective was to provide incorrect data to an unmanned aerial vehicle (UAV) in flight to force the aircraft to land (Psiaki, Humphreys, and Stauffer 2016; Farivar 2013). The team successfully injected their false data into the system causing the UAV to adjust its altitude as it was led to believe it was executing a climb. It took a manual override of the controls to prevent the aircraft from crashing according to the report (Psiaki, Humphreys, and Stauffer 2016).

The details of the White Sands exercise were presented at the 2013 South by Southwest Interactive conference and generated interest in both the public and private sectors. One attendee who took a keen interest in the presentation was Andrew Schofield, the "Master of the *White Rose of Drachs*." The *White Rose of Drachs* is an $80 million 213-foot British owned super yacht that was offered up to Professor Humphreys to conduct a spoofing exercise. That same year, Professor Humphreys once again led his research team from UT in conducting this experiment in the Mediterranean Sea. Armed with a customized GPS spoofer, the team set out to demonstrate that they could steer the White Rose off its designated course. To execute a successful spoofing attack, the spoofer had to mimic the true GPS signal from the satellite constellation to user equipment, which would then accept the false signal as being the true PNT solution. A spoofer, to be effective, must have line of sight with the target, knowledge of the satellites in view of the target, and the knowledge of the publicly available PRN code for the satellites. The spoofer must be able to transmit the PRN codes to the receiver in a series of steps to coerce the receiver into accepting the illegitimate signals. In the end, the team successfully forced the $80 million vessel to unknowingly stray off the planned course by spoofing the GPS signal using what was described in the literature as "the world's first openly acknowledged GPS spoofing device" (UT News 2013). According to the UT News, the team wanted to assess whether the ship's sensors could detect that spoofing was occurring. The team wanted to determine if carrying out such a spoofing attack would be possible. On the surface, one would believe that with a high value target such as the *White Rose of Drachs*, the on-board electronics would be robust and resilient to such interference. Nonetheless, the team was successful in carrying out this spoofing attack. The vessel tracked off the intended course by a kilometer without the captain's knowledge, according to the reports (Farivar 2013; Psiaki, Humphreys, and Stauffer 2016).

Spoofing attacks on a military system are, in theory, a more difficult proposition than the spoofing of the system on the *White Rose of Drachs*, due to the safeguards used by military GPS operations. Military GPS operates on the same L1 and L2 signals as commercial systems. However, the keying or encryption of military systems using P(Y) Code and M-Code makes these systems more resilient to interference. The concept of operations (CONOPs) for fielding USMC GPS equipment dictate that said equipment uses communications security (COMSEC) encryption keys. NIST defines COMSEC as “a component of information assurance (IA) that deals with measures and controls taken to deny unauthorized persons information derived from telecommunications and to ensure the authenticity of such telecommunications” (NIST 2021). Documentation on the DAGR also decree that GPS equipment are loaded with valid COMSEC keys prior to operation in the field (logsa.army.mil n.d.). An unencrypted DAGR is as susceptible to a spoofing attack as a commercial GPS user device. The reports of the Navy ship in Iranian waters and the UAV captured by Iran are two of the documented incidents of spoofing U.S. military assets in the public domain. Even though most of the reports of GPS spoofing have been on commercial assets, the gravity of the consequences of a spoofed system should not be discounted.

The U.S. Maritime Administration maintains a web portal that tracks reports of GPS incidents across the world and has documented multiple recent spoofing attempts. One report entitled “Did Russia make this ship disappear?” documents the experience of the captain of the Atria oil tanker, who recognized an anomaly in his navigation showing his ship being at an airport inland when, in actuality, the vessel was heading to a port 20 nautical miles away (Darwish 2017). After reporting this incident to the U.S. Coast Guard, an advisory was issued showing that about 20 other vessels in the vicinity experienced similar issues wherein their navigation systems placed them in locations that were different from the true position (McLaughlin 2020). The evidence from these episodes pointed to Russia as the agitator. As mentioned earlier, GPS spoofing is no longer limited to nation states, but also to individuals who, with the right mix of skills and access, can execute these deleterious acts. Nation states and individuals with an agenda can wreak havoc in the transportation sector by using vehicular system spoofing to turn vehicles into deadly weapons. The possibilities are endless.

C. ARTIFICIAL INTELLIGENCE METHODS

The work on AI started in earnest shortly after World War II. The term AI, by all accounts, was coined in 1956. Artificial intelligence is no longer a new area of science as evidenced by the number and diversity of organizations using AI as a tool. Russell and Norvig write that "*homo sapiens*, for thousands of years, have tried to understand how we are able to perceive, understand, predict, and manipulate a world far larger and more complicated than itself" (Russell and Norvig 2003, 1). The authors further state that AI takes the human attempts further and attempts not just to understand, but to also build intelligent entities. The field of AI has experienced significant growth. As stated in the literature, AI systemizes and automates intellectual tasks and can be applied to almost any sphere of human intellectual activity (Russell and Norvig 2003). AI also refers to computer systems trained to simulate human intelligence and can perceive the environment, make decisions, and take actions.

Artificial intelligence encompasses ML, DA, and deep learning principles. The literature search underscores the concept that AI is one of the emerging technologies most used by large enterprises through ML and predictive analytics (Abramovich 2021). ML covers the concept of machines being able to learn from given data without human assistance. Deep learning is a subset of ML that is modeled after the neurons in the human brain. Deep learning uses "a hierarchical level of artificial neural networks (ANN) to carry out the process of machine learning" (Hargrave and Anderson 2021). Deep learning algorithms learn what features from a dataset are important. Deep learning functions on both structured and unstructured data. AI applications find usefulness in wide-ranging industries and market segments. An Internet search of AI applications generates any number of lists. For example, a 2020 blog published a list of top 10 real world AI applications in which marketing, banking, finance, agriculture, healthcare, space exploration, and autonomous vehicles are listed (Johari 2020). Recommender systems are used in the field of marketing to recommend items to customers based on demographic data and previous purchases. AI systems are being used in healthcare, to help practitioners more quickly and accurately detect and diagnose patients' maladies. For autonomous vehicles, AI is used for vision systems and natural language processing. This list shows the diversity of areas that are occupying the AI/ML technology space. With all the discussion presented about AI/ML, it should be reiterated that ML is considered a subset of AI. All ML applications are AI applications; however, the converse does not hold true.

In the DOD, AI and ML are becoming widely adopted tools. According to Grady, AI and ML tools are "the new arms race among the great powers" (Grady 2020). In 2017 the DOD launched Project Maven; an AI project conceived to take advantage of the abundance of image data captured by UAVs in theater. Shultz and Clarke (2020) in a publication on the Modern War Institute at West Point discuss the genesis of Project Maven. In the article, the authors state that one of the project's objectives is "to automate the processing, exploitation, and dissemination of massive amounts of full-motion video collected by intelligence, surveillance, and reconnaissance (ISR) assets in operational areas around the globe" (Shultz and Clarke 2020). Project Maven is not without its share of dissenters. In 2018, it was reported that more than 3,000 employees at Google voiced their opposition to the company's involvement in the project. Their reasoning was that by being involved with the project, Google's brand and image would suffer in the eyes of the public. Meanwhile, it has been reported that researchers at the Army Research Laboratory (ARL) are working on a recommender system for use by the Army (Knapp 2018). Applying the underlying AI/ML principles of recommender systems used by commercial enterprises such as Netflix, Amazon, and a host of others, this system is being considered as a force multiplier in helping soldiers more quickly assess a situation and respond in kind. A RAND Corporation report published in 2020 brings to the fore the ethical concerns of using AI/ML tools in prosecuting this new arms race. There are public concerns "about the legal and ethical implications of using AI in war or even to enhance security in peacetime" (Morgan, et al. 2020). Other documented findings show that the likelihood integration and employment of AI in military systems is inevitable, the risks of deploying AI/ML systems in warfare need to be addressed, and that there are apprehensions about the operational risks related to the reliability, fragility, and security of these systems (Morgan et al. 2020). The authors recommend that there should be a public outreach campaign "to inform stakeholders of the U.S. military's commitment to mitigating ethical risks associated with AI to avoid a public backlash" (Morgan et al. 2020).

With the abundance of data being captured and stored in almost every human and machine activity, it seems logical to find methods to harness the power of the available data using AI tools. The findings of the literature review demonstrate that application of AI methods in the detection of spoofing activities on GPS is one area that has gained traction and is one of the areas being studied in earnest. Our literature search has found several commercial products that are being marketed as AI-based solutions for addressing the GPS spoofing attack issue. A 2019 publication from Mitsubishi Electric Research Laboratories documents a "wide-area algorithm" used in protecting GPS from spoofing attacks (Bhamidipati et al. 2019). The paper focused on signal-level spoofing using the methods of belief propagation (BP) and recurrent neural network (RNN). BP is used to estimate the states of unobserved variables in a system or perform inferences on models using Bayesian networks. A RNN is a deep learning tool that remembers important characteristics of the input stimuli thereby producing more accurate predictions. RNN lends itself well to working on sequential data such as time series, audio, and weather data (Donges 2019).

A 2017 article in *The Journal of Navigation* describes using ML for detection of spoofing attacks by employing *k*-nearest neighbor (KNN) and naïve Bayesian classifier techniques. The paper illustrates a spoofing detection method implemented in an SDR. The algorithm uses neural networks (NN) to identify any unusual distortions of correlation to detect spoofing attempts (Mosavi and Moazedi 2017). The model in this solution is trained to apply a NN to identify when the signal index moves beyond the allowed threshold. A spoofing attempt is identified by a model that detects the nefarious action of an attacker that is attempting to "occupy the receiver's correlation peak" (Mosavi and Moazedi 2017). The solution presented in this paper is shown to be a software add-on and requires no additional hardware, thereby not impacting the size, weight, and power (SWaP) of the current system.

The literature search has shown that conducting scientific research can be accomplished in many ways. Dietterich state that "science-in-the-small" occupies one end of the spectrum. Science-in-the-small captures the work of individual scientists where they "formulate hypotheses, perform experiments, gather data, and analyze that data to test and refine their hypotheses.” The article suggests that this approach “provides profound scientific understanding.” One of the weaknesses of science-in-the-small is that it yields progress slowly because each scientist has to overcome the limits of time and space in conducting their study (Dietterich 2009). Dietterich calls the other end of the spectrum “science-in-the-large.” Science-in-the-large works on a much larger scope than science-in-the-small. Because of the larger scope of this science-in-the-large, automated tools and instruments are used to collect the data. Analysis of the data along with formulation and refining of the hypotheses are conducted using ML algorithms and data mining tools (Dietterich 2009). The approach to this study used both principles. Science-in-the-small was used to advance the hypothesis of the study, while “science-in-the-large” was used to capture data and conduct the data analysis portion of the study. According to the literature, ML methods often have been employed for pattern matching and discovery. ML can be viewed as a collection of tools and methods that can be applied to disciplines such as statistics, banking, agriculture, control theory, transportation, and AI. According to Ao, “the spectrum of AI tools includes inductive logic programming, genetic algorithms, neural network, Bayesian networks, and hidden Markov Models, etc.” (Ao 2010).

D. SUMMARY

This chapter presented the results of a literature review in three areas: a description of GPS and its vulnerabilities, documented cases of GPS spoofing, and an exploration of AI/ML methods including current commercial and defense applications. From the literature survey, we can conclude that GPS, as a system, is vulnerable to both jamming and spoofing. Jamming impedes the passage of the low power signal from reaching the GPS user device, while spoofing results of misinformation being ingested by the user device. Our experiments have demonstrated that GPS systems are vulnerable to spoofing and from these demonstrations, we can extrapolate that nefarious spoofing can lead to catastrophic consequences. While AI techniques, and ML specifically, offers potential solutions to the GPS spoofing problem, the work on identifying these cyber-attacks and informing an end user needs to be pushed to the forefront as a matter of urgency.

THIS PAGE INTENTIONALLY LEFT BLANK

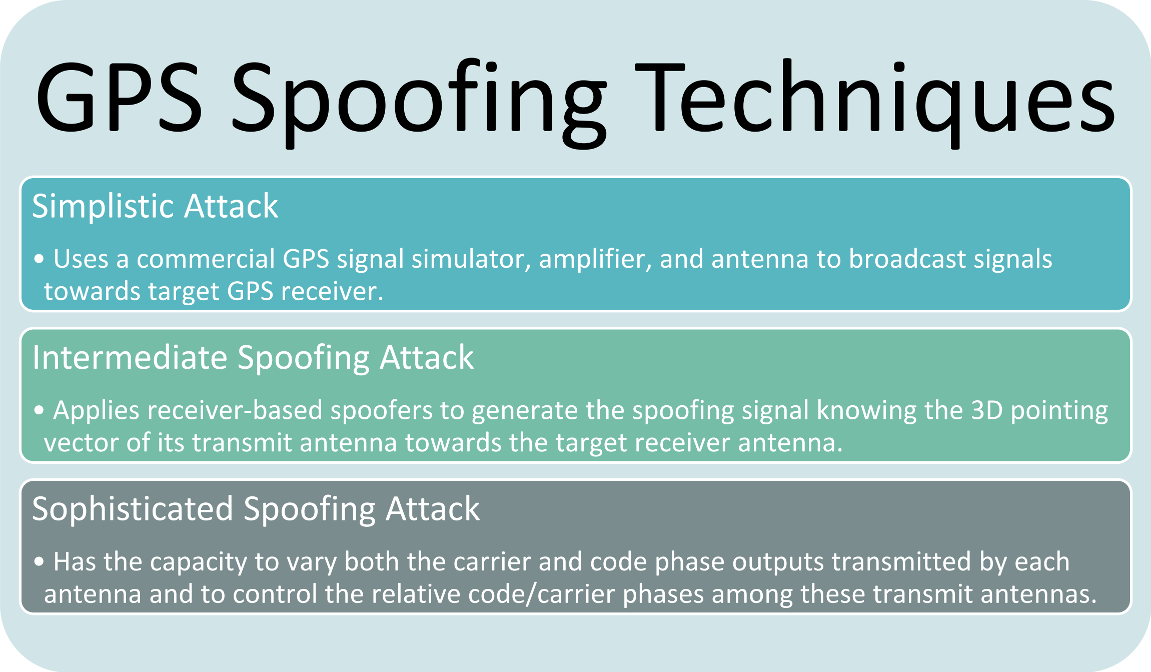
III. GPS-SPOOFING–DETECTION AI SYSTEM

This chapter provides descriptions of a high-level concept and a prototype experiment for a GPS-spoofing–detection AI system that is intended to detect GPS spoofing attempts. The chapter begins with a GPS spoofing taxonomy that was developed to characterize different types of GPS spoofing attempts. Next, the chapter presents a proposed concept for the envisioned AI system and how it might interact with the existing GPS system to alert operators of nefarious spoofing attempts. Finally, the chapter describes the prototype GPS spoofing detection AI system that was developed for this thesis to support experimentation and demonstration of the concept.

A. TAXONOMY OF GPS SPOOFING

Cuntz et al. (2012) puts forward the theory that a successful GPS spoofing attack impacts time, frequency, and space domains. An effective spoofing detection system must therefore consider, at least, all three of these domains in the detection of an adversarial GPS attack. The structure of both NMEA 0183 and ICD-153 messages contain other data items that can also be impacted by false data injection (FDI), so it is imperative that we take a holistic view of the total message structure in our detection work. In the power industry, there is growing concern about cyber-attacks on the smart grid (SG); the term FDI attack is being used to represent one mode of attack (Oozer and Haykin 2019; Lakshminarayana et al. 2020). The challenge in the SG operations, as in the GPS interference arena, is making the supporting infrastructure resilient to such attacks. In the power industry, the supporting infrastructure is "the part equipped with intelligence, dealing mainly with the control and monitoring aspect of the fundamental operations of the SG using software, hardware and communication networks" Oozer and Haykin (2019, 48). We can see parallels between FDI attacks on the SG and interference on GPS systems, leading us to adopt the FDI acronym to represent the spoofing attack problem being studied. Although GPS spoofing attacks can be executed by different methods, the consequence of these attacks, if successful, results in deceiving users and placing them in locations that are different from their preplanned route or destination. (Cuntz et al. 2012) present spoofing attacks in comparison to jamming events and concur with the statements put forward in other places that spoofing attacks require a more concerted effort than a GPS jamming episode. The impact, consequences, and the risk of not detecting the spoofing attack can be extremely detrimental to an operation (Cuntz et al. 2012). It is generally accepted that GPS spoofing techniques fall into three categories. The concept of three categories was presented in Humphreys et al. (2008). The authors identified simplistic, intermediate, and sophisticated techniques as known GPS spoofing methods. Figure 11 shows the taxonomy of GPS spoofing with the three categories of spoofing. The discussion of each of these methods follows in subsequent paragraphs.

It is widely held that most connected devices in use today are inherently vulnerable to malicious attacks. GPS devices are no exception to this interference whether by jamming or spoofing. To lend more credence to this notion, (Humphreys et al. 2008) assert that "all stand-alone commercial civilian GPS receivers available today are trivial to spoof" (3). Their rationale stems from the fact that all that is required to execute an attack in this mode are the motivation, a rudimentary set of skills, basic tools, and some level of effort. A GPS simulator, a power amplifier, and an antenna that radiates RF signal towards the user device, put in the hands of an adversary presents a clear and present danger to an operation that relies on an accurate GPS solution. This construct describes the simplistic spoofing attack illustrated in Figure 11 and presented by (Humphreys et al. 2008).



GPS Spoofing Techniques

Simplistic Attack

- Uses a commercial GPS signal simulator, amplifier, and antenna to broadcast signals towards target GPS receiver.

Intermediate Spoofing Attack

- Applies receiver-based spoofers to generate the spoofing signal knowing the 3D pointing vector of its transmit antenna towards the target receiver antenna.

Sophisticated Spoofing Attack

- Has the capacity to vary both the carrier and code phase outputs transmitted by each antenna and to control the relative code/carrier phases among these transmit antennas.

Figure 11. Taxonomy of GPS Spoofing

The simplistic attack aligns with the GPS simulator technique put forward by Jafarnia-Jahromi et al. (2016, 2). In that work, the authors concur with other disciples who have been advancing the study of GPS interference, that the simplistic type of spoofer can be effective against commercial GPS receivers “especially if the spoofing signal power is higher than the authentic signals” (2). These concepts were further amplified in a paper by Haider and Khalid (2016) describing a simplistic attack as one that “uses a commercial GPS signal simulator, amplifier and antenna to broadcast signals towards target GPS receiver” (573). By consensus, the simplistic technique is relatively easy to implement. However, getting a spoofer close enough to the targeted user system to make this approach work presents one obstacle to the effective execution of such an attack. For this approach to be effective, the spoofer must be co-located with the targeted user system. This approach is also referred to as a proximity-based attack in the literature.

The intermediate spoofing attack (described in Figure 11) works via a portable receiver-spoofer – the receiver component captures the authentic GPS signals to approximate its own PNT (Humphreys et al. 2008). Using these approximations, the receiver-spoofer then generates similarly structured bogus signals to carry out the spoofing attack. The receiver-based spoofer presented by Jafarnia-Jahromi et al. (2012) also falls within the intermediate spoofing category. The authors believe that a spoofer needs detailed information such as the antenna position and velocity of the targeted system as they are aligned with the authentic signals. The authors further suggest that a spoofer has to synchronize their system with the GPS signal to be able to extract satellite information such as position, time, and ephemeris which it then uses to "generate the spoofing signal knowing the 3D pointing vector of its transmit antenna toward the target receiver antenna" (Jafarnia-Jahromi et al. 2012, 2). Humphreys et al. (2008) also believe that for this method to be effective in impacting a valid GPS solution, the spoofer must be located near the user equipment to align the spoofing signal with the authentic signal. One drawback to the intermediate spoofing is the lack of intimate knowledge of the genuine GPS signal. This weakness in the intermediate attack technique allows this type of spoofing to be detected easily (Humphreys et al. 2008). An intermediate attack using the methods described here can be more difficult to detect than an attack carried out by the simplistic method. Both the simplistic and intermediate spoofing methods are proximity-based spoofing attacks. An appropriately configured DAGR should identify and discriminate both simplistic attack and intermediate attack methods and reject such efforts. In addition, an operational military GPS receiver will employ multiple layers of security to include physical security, protection of data in transit (DIT), and data at rest (DAR), thereby decreasing the likelihood of such an attack being successful.

The third, and most advanced, type of spoofing technique, sophisticated spoofing attacks (listed in Figure 11) are more complex and more effective than the previous two methods. Humphreys et al. (2008), presented a sophisticated attack using multiple phase-locked portable receiver-spoofers. Montgomery, Humphreys and Ledvina (2009) further describe a sophisticated spoofer design as one that expands on the principles of the GPS receiver-based spoofer. The sophisticated receiver-based spoofer adds multiple transmit antenna functions to the intermediate concept. The sophisticated spoofer has the capacity "to vary both the carrier and code phase outputs transmitted by each antenna and to control the relative code/carrier phases among these transmit antennas" (Montgomery, Humphreys, and Ledvina 2009, 126). The sophisticated receiver-based spoofer "is assumed to know the centimeter level position of the target receiver antenna phase center to perfectly synchronize the spoofing signal code and carrier phase to those of authentic signals at the receiver" (Jafarnia Jahromi 2013, 21). Tippenhauer et al. (2011) describe a sophisticated attack in which an adversary can inject shifts in the angle of arrival of GPS signals. Jafarnia-Jahromi et al. (2012) further stress that the sophisticated spoofer can exploit the concept of multiple transmit antennas used in this method to defeat direction of arrival anti-spoofing techniques.

The consensus from these works is that GPS spoofing can be executed in three common ways from simplistic to sophisticated. The effort in creating these interference methods become more difficult as the attacks move up the scale. It is recognized that the spoofing detection methods increase in sophistication along with the attack vectors. This is not lost on the authors and researchers; when discussing sophisticated receiver-based spoofers, Jafarnia-Jahromi et al. (2012) declare that "the realization of these types of spoofers is very difficult and, in many cases, impossible due to the geometry and movement of the target receiver antenna." At this point, the researchers (us), did not have the tools and resources to execute the sophisticated spoofing attack. However, it is well worth noting that this level of sophistication exists. With adversaries persistently looking for fissures of vulnerabilities within our infrastructure, it is conceivable that sophisticated attack modes are being studied with the goal of executing a clandestine attack at the opportune moment with maximum effect.

B. HIGH LEVEL CONCEPT FOR A GPS-SPOOFING–DETECTION AI SYSTEM

The work documented in this study was undertaken with the vision of applying to an operation construct. The goal of any new development should be to add to the efficiency of the operating forces without adding any new physical or mental load to the operator. Considering that SWaP should be a design criterion, during the development phase, we imagined tools that can positively impact operations while keeping said design criteria in focus. The spoofing detection is conceived and developed as a software application that can be installed in an operations center from which active monitoring of force movement is dictated by CONOPS. The vision for the system is captured in Figure 12. The figure shows several Command and Control (C2)/ Situational Awareness (SA) systems currently operated by U.S. Forces. The Joint Battle Command Platform (JBC-P) Family of Systems (FoS) is a joint C2/SA system, marketed as "the Army's next-generation friendly force tracking system, equipping Soldiers with a faster satellite network, secure data encryption and advanced logistics" (Program Executive Office Command, Control and Communications-Tactical (PEO C3T) n.d.). The Marine Corps developed developed the Marine Air-Ground Task Force Common Handheld (MCH) to be fielded as part of the JBC-P FoS. MCH uses commercial off the shelf (COTS) hardware to serve as "a tablet-based communication system that enhances situational awareness on the battlefield" (Gonzales 2019b). As a portable handheld system, the device enables dismounted Marines to plot and display "own position" as well as be made aware of friendly and enemy positions. Another dismounted system in the battlespace is the Army's Nett Warrior (NW) system. Nett Warrior also leverages COTS hardware on a portable platform "to provide situational awareness to the dismounted leader" (Program Executive Office Soldier 2019).

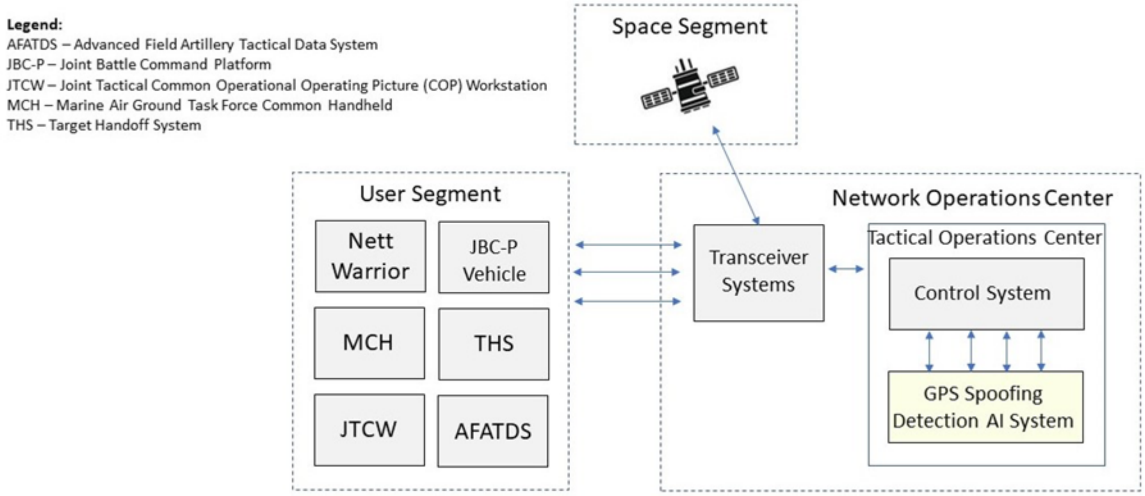

Legend:
AFATDS – Advanced Field Artillery Tactical Data System
JBC-P – Joint Battle Command Platform
JTCW – Joint Tactical Common Operational Operating Picture (COP) Workstation
MCH – Marine Air Ground Task Force Common Handheld
THS – Target Handoff System
Space Segment
User Segment
Nett Warrior
JBC-P Vehicle
MCH
THS
JTCW
AFATDS
Network Operations Center
Transceiver Systems
Tactical Operations Center
Control System
GPS Spoofing Detection AI System

Figure 12. GPS Spoofing Detection Operational Concept

Digital Fires systems are represented by the advanced field artillery tactical data system (AFATDS) and the Target Handoff System (THS). Both the Marine Corps and U.S.

Army use AFATDS "to provide automated support for planning, coordinating, controlling, and executing fires and effects" (Raytheon 2021). THS v2 also leverages COTS devices to provide "a lightweight, fire control system to perform targeting functions" (Gonzales 2019a). The Joint Tactical Common Operational Picture (COP) Workstation (JTCW) "combines seven tactical applications into one user interface" thereby providing "a single digital display of relevant operational information shared by battalion and higher leadership connected on the Global Combat and Control System Tactical Combat Operations network (GCCS-TCO)" (Browne 2016). The systems described in this section are a mix of portable handheld devices and vehicle mounted systems. The common thread through these systems is the need to always provide C2 and SA to friendly operating forces whether in mounted or dismounted platforms. GPS becomes a critical component in this environment as it can be a force multiplier when providing legitimate information.

The AI system can be installed on a monitoring system located in the operations center where tracks are monitored. Most of the systems discussed communicate using some variant of the MIL-STD-6017 variable message format (VMF) a K05.1 position report message can be used to report "own position" on a periodic basis as dictated by doctrine. The study shows that a Marine in the field will not be aware that the position on being tracked has been spoofed, but with the spoofing detection system the operators in the operations center will have visibility on the wayward nature of the navigation and reach. We have investigated and discussed spoofing methods and the seriousness of their impact. The results of this study will provide an enhancement to the current system.

C. PROTOTYPE GPS-SPOOFING–DETECTION AI SYSTEM FOR EXPERIMENTATION

The design concept for the system was developed based on the research done and with the understanding that FDI can be effected in different ways. The methodology and tools used for the experimental system development are covered in this section. This section adds the systems engineering model and the requirements of the system. Success of the experiment will serve to validate the system design concept as well as the objectives of the study.

1. GPS Data for Experimentation

The development of decision aids using AI requires datasets that represent specific mission threads identifying anomalous activities in the GPS data. The GPS data for this study contained both over-the-air signals, also called “live sky” data, and data from a GPS simulator. Orolia is one of the companies that manufactures and sells GNSS simulators. The company markets their family of GNSS simulators as “radio frequency generating instruments that can transmit the same data as GPS satellites” (Orolia 2018, 1). GPS simulators give users the option to create scenarios to manipulate parameters of interest from a test bench in evaluating a system under test. With a GNSS simulator, a user can simulate satellite signals that are then used to verify receiver signal acquisition (Orolia 2018). Using GNSS simulators as a data source provided the option and flexibility to manipulate GPS data and message information to create mission scenarios that capture anomalous activity. For this study, we used a local data repository to obtain operational data as one source of truth data. Simulated data supplemented the operational data used in the study. Data pre-processing and processing were conducted using data mining tools before serving as an input to the AI/ML system.

2. Prototype System Description

The DAGR served as the device under interrogation and along with the equipment needed to support the execution of the experiment, formed the system under evaluation as depicted in Figure 12. The system under evaluation is also referred to as the engineered system and is comprised of the components listed in Table 2.

Table 2. Engineered System Equipment List

Component	Manufacturer	Serial Number	Software version
AN/PSN-13 DAGR	Rockwell Collins	92840004	984-3006-009
GPS Active DAGR Antenna	Rockwell Collins	N/A	N/A
GSG-62 Multiband GNSS Simulator	Orolia	202189	V8.1.1
C21 Combiner	General Dynamics	N/A	N/A

Component	Manufacturer	Serial Number	Software version
Computer	National Instrument	00371-OEM-8992671-0040	Windows 7 1.0.0f4

The DAGR is primarily a handheld unit, but it can be installed in host platforms such as vehicles or control centers. The manufacturers of the DAGR markets the device as "a portable, versatile and precise, guiding tool used in vehicular, handheld, sensor and gun-laying applications" (BAE Systems 2020, 1). As a force multiplier, the DAGR enables military personnel to execute tasks such as navigation, hazard avoidance, and location marking. The DAGR contains four external connectors, marked as J1–J4. The J1 port serves as an RS-232 compatible 2-way serial data I/O port, while the J2 port serves as both RS-232 and RS-422 compatible 2-way serial data I/O port. The J3 port provides an external antenna input, while the J4 port serves as an external power input port. For this study, the DAGR was used to track, collect, and process GPS signals on both the L1 and L2 frequency bands. The J2 port was used to facilitate two-way communication between the DAGR and the computer terminal shown in Figure 13. The computer terminal was used as the local data collection and processing hub. The J2 port integrates both COM1 and COM2 feeds via the RS-232 and RS-422 compatible 2-way serial data I/O port. This set up allowed for the collection of both NMEA and ICD 153 traffic on the laptop terminal through the J2 port. The DAGR, through the 3 port RF combiner, received input signals on the J3 antenna port from both the GPS antenna and the GNSS simulator. It should be reiterated here that GPS signals at a receiver interface are low powered signals, in the region of -160 dBm to -130 dBm. The active L1/L2 RF antenna port on the DAGR is used to amplify the input GPS signal above the noise floor. The configuration used in this setup allowed for the creation of scenarios with the GNSS simulator used to inject an alternate signal feed to the DAGR. Power was supplied to the DAGR through port J4 from the DC Power supply.

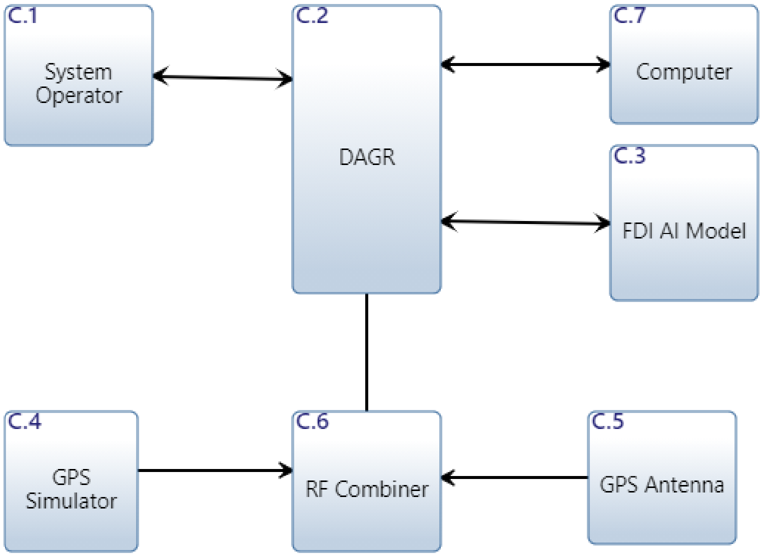


Figure 13. Engineered GPS System

GPS signal generators are radio frequency generating instruments that can transmit the same data as GPS satellites (Perdue 2017). Using this principle, the experiment was designed to take advantage of this operational principle of the simulator. Using the simulator allowed for the creation of scenarios that mimicked the satellite constellation and a variety of parameters as seen by the DAGR. The simulator was used as one signal source, while the live sky signal was used as a second source. The GPS 3 port RF combiner consists of two input ports and one output port and is typically used in scenarios "where two inputs from active GPS antennas are combined evenly into a single receiving GPS unit" (General Dynamics Mission Systems 2021, 1). The GNSS simulator signal was fed to one input, while the input from the active GPS antennas was fed to the second input port. The output port served as the source to the DAGR antenna port. The time synchronization provided the timing function to the system, ensuring that all devices were synchronized to the same time base. The GPS jammer is not an active component in the scenarios used for this experiment but was included in the system to provide the option to exercise the second GPS interference method—jamming—mentioned elsewhere in this study.

3. Experiment Concept of Operations

The system setup shown in Figure 13 formed the basis of the experiment. Multiple scenarios were created to add FDI to the system using the simplistic attack method described earlier. INCOSE defines a system as "an arrangement of parts or elements that together exhibit behavior or meaning that the individual constituents do not" (INCOSE n.d). It is understood that systems can exist in a physical context, a conceptual context, or any combination of both, according to INCOSE. For this study, systems engineering principles were applied to the physical and conceptual model. INCOSE defines systems engineering as "a transdisciplinary and integrative approach to enable the successful realization, use, and retirement of engineered systems, using systems principles and concepts, and scientific, technological, and management methods" (INCOSE n.d). The result of the systems engineering process is an engineered system. The engineered system used in this experiment was designed to function in a simulated operational environment to prove the stated hypothesis of the study while advancing the study of GPS spoofing on military systems. As both jamming and spoofing attacks on GPS are illegal in the U.S., the experiments were carried out in a controlled environment with standalone systems. The FCC portal clearly states that "Federal law prohibits the operation, marketing, or sale of any type of jamming equipment that interferes with authorized radio communications, including cellular and Personal Communication Services (PCS), police radar, GPS" (FCC 2020). Although the FCC statement did not specifically list GPS spoofing, we believe that the law also applies to this type of interference of the GPS devices.

The systems engineering approach used in this study applied model-based systems engineering (MBSE) concepts to the overall system design. One definition of the MBSE concept is "a formalized methodology that is used to support the requirements, design, analysis, verification, and validation associated with the development of complex systems" (Shevchenko 2020). The position taken by Shevchenko (2020), who postulates that MBSE coalesces around the concepts of model, systems thinking, and systems engineering (Shevchenko 2020), finds concurrence in this study: In the text *Application of Artificial Intelligence for Decision-Making*, the authors explain that "systems engineering is based on the definition, allocation, and satisfaction of requirements" (Talbot and Ellis 2015, 14).

Requirements in the systems engineering arena are manifested in multiple categories and formats. For the engineered system concept applied to this study, the requirements included functional, performance, interface, and mission requirements. Functional requirements are tied to operations that the engineered system must execute in an operational construct. Performance requirements describe the measured parameters used to assess the system's performance. Interface requirements map to interconnectivity and interface functions while facilitating the exchange of material, energy, or information with other systems (Talbot and Ellis 2015; Faisandier et al. 2021). Mission requirements define the "activities that constitute the discharge of an agency's official responsibilities" (lawinsider.com 2021). A general concept of MBSE is the idea of an artifact, which is conceived as a source that provides the seed that germinated the concept being explored. Concept exploration was undertaken in the early stages followed by feasibility studies to identify and scope both the problem and the solution space. The AI/ML approach was identified as a viable candidate to tackle the spoofing problem.

It is universally accepted that a model is a conceptual representation of an entity and can exist in various forms to include, graphical, mathematical, or physical forms. Software tools such as Innoslate describe the MBSE structure as one where "models describe physical components performing functions exchanging input/output (I/O) over links, as specified by requirements" (Kristin Giammarco, lecture notes, September 21 2019). Applying this definition to the experiment, the engineered system receives navigation messages through the GPS input applied through user input as specified by the requirements. Innoslate uses the Life cycle Modeling Language (LML) in its operations. In the LML schema, an asset performs an action received by an input and generates an output. The underlying LML model used by Innoslate comprises the following seven classes of entities: "Requirement, Artifact, Action, Asset, I/O, Conduit, and Characteristic entities" (Innoslate 2021). The Universe, Figure 14, and the Metafunction, Figure 15, decompose to the physical and conceptual assets of the engineered system. In Innoslate, I/ O entities provide the means of communication between actions through conduit entities. Innoslate provides a list of definitions for the LML entities used in the tool. These definitions were adopted and applied to the engineered system. Another artifact of the

MBSE process is a Spider diagram. A spider diagram maps the engineered system in both the physical and conceptual contexts. Innoslate describes the spider diagram as "a type of hierarchical organizational chart used as a means of visualizing traceability" (Innoslate 2021). A spider diagram in this context would show the Universe "decomposed by" path to the entities in Figure 14. The entities in the Universe then "performs," through the conduits, the functions in the Metafunction; the DAGR performs the "provide navigation" function, while the FDI AI model performs the "detect spoofing activity" function.

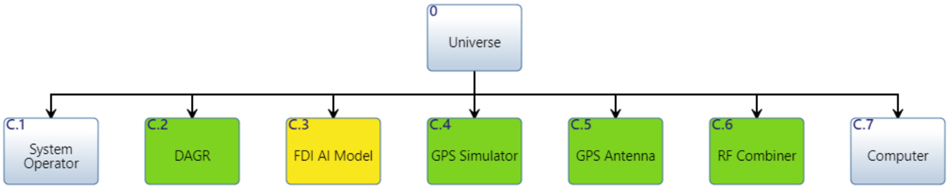


Figure 14. Engineered System Universe Diagram

The engineered system Universe is composed of the elements required for the successful operation of the GPS spoofing generation and the detection of FDI. The external systems do not all exist in the physical domain. The DAGR exists in the physical domain, while the FDI AI model functions in the conceptual domain represented as a software application. The green blocks in both Figure 14 and Figure 15 represent the functional entities that are directly involved in getting the GPS signal to the DAGR. The yellow block represents the FDI AI model containing the algorithms that perform the spoofing detection function.

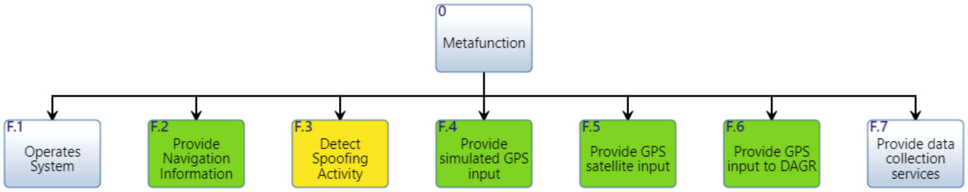

0
Metafunction
F.1
Operates System
F.2
Provide Navigation Information
F.3
Detect Spoofing Activity
F.4
Provide simulated GPS input
F.5
Provide GPS satellite input
F.6
Provide GPS input to DAGR
F.7
Provide data collection services

Figure 15. Engineered System Metafunction Diagram

The engineered system Metafunction represents the Asset entity and specifies the object that performs the Actions of "detect spoofing activities." An I/O entity transfers the information and input data that triggers the output from the spoofing Action. A Conduit entity serves as the transfer medium between the system Assets. The Risk entity is tangentially represented by the FDI AI model as it also captures the combined probability and consequence of not detecting the presence of FDI on the system.

At the functional and operational level, a basic requirement of a GPS device is to provide navigation. As GPS finds applications in industries as diverse as communications, banking systems, package delivery, and financial markets, precise time synchronization is required for the smooth functioning of these operations. Some of the operations listed are time critical, thereby making timing another important GPS requirement. These basic requirements are at the heart of the PNT concept on which the DAGR operates. In the MBSE construct, "a requirement entity identifies a capability, characteristic, or quality factor of a system that must exist for the system to have value and utility to the user" (Innoslate 2021). The requirements for the engineered system were tailored to both the operational and functional contexts of the experiment. The requirements include:

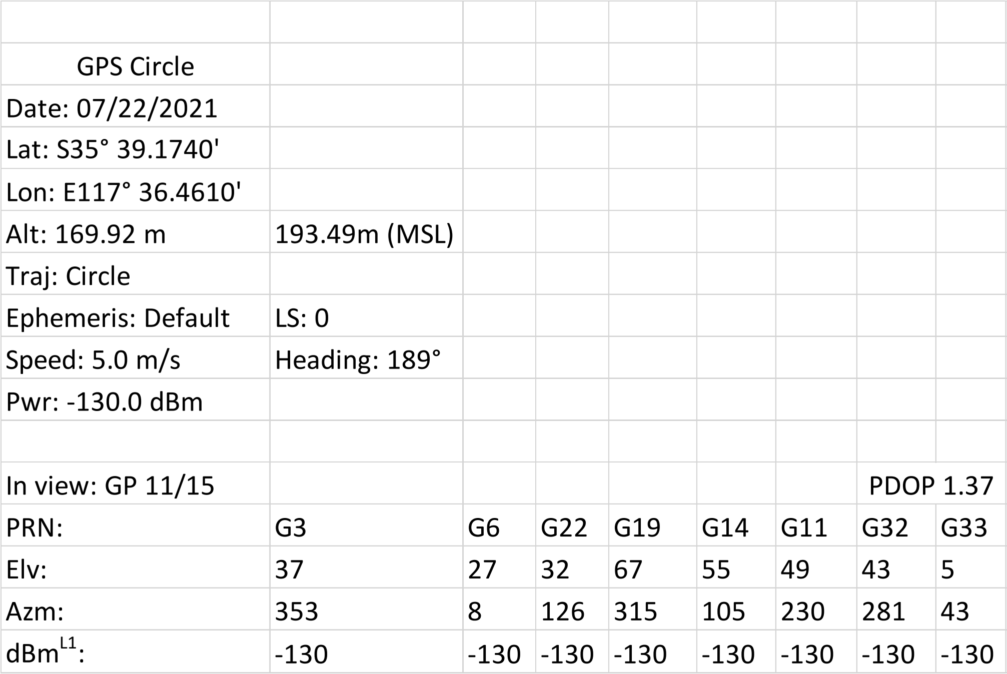
- The system shall operate as a software application.
- The system shall receive NMEA 0183 navigation messages.
- The system shall receive ICD 153 navigation messages.
- The system shall receive GPS synchronization timing information.
- The system shall receive GPS input signals through the antenna port.

- The system shall reject a spoofing attack.
- The system shall accept navigation input from the user.
- The system shall operate in accordance with USMC CONOPS.

For this study, the Orolia GSG-62 Multiband GNSS simulator was used to generate GPS spoofing signals through the process of injecting FDI into the engineered system. The message traffic data was summarily captured for post processing, data analysis, and training of the AI/ML algorithm. To insert FDI on the engineered system, several spoofing scenarios were created and introduced to the system by the GNSS simulator. The operators pushed FDI on the system to manipulate parameters within the message structure with the goal of impacting the GPS solution, PNT, presented to the user. Message traffic used by military navigation systems, including the DAGR, are implemented in accordance with both NMEA 0183 and ICD-153-GPS message standards. The standards recommend that GPS messages should be received at one second intervals. For the engineered system, the GPS message received on the J3 port of the DAGR represented message traffic from both live sky feed and the GNSS simulator. The message traffic from both sources is directed to the DAGR by way of the RF 3-port combiner. The message traffic from the DAGR was delivered to the computer terminal by way of the J2 serial data I/O port. Software applications on the computer terminal were used to record the message traffic data stream. A LabVIEW program resident on the computer terminal was used to visualize, capture, and process the NMEA 0183 message traffic. LabVIEW is a software application tool that aids in visualizing message traffic in real time. The message traffic captured by the LabVIEW application was post-processed and used as an input to the AI/ML algorithm. The ICD-153 message traffic was similarly captured using the compliance tester software for ICD-GPS-153 (CTS-153) software. Similar to the LabVIEW application used for the NMEA 0183 messages, the CTS-153 software application was used to visualize, monitor, and log the ICD-153 message traffic from the DAGR.

The initial phase of the experiment was structured to capture message traffic in a typical operational construct. The normal operations construct in this context refers to a situation wherein the DAGR receives GPS signals from the operational constellation without any artificial interference. As discussed earlier, the GPS signal presented to the receiver is a low power signal. To add some context to the GPS power levels, Warner and Johnston (2012) describe the GPS signal strength as "roughly equivalent to viewing a 25-Watt light bulb in Japan from Los Angeles, California." Imagine that for a moment and think of the inherent vulnerabilities at play in this arena. This finding aligns perfectly with the argument presented by multiple sources from the literature review on GPS spoofing, which demonstrated that by changing power levels of a spoofing signal, FDI can be generated on GPS.

The concept of varying power level on the GPS signal was used as a starting point for this exercise. The GSG-62 output signal level ranges from -65 to -160 dBm. The starting point for the scenarios in this study was -130 dBm. The main menu on the GSG-62 includes information such as the name of the scenario, the scenario start date, the transmit power, the trajectory, and the current position of the scenario. A total of four screen views are available on the GSG-62 during the execution of the scenario. Figure 16 shows the display panel settings for one of the scenarios run in this study. In view shows GP 11/15 indicating that 11 of 15 GPS satellites are in view for this scenario. The PRN field represents the IDs of the satellites, where the 'G' represents the U.S. system. The set of valid IDs for GPS are 1–32. The satellite elevation (Elv) ranges between 0° and 90° and shows the angle between the horizontal plane and the satellite position. Azimuth (Azm) represents a direction measure ranging between 0° and 360°. The dBm parameter represents the transmit power ratio for the L1 frequency band. Only 8 channels are shown in Figure 16 in keeping with the GSG-62 function of displaying a maximum of 8 channels per view. The PDOP value of 1.37 represents an accurate positional measure. With all DOP measures, lower is better and values between 1 and 2 are desirable for navigation.

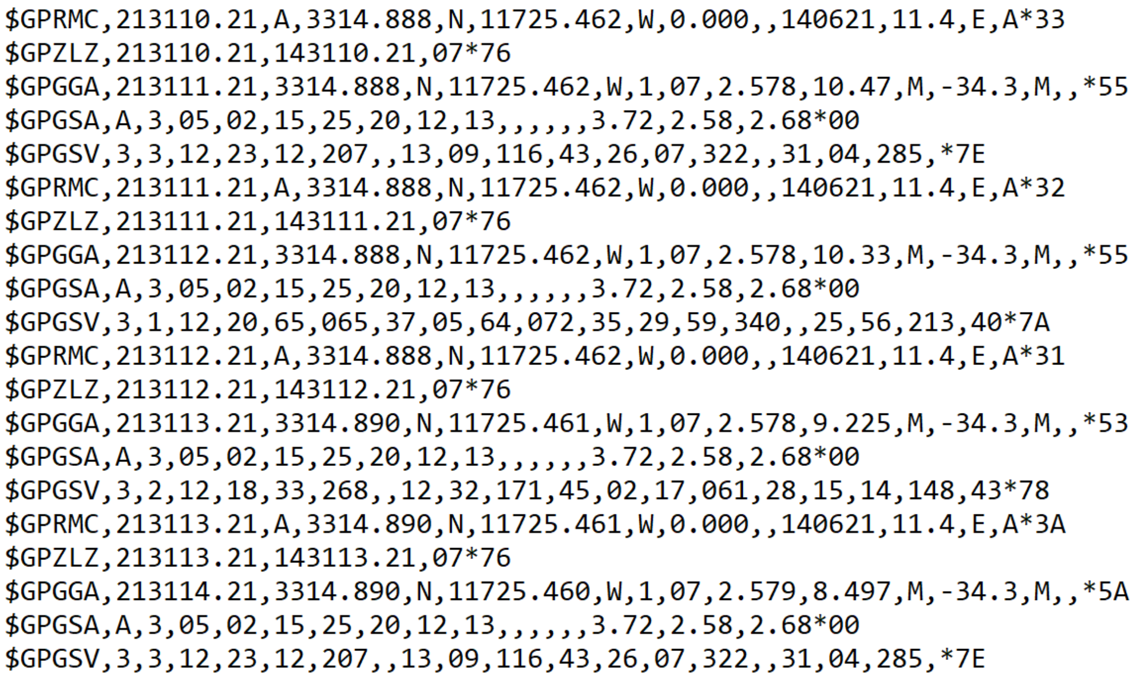


GPS Circle								
Date: 07/22/2021								
Lat: S35° 39.1740'								
Lon: E117° 36.4610'								
Alt: 169.92 m	193.49m (MSL)							
Traj: Circle								
Ephemeris: Default	LS: 0							
Speed: 5.0 m/s	Heading: 189°							
Pwr: -130.0 dBm								
In view: GP 11/15							PDOP 1.37	
PRN:	G3	G6	G22	G19	G14	G11	G32	G33
Elv:	37	27	32	67	55	49	43	5
Azm:	353	8	126	315	105	230	281	43
dBm^{L1}:	-130	-130	-130	-130	-130	-130	-130	-130

Figure 16. GSG-62 Display Parameters

Monitoring message ID 5040 at the start of the application of FDI, the *EPE* was recorded as 104.04 m, the operating mode was *0h* (continuous), Encryption Type was set to 1 (Y-code), Tracking State set to 5 = Carrier track/data (Good Track Mode), Code Type set to 1=C/A-code, Chan Status A/B Valid set to 0=Data Valid, Channel 1 Status B J/S = 28 dB, Channel 1 Status B C/No = 32 dB-Hz. Similar values were recorded for Channel 2; however, Channels 3 – 5 all recorded J/S and C/No = 0. Antenna Source was recorded as 1=External. There were nine satellites representing both the Number of Visible space vehicles (SVs) and the Number of Healthy SVs. An Almanac Age of one day was recorded for the duration of the experiment, in keeping with the operational concepts of the system. These parameters were verified by looking at both NMEA 0183 and ICD 153 message data traffic.

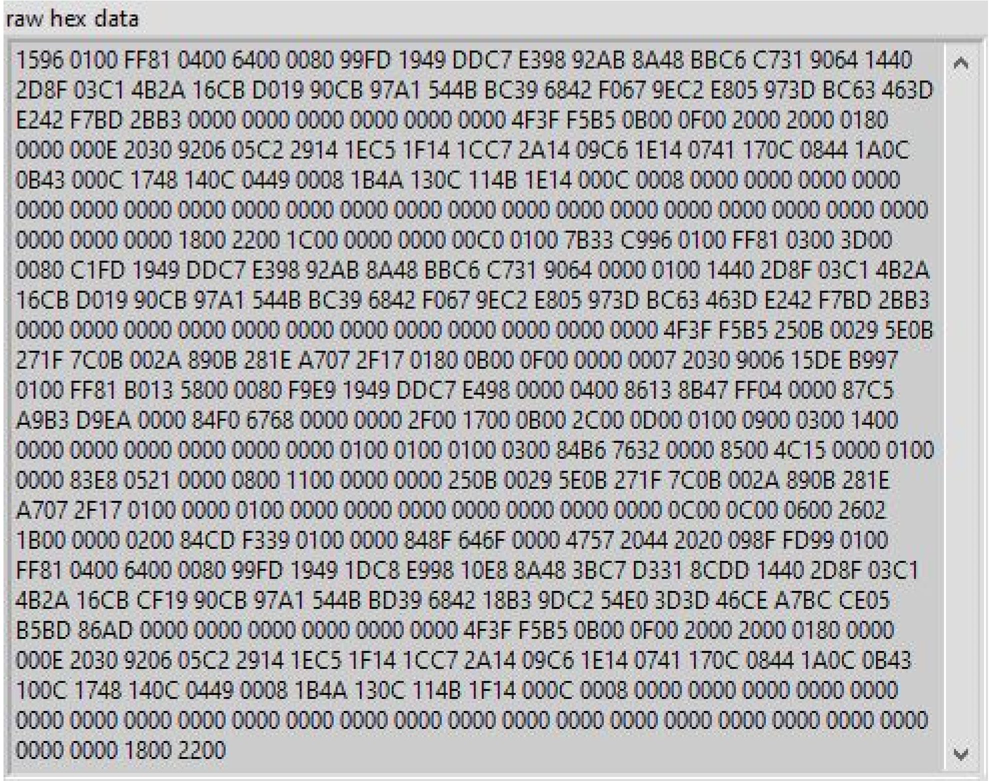
A valid GPS solution will place a user in the pre-planned place and time that was input to the GPS equipment, within the statistical boundaries prescribed by the message standards. Extraneous factors such as environmental conditions and the interference methods studied in this work also impact a valid GPS solution. Moving on with this framework, the experiment was structured to present scenarios that put FDI in the GPS message traffic data with the goal of impacting the valid GPS solution. The "P" in the PNT represents position and was accepted as a likely candidate on which to focus in this study. To capture the essence of normal GPS message traffic, the engineered system was configured and run with the live sky feed as the sole GPS message traffic, thereby providing the basis of the training dataset. Following this, the GSG-62 GNSS simulator was introduced into the configuration. The simplistic GPS spoofing method was then started with scenarios mimicking the power levels, position, and timing factors of the live sky feed. The operational concept for infecting the system with FDI involved incrementally increasing the input signal levels on the GSG-62 and monitor the system for any changes. The message parameters were monitored on the LabVIEW and CTS-153 applications. During the scenario execution, adjusting the dBm parameter applies the changes to all the satellites at the same time. A sample of the message traffic data for the NMEA 0183 message is shown in Figure 17. The file shows five types of messages, GPRMC, GPZLZ, GPGGA, GPGSA, and GPGSV. As described earlier in this report and displayed in Figure 17, the NMEA messages begin with *$* and end with the * before the value representing the checksum. Other message parameters such as GPS time, latitude, longitude, and date can be seen in the GPRMC message.



```
$GPRMC,213110.21,A,3314.888,N,11725.462,W,0.000,,140621,11.4,E,A*33
$GPZLZ,213110.21,143110.21,07*76
$GPGGA,213111.21,3314.888,N,11725.462,W,1,07,2.578,10.47,M,-34.3,M,,*55
$GPGSA,A,3,05,02,15,25,20,12,13,,,,,,3.72,2.58,2.68*00
$GPGSV,3,3,12,23,12,207,,13,09,116,43,26,07,322,,31,04,285,*7E
$GPRMC,213111.21,A,3314.888,N,11725.462,W,0.000,,140621,11.4,E,A*32
$GPZLZ,213111.21,143111.21,07*76
$GPGGA,213112.21,3314.888,N,11725.462,W,1,07,2.578,10.33,M,-34.3,M,,*55
$GPGSA,A,3,05,02,15,25,20,12,13,,,,,,3.72,2.58,2.68*00
$GPGSV,3,1,12,20,65,065,37,05,64,072,35,29,59,340,,25,56,213,40*7A
$GPRMC,213112.21,A,3314.888,N,11725.462,W,0.000,,140621,11.4,E,A*31
$GPZLZ,213112.21,143112.21,07*76
$GPGGA,213113.21,3314.890,N,11725.461,W,1,07,2.578,9.225,M,-34.3,M,,*53
$GPGSA,A,3,05,02,15,25,20,12,13,,,,,,3.72,2.58,2.68*00
$GPGSV,3,2,12,18,33,268,,12,32,171,45,02,17,061,28,15,14,148,43*78
$GPRMC,213113.21,A,3314.890,N,11725.461,W,0.000,,140621,11.4,E,A*3A
$GPZLZ,213113.21,143113.21,07*76
$GPGGA,213114.21,3314.890,N,11725.460,W,1,07,2.579,8.497,M,-34.3,M,,*5A
$GPGSA,A,3,05,02,15,25,20,12,13,,,,,,3.72,2.58,2.68*00
$GPGSV,3,3,12,23,12,207,,13,09,116,43,26,07,322,,31,04,285,*7E
```

Figure 17. NMEA 0183 Message Data

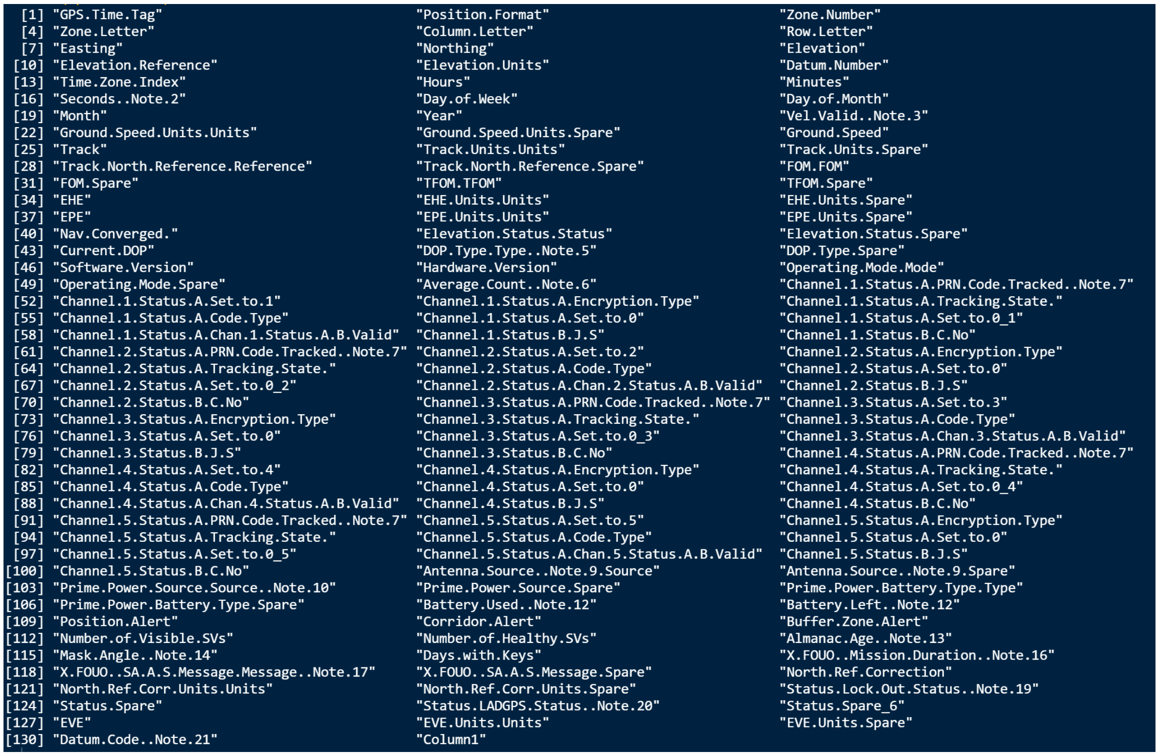
The raw ICD-153 message traffic is captured as a hexadecimal file representing the message, a sample of which is shown in Figure 18. The raw data file was post-processed using the CTS-153 software and presented as a **.prn* file that can be consumed by most generalized text editors. The post-processed data file consists of rows and columns, where each row represents one second interval of message traffic. The columns represent the various data parameters of the dataset. The ICD -153 message ID 5040, for instance, consists of 131 columns of data.



```
1596 0100 FF81 0400 6400 0080 99FD 1949 DDC7 E398 92AB 8A48 BBC6 C731 9064 1440
2D8F 03C1 4B2A 16CB D019 90CB 97A1 544B BC39 6842 F067 9EC2 E805 973D BC63 463D
E242 F7BD 2BB3 0000 0000 0000 0000 0000 0000 4F3F F5B5 0B00 0F00 2000 2000 0180
0000 000E 2030 9206 05C2 2914 1EC5 1F14 1CC7 2A14 09C6 1E14 0741 170C 0844 1A0C
0B43 000C 1748 140C 0449 0008 1B4A 130C 114B 1E14 000C 0008 0000 0000 0000 0000
0000 0000 0000 0000 0000 0000 0000 0000 0000 0000 0000 0000 0000 0000 0000 0000
0000 0000 0000 1800 2200 1C00 0000 0000 00C0 0100 7B33 C996 0100 FF81 0300 3D00
0080 C1FD 1949 DDC7 E398 92AB 8A48 BBC6 C731 9064 0000 0100 1440 2D8F 03C1 4B2A
16CB D019 90CB 97A1 544B BC39 6842 F067 9EC2 E805 973D BC63 463D E242 F7BD 2BB3
0000 0000 0000 0000 0000 0000 0000 0000 0000 0000 0000 0000 4F3F F5B5 250B 0029 5E0B
271F 7C0B 002A 890B 281E A707 2F17 0180 0B00 0F00 0000 0007 2030 9006 15DE B997
0100 FF81 B013 5800 0080 F9E9 1949 DDC7 E498 0000 0400 8613 8B47 FF04 0000 87C5
A9B3 D9EA 0000 84F0 6768 0000 0000 2F00 1700 0B00 2C00 0D00 0100 0900 0300 1400
0000 0000 0000 0000 0000 0000 0100 0100 0100 0300 84B6 7632 0000 8500 4C15 0000 0100
0000 83E8 0521 0000 0800 1100 0000 0000 250B 0029 5E0B 271F 7C0B 002A 890B 281E
A707 2F17 0100 0000 0100 0000 0000 0000 0000 0000 0000 0000 0C00 0C00 0600 2602
1B00 0000 0200 84CD F339 0100 0000 848F 646F 0000 4757 2044 2020 098F FD99 0100
FF81 0400 6400 0080 99FD 1949 1DC8 E998 10E8 8A48 3BC7 D331 8CDD 1440 2D8F 03C1
4B2A 16CB CF19 90CB 97A1 544B BD39 6842 18B3 9DC2 54E0 3D3D 46CE A7BC CE05
B5BD 86AD 0000 0000 0000 0000 0000 0000 4F3F F5B5 0B00 0F00 2000 2000 0180 0000
000E 2030 9206 05C2 2914 1EC5 1F14 1CC7 2A14 09C6 1E14 0741 170C 0844 1A0C 0B43
100C 1748 140C 0449 0008 1B4A 130C 114B 1F14 000C 0008 0000 0000 0000 0000 0000
0000 0000 0000 0000 0000 0000 0000 0000 0000 0000 0000 0000 0000 0000 0000 0000
0000 0000 1800 2200
```

Figure 18. ICD-153 Raw Hexadecimal Data Sample

The columns represent the various data parameters of the dataset. The ICD -153 5040 “current status” message, for example, consists of 131 columns of data shown in Figure 19. One of the objectives of this study is to apply AI/ML tools to this dataset to discern which combination of predictors from the 131 entities in this dataset are impacted by a spoofing attack.



```
  [1] "GPS.Time.Tag"                                "Position.Format"                             "Zone.Number"
  [4] "Zone.Letter"                                 "Column.Letter"                               "Row.Letter"
  [7] "Easting"                                     "Northing"                                    "Elevation"
 [10] "Elevation.Reference"                         "Elevation.Units"                             "Datum.Number"
 [13] "Time.Zone.Index"                             "Hours"                                       "Minutes"
 [16] "Seconds..Note.2"                             "Day.of.Week"                                 "Day.of.Month"
 [19] "Month"                                       "Year"                                        "Vel.Valid..Note.3"
 [22] "Ground.Speed.Units.Units"                    "Ground.Speed.Units.Spare"                    "Ground.Speed"
 [25] "Track"                                       "Track.Units.Units"                           "Track.Units.Spare"
 [28] "Track.North.Reference.Reference"             "Track.North.Reference.Spare"                 "FOM.FOM"
 [31] "FOM.Spare"                                   "TFOM.TFOM"                                   "TFOM.Spare"
 [34] "EHE"                                         "EHE.Units.Units"                             "EHE.Units.Spare"
 [37] "EPE"                                         "EPE.Units.Units"                             "EPE.Units.Spare"
 [40] "Nav.Converged."                              "Elevation.Status.Status"                     "Elevation.Status.Spare"
 [43] "Current.DOP"                                 "DOP.Type.Type..Note.5"                       "DOP.Type.Spare"
 [46] "Software.Version"                            "Hardware.Version"                            "Operating.Mode.Mode"
 [49] "Operating.Mode.Spare"                        "Average.Count..Note.6"                       "Channel.1.Status.A.PRN.Code.Tracked..Note.7"
 [52] "Channel.1.Status.A.Set.to.1"                 "Channel.1.Status.A.Encryption.Type"          "Channel.1.Status.A.Tracking.State."
 [55] "Channel.1.Status.A.Code.Type"                "Channel.1.Status.A.Set.to.0"                 "Channel.1.Status.A.Set.to.0_1"
 [58] "Channel.1.Status.A.Chan.1.Status.A.B.Valid"  "Channel.1.Status.B.J.S"                      "Channel.1.Status.B.C.No"
 [61] "Channel.2.Status.A.PRN.Code.Tracked..Note.7" "Channel.2.Status.A.Set.to.2"                 "Channel.2.Status.A.Encryption.Type"
 [64] "Channel.2.Status.A.Tracking.State."          "Channel.2.Status.A.Code.Type"                "Channel.2.Status.A.Set.to.0"
 [67] "Channel.2.Status.A.Set.to.0_2"               "Channel.2.Status.A.Chan.2.Status.A.B.Valid"  "Channel.2.Status.B.J.S"
 [70] "Channel.2.Status.B.C.No"                     "Channel.3.Status.A.PRN.Code.Tracked..Note.7" "Channel.3.Status.A.Set.to.3"
 [73] "Channel.3.Status.A.Encryption.Type"          "Channel.3.Status.A.Tracking.State."          "Channel.3.Status.A.Code.Type"
 [76] "Channel.3.Status.A.Set.to.0"                 "Channel.3.Status.A.Set.to.0_3"               "Channel.3.Status.A.Chan.3.Status.A.B.Valid"
 [79] "Channel.3.Status.B.J.S"                      "Channel.3.Status.B.C.No"                     "Channel.4.Status.A.PRN.Code.Tracked..Note.7"
 [82] "Channel.4.Status.A.Set.to.4"                 "Channel.4.Status.A.Encryption.Type"          "Channel.4.Status.A.Tracking.State."
 [85] "Channel.4.Status.A.Code.Type"                "Channel.4.Status.A.Set.to.0"                 "Channel.4.Status.A.Set.to.0_4"
 [88] "Channel.4.Status.A.Chan.4.Status.A.B.Valid"  "Channel.4.Status.B.J.S"                      "Channel.4.Status.B.C.No"
 [91] "Channel.5.Status.A.PRN.Code.Tracked..Note.7" "Channel.5.Status.A.Set.to.5"                 "Channel.5.Status.A.Encryption.Type"
 [94] "Channel.5.Status.A.Tracking.State."          "Channel.5.Status.A.Code.Type"                "Channel.5.Status.A.Set.to.0"
 [97] "Channel.5.Status.A.Set.to.0_5"               "Channel.5.Status.A.Chan.5.Status.A.B.Valid"  "Channel.5.Status.B.J.S"
[100] "Channel.5.Status.B.C.No"                     "Antenna.Source..Note.9.Source"               "Antenna.Source..Note.9.Spare"
[103] "Prime.Power.Source.Source..Note.10"          "Prime.Power.Source.Spare"                    "Prime.Power.Battery.Type.Type"
[106] "Prime.Power.Battery.Type.Spare"              "Battery.Used..Note.12"                       "Battery.Left..Note.12"
[109] "Position.Alert"                              "Corridor.Alert"                              "Buffer.Zone.Alert"
[112] "Number.of.Visible.SVs"                       "Number.of.Healthy.SVs"                       "Almanac.Age..Note.13"
[115] "Mask.Angle..Note.14"                         "Days.with.Keys"                              "X.FOUO..Mission.Duration..Note.16"
[118] "X.FOUO..SA.A.S.Message.Message..Note.17"     "X.FOUO..SA.A.S.Message.Spare"                "North.Ref.Correction"
[121] "North.Ref.Corr.Units.Units"                  "North.Ref.Corr.Units.Spare"                  "Status.Lock.Out.Status..Note.19"
[124] "Status.Spare"                                "Status.LADGPS.Status..Note.20"               "Status.Spare_6"
[127] "EVE"                                         "EVE.Units.Units"                             "EVE.Units.Spare"
[130] "Datum.Code..Note.21"                         "Column1"
```

Figure 19. ICD-153 message ID 5040 Data Fields

Taking a cursory look at the list of parameters in Figure 19, we were able to filter the number of parameters factored in the solution by eliminating the items deemed not relevant to the study. Principal Components Analysis (PCA) is one data reduction tool that can be applied to the dataset to perform dimension reduction. The findings based on the data reduction theory are presented later in this study.

D. MACHINE LEARNING EXPERIMENTATION AND DEMONSTRATION

In the previous section we discussed the systems engineering process and the creating of the experiment and requirements following the principles. This section moves into a description of the tools that will be used to create the models and conduct data analysis of the results. The distinction between "normal" and spoofed data will be revealed and analyzed to identify the impact of undetected application of FDI on the DAGR.

1. Machine Learning Tools

Machine learning is a subfield of AI that facilitates learning about a dataset by feeding it a training set. At its core, ML builds on the concept of teaching computers to learn from experience, and in so doing, to mimic what humans and animals naturally do (MathWorks 2020). Simeone (2018) describes ML as the approach taken to train a system to predict an outcome based on a large dataset, and then using the trained algorithm to explore ways to produce more advantageous outcomes (Simeone 2018). To concretize the concept of ML, Huddleston and Brown (2019) puts forward the notion that “the defining characteristic of ML is the focus on using algorithmic methods to improve descriptive, predictive, and prescriptive performance in real-world contexts.” Burkov (2019) expands on this definition by describing ML as “the process of solving a practical problem by 1) gathering a dataset, and 2) algorithmically building a statistical model based on that dataset.” Hall and Gill (2018), in discussing the social and commercial motivations for ML interpretability, acknowledge that “usage of AI and ML models is likely to become more commonplace as larger swaths of the economy embrace automation and data-driven decision making.” The results of this study will lend itself well to the concept of embracing automation and aid in the data-driven decision-making process. Hall and Gill’s perspective that increased convenience, coupled with automation and better structure in one’s daily activities, being one of the greatest hopes of applying data science and ML confirms the conclusions of this study. Having an automated system that detects spoofing activities reduces the burden on an operator by producing unbiased decision making in the face of FDI.

The explanations of ML given in the previous paragraph are but a small sample of the definitions being applied to the AI/ML sphere. There is an overall belief that there are three groups of ML algorithms: supervised learning, unsupervised learning, and reinforcement learning (Russell and Norvig 2003; Han and Kamber 2006; Huddleston and Brown 2019; Simeone 2018). Huddleston and Brown (2019) highlight that the three ML algorithms are not mutually exclusive. Both supervised and unsupervised learning ML paradigms can be applied to a dataset to resolve tasks by learning from experience and performance measures applied. One definition of reinforcement learning put forward by

Microsoft is “the study of decision making over time with consequences” (Microsoft 2021). Burkov (2019), states that “the goal of a reinforcement learning algorithm is to learn a policy.” Sutton and Barto (2015) further add that reinforcement learning problems involve learning what actions are required to maximize a numerical reward signal. The dataset used in this study is not readily applicable to reinforcement learning and therefore will not be discussed further.

Supervised learning is a ML technique used to develop a predictive model from labeled training data by mapping an input to an output based on input and output pairs. Supervised learning requires a teacher to help guide the learning process (Russell and Norvig 2003). Sutton and Barto (2015) in describing supervised learning, state that “the objective of supervised learning is for the system to extrapolate, or generalize, its responses so that it acts correctly in situations that are not present in the training set.” Burkov (2019, 2) echoes Sutton and Barto (2015) by pointing out that “the goal of a supervised learning algorithm is to use the dataset to produce a model that takes a feature vector ***x*** as input and outputs information that allows deducing the label for this feature vector.” This fact was amplified by (Huddleston and Brown 2019, 2), where they stated that the goal of supervised learning is “to use available observations to predict the response variable associated with new observations.” Supervised learning is generally applied to a dataset that contains a specific variable that the model is trying to predict. Supervised learning occurs when a machine is trained using labeled data.

Supervised learning problems are categorized as regression or classification problems. Han and Kamber (2006, 286) describe classification as the output of a model or classifier that is constructed to predict categorical labels such as “yes” or “no,” “spam” or “legit.” This point is also restated by (Hastie, Tibshirani, and Friedman 2008) in their discussion on the classification construct. James et al. (2017) report that classification problems usually involve the prediction of a qualitative response (James et al. 2017). The IBM Cloud Education portal in defining classification, concur with the descriptions given in the current discussion by adding that “classification uses an algorithm to accurately assign test data into specific categories.” Huddleston and Brown (2019, 23) provide the following list as some of the common classification algorithms in current use:

- K-nearest neighbors (KNN)—“Given a positive integer *K* and a test observation x_0, the KNN classifier first identifies the *K* points in the training data that are closest to x_0, represented by N_0” (James et al. 2017, 39).
- Regression—“widely used in machine learning for prediction, time series forecasting, and classification problems” (Huddleston and Brown 2019, 25).
- Classification and Regression Trees (CART)—“provide a highly visual and human interpretable method for regression and classification problems” (Huddleston and Brown 2019, 27).
- Time Series Forecasting—uses time as the independent variable and predicts future actions based on historical patterns in the dataset.
- Decision Trees—“are an ideal base learner for data mining applications of boosting” (Hastie, Tibshirani, and Friedman 2008, 341).
- Support Vector Machines (SVM)—“the motivating principle for support vector machines is to find a maximum separating hyperplane in feature space that separates classes” (Huddleston and Brown 2019, 30).
- Artificial Neural Networks (ANN)—“are motivated by trying to mimic the way neurons in the brain fire to influence human learning and decision-making” (Huddleston and Brown 2019, 31).
- Ensemble Methods—“combines predictions from several (or many) machine learning algorithms” (Huddleston and Brown 2019, 33).

The discussion of classification shows that the result of classification is a discrete value. On the other hand, “learning on a continuous function is called regression” (Russell and Norvig 2003, 653). This definition of regression is also widely supported in the sphere of mathematics and statistics. James et al. (2017, 28) states that “we tend to refer to problems with a quantitative response as regression problems." Regression problems are varied and include tasks such as predicting the weather based on current conditions or predicting the cost of real estate based on location and size, among other factors. Regression algorithms include generalized linear models, logistic regression, decision trees, linear regression, nonlinear regression, and neural networks. James et al. (2017) posted a summary statement of regression: "in the supervised learning setting, we typically have access to a set of p features $X_1, X_2, \ldots, X_p$, measured on n observations, and a response Y also measured on those same n observations. The goal is then to predict Y using $X_1, X_2, \ldots, X_p$" (James et al. 2017, 373).

While supervised learning involves learning from a training dataset using the input and output data, unsupervised learning involves identifying patterns in the input where there are no predetermined labels. Delua (2021) asserts that "unsupervised learning uses machine learning algorithms to analyze and cluster unlabeled data sets." Huddleston and Brown (2019, 3) state that unlike supervised learning where labeled data is required, "unsupervised learning seeks to identify latent (underlying or hidden) structures in a dataset." Unsupervised learning works without having a labeled response variable. As unsupervised learning attempts to understand the structure implied by a set of variables without a response variable, uncovering patterns or relationships in a dataset is one of the strengths of this construct. This notion lends itself nicely to the detection of anomalies or outliers in this study. James et al. (2017) believe that unsupervised learning presents a more challenging situation than supervised learning since there is no label or teacher with which to evaluate the output. "For every observation $i = 1, \ldots, n,$ we observe a vector of measurements x_i but no associated response y_i, thereby making it impossible to fit a linear regression model, since there is no response variable to predict" (James et al. 2017, 26). "In this setting, we are in some sense working blind; the situation is referred to as unsupervised because we lack a response variable that can supervise our analysis" (James et al. 2017, 26). Huddleston and Brown (2019, 35) list "kernel density estimation, association rule mining, principal component analysis (PCA), clustering methods, and bag-of-words" as some of the most frequently employed unsupervised learning techniques.

James et al. (2017) presented the clustering method as one of the tools that can be used to discern the relationships in the dataset. Clustering was also presented by Huddleston and Brown (2019), amplifying the concept. Clustering assembles data items that have similar properties and exhibit similar qualities together. Data items that are dissimilar are placed in a different cluster. James et al. (2017, 27) acknowledges that "the goal of cluster analysis is to ascertain, on the basis of $x_1, \ldots, x_n$, whether the observations fall into relatively distinct groups." Clustering can be a useful tool for anomaly/outlier detection because in a normally distributed dataset, most of the data is clustered around a mean; any data points outside of this normal distribution should be investigated as potential outliers. Not all parameters in our dataset display properties of a normally distributed set. In the text, *An Introduction to Statistical Learning*, the authors present *K*-means clustering and hierarchical clustering as the two most used clustering methods in use (James et al. 2017). Hastie, Tibshirani and Friedman (2008) introduces *K*-means clustering as one method used for finding clusters and cluster centers in a set of unlabeled data. Meanwhile, Simeone (2018) describes *K*-means as a heuristic method used to cluster together points that are mutually close in Euclidean distance. In the literature, the *K*-means clustering algorithm is presented as one of the most popular iterative descent clustering methods. *K*-means clustering is used in situations in which all variables are quantitative. One of the objectives of *K*-means clustering is the creation of partitions in the dataset into a pre-determined number (*K*) of clusters.

In hierarchical clustering, the number of clusters is not determined prior to the clustering operation. The result of hierarchical clustering is generally represented by a tree-like structure called a *dendrogram.* A dendrogram shows the relationship between clusters and provides an easily digestible representation of the cluster in a graphical format. In some circles, it is believed that the main point of hierarchical clustering is to create the dendrogram because it is the dendrogram that creates the methodology used to extract the most optimal clustering configuration (Eremenko and de Ponteves 2021). At each level of the hierarchy, clusters are created by merging clusters at the next lower level (Michael Atkinson, lecture notes, May 19, 2021). A single cluster representing the full dataset resides at the highest level of the hierarchy, while at the lowest level, each cluster contains a single entity from the dataset. Hierarchical clustering exists in either agglomerative or divisive forms. In the agglomerative form, the starting point represents each observation in its own cluster, which is then merged with other clusters. In the divisive form, the starting point contains all entries of the dataset represented as a single cluster, which is then decomposed into smaller sub-clusters.

A major task in this study is the design of the experiment and data collection. The data collected from the experiment described earlier captures the GPS message traffic that formed the basis of the study. Having good data was crucial to the study because, in the words of W. Edwards Deming "without data, you're just another person with an opinion." This study was built around facts demonstrated throughout this report. Kotsiantis (2007) proposes one method of data collection as depicted in Figure 20. The process starts with identification of the problem. In this study, we identified the problem as "identifying spoofing activities on military GPS." In keeping with the construct proposed by Kotsiantis, we identified the data necessary for the study and proceeded to construct a data collection plan as described in Figure 13. It has long been held that one of the biggest efforts in ML is the preprocessing of data. Data preprocessing is described as a data mining technique that transforms raw data into a clear and useable format (Techopedia 2021).

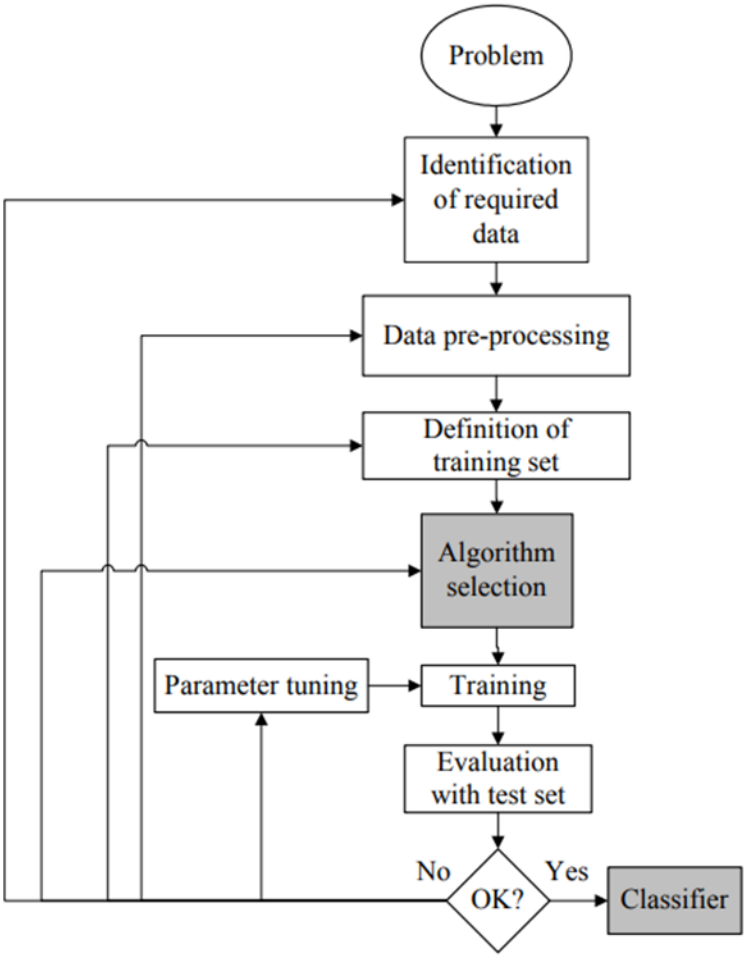


Figure 20. The Process of Supervised Learning. Source: Kotsiantis (2007, 250)

As ML algorithms learn from data, it is vitally important that the data being applied is representative of the problem being studied. Data preprocessing involves verifying that the most meaningful features of the data set are captured and that the appropriate format for the data is being observed. A subset of the features is selected, removing as many irrelevant and redundant features as possible. For the dataset used in this study, we identified the features and parameters that are correlated and that influenced the integrity of the dataset. Recognizing that the selection of the right algorithm to solve the problem is critical, one approach was to exercise multiple ML software applications and algorithms to process the data and use statistical tools to measure and compare the performance of the algorithms. Kotsiantis (2007) in discussing supervised learning, states that “A common method for comparing supervised ML algorithms is to perform statistical comparisons of the accuracies of trained classifiers on specific datasets" (Kotsiantis 2007, 251). This advice equally applies to the unsupervised learning paradigm and has been adopted for this study.

Chen and Lui (2018) opines that "the current dominant paradigm for ML is to run an ML algorithm on a given dataset to generate a model" (1). In this instance, the model is then applied to the dataset for performing the FDI detection tasks. Chen and Lui (2018, 1) refer to this concept as "isolated learning because it does not consider any other related information or the previously learned knowledge." The authors identify deficiencies within the isolated learning paradigm stating that "it does not retain and accumulate knowledge learned in the past and use it in future learning" (Chen and Liu 2018, 1). This statement is contrary to human learning as humans seldom "learn in isolation or from scratch" (Chen and Liu 2018, 1). The concept of human learning is captured by the constructivism theory. The main tenet of the theory is that humans learn by adding "new knowledge upon the foundation of previous learning" (McLeod 2019, 1). Expanding on the constructivism principle, the model should also exhibit perspicacious properties. *Merriam-Webster* gives this formal definition of perspicacious: "having or showing an ability to notice and understand things that are difficult or not obvious" (Merriam-Webster 2021). Former chess grandmaster Garry Kasparov (2018), in an article titled *Intelligent Machines Will Teach Us—Not Replace Us,* recommends that humans should think of AI as "augmented intelligence" because the "intelligent machines are making us smarter, just as our past technology made us stronger and faster" (Kasparov 2018). He goes on to state that "for the first time, machines aren't just giving us answers more quickly and accurately; they are generating new knowledge that helps us better understand the world" (Kasparov 2018). These are sublte, yet important points, that we endeavoured to implement in our FDI system to complement to the human element.

2. Anomaly Detection

A spoofing attack can be represented as an anomaly in a dataset. Anomalies can show up as outliers in a dataset. Outliers are described as "an entry in a dataset that is anomalous behavior relative to the majority of the other entries in the dataset" (Pearson 2005, 23). Outliers can take different forms, with the simplest case being called univariate outliers. In this instance, in a dataset which contains a series of values $\{y_k\}$, each observation y_k may be represented as

$$y_k = y_{nom} + e_k,$$

where y_{nom} represents the nominal value for this data sequence. The factor $\{e_k\}$ represents the error measure of the distance from the nominal value y_k (Pearson 2005).

3. Data Analytics Tools

Investopedia defines data analytics as "a broad term that encompasses many diverse types of data analysis" (Frankenfield 2021). Frankenfield further states that "any type of information can be subjected to data analytics techniques to get insight that can be used to improve things." Data analytics serves as a great tool as it presents techniques that can be used to reveal latent trends and metrics that are not immediately obvious to the untrained eye. This information gleaned from data analytics can be used as an input to streamline and optimize processes resulting in increased efficiency of a business or system (Frankenfield 2021).

There is an abundance of ML tools available for data extraction and decision making in the current environment. MathWorks, in answering the question, "*when should you use machine learning?*" responds with "Consider using machine learning when you have a complex task or problem involving a large amount of data and lots of variables, but no existing formula or equation" (MathWorks 2020, 10). The study being reported in this work certainly fits the profile of a system that will benefit from a ML intervention. A survey of ML shows a strong linkage between ML and statistics as evidenced by the application of statistical tools in the analysis of the output of ML algorithms. Hastie, Tibshirani, and Friedman (2008, 1) write that "the science of learning plays a key role in the fields of statistics, data mining and artificial intelligence, intersecting with areas of engineering and other disciplines." The algorithm of ML takes input variables, manipulates the data, and produces output variables.

In the text, *R for Data Science: Import, Tidy, Transform, Visualize, and Model Data*, the authors posit that data science is a multi-step process as represented in Figure 21. The importance of data quality is captured in the import and tidy functions as this step prepares the dataset from the raw form for presentation to the data science tools. Wickham and Grolemund (2017) explain that "tidying your data means storing it in a consistent form that matches the semantics of the dataset with the way it is stored" (Wickham and Grolemund 2017, 1). Data tidying is complemented by the data preprocessing recommended by Kotsiantis. Applying the concept of tidying to our dataset, the ICD-153 messages have a file structure containing 131 columns or variables. For the NMEA 0183 messages, the number of columns or vary by the message type. For both message standards, each row represents a single observation, because the messages are captured at 1 second intervals. Therefore, the number of observations is determined by the duration of the run time required to capture the data.

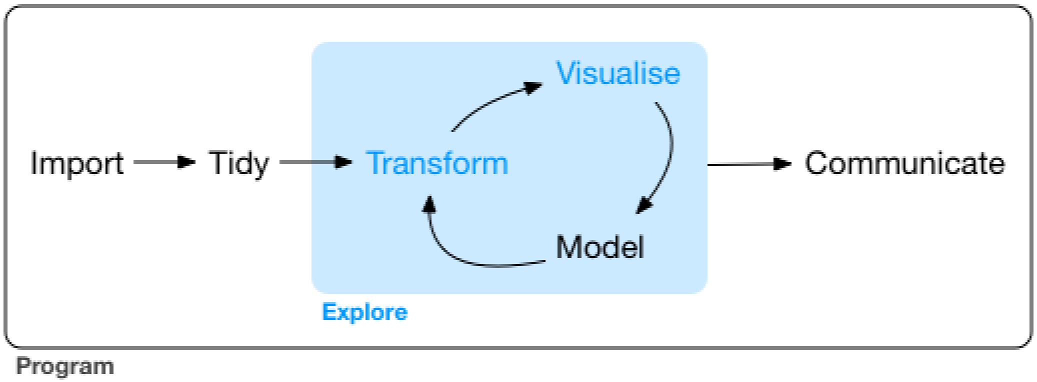
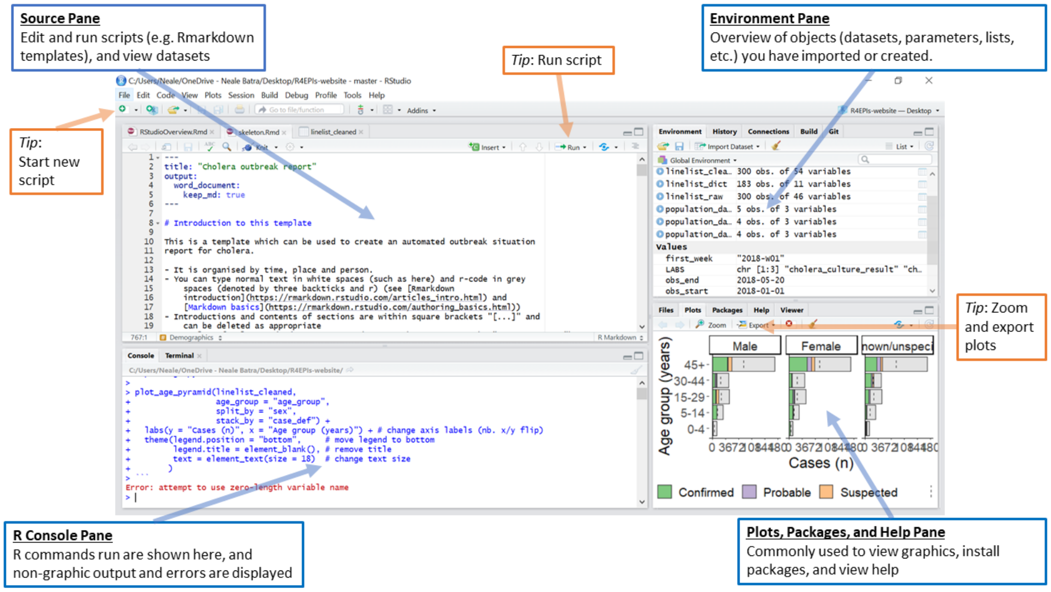


Figure 21. Data Science Construct Source: Wickham and Grolemund (2017)

Wickham and Grolemund (2017) also believe that after the tidy data step, the visualization and modeling steps follow. Visualization and modelling are two main engines of knowledge generation according to Wickham and Grolemund (2017). Visualization and modeling are regarded as complementary tools as they blend the human element with the machine element. Some of the visualization tools applied to the dataset include the use of statistical tools such as PCA, histograms, and scatter plots. PCA was used to reduce the dimensionality of the dataset while histograms and scatter plots were used to observe relationships between, and among, variables. Following the visualization step, the next logical step in the process is to get to the model phase. This path forward is in keeping with the structure presented by Wickham and Grolemund (2017).

a. R Programming Language

As discussed earlier, there are many AI/ML tools in use that can be applied to the dataset documented in this study. One such tool is the R programming language, a free open-source software used for data science, statistics, and visualization projects. R is a programming language and environment that finds applicability in statistical computing as well as graphics. The R programming language was built specifically for statistical analysis and data mining. The R programming language interfaces with RStudio a free, open-source integrated development environment (IDE) (Kent State University 2021). RStudio "includes a console, syntax-highlighting editor that supports direct code execution, as well as tools for plotting, history, debugging and workspace management" (RStudio 2021). The graphical user interface of R Studio shown in Figure 22, is organized to present the user with visuals of graphs, data tables, R code, and output in a single view. According to The R Foundation, the R programming language is a powerful and versatile tool that can be easily integrated with other platforms. The R programming language is a versatile tool that can be used to execute ML and statistics tasks such as linear and nonlinear modelling, classical statistical tests, time-series analysis, classification, and clustering (The R Foundation 2021). The R programming language was evaluated as a potential candidate for use in the model creation and analysis of the dataset.



Source Pane
Edit and run scripts (e.g. Rmarkdown templates), and view datasets

Tip: Run script

Environment Pane
Overview of objects (datasets, parameters, lists, etc.) you have imported or created.

Tip:
Start new script

Tip: Zoom and export plots

R Console Pane
R commands run are shown here, and non-graphic output and errors are displayed

Plots, Packages, and Help Pane
Commonly used to view graphics, install packages, and view help

```
---
title: "Cholera outbreak report"
output:
  word_document:
    keep_md: true
---

# Introduction to this template

This is a template which can be used to create an automated outbreak situation
report for cholera.

- It is organised by time, place and person.
- You can type normal text in white spaces (such as here) and r-code in grey
    spaces (denoted by three backticks and r) (see [Rmarkdown
    introduction](https://rmarkdown.rstudio.com/articles_intro.html) and
    [Markdown basics](https://rmarkdown.rstudio.com/authoring_basics.html))
- Introductions and contents of sections are within square brackets "[...]" and
    can be deleted as appropriate
```

```
>
> plot_age_pyramid(linelist_cleaned,
+                  age_group = "age_group",
+                  split_by = "sex",
+                  stack_by = "case_def") +
+   labs(y = "Cases (n)", x = "Age group (years)") + # change axis labels (nb. x/y flip)
+   theme(legend.position = "bottom",     # move legend to bottom
+         legend.title = element_blank(), # remove title
+         text = element_text(size = 18)  # change text size
+        )
+ ...
>
Error: attempt to use zero-length variable name
> |
```

```
linelist_clea…   300 obs. of 54 variables
linelist_dict    183 obs. of 11 variables
linelist_raw     300 obs. of 46 variables
population_da…   5 obs. of 3 variables
population_da…   4 obs. of 3 variables
population_da…   4 obs. of 3 variables
Values
  first_week     "2018-W01"
  LABS           chr [1:3] "cholera_culture_result" "ch…
  obs_end        2018-05-20
  obs_start      2018-01-01
```

Figure 22. R Studio: Source R4epis (2021).

Most of the user interaction takes place in the Source Pane where "R Scripts" can be created and edited. According to Philips (2019), the R Console pane is the heart of R and serves as the area where the R code gets evaluated. A user can type code directly into the console after the prompt and get an immediate response (Phillips 2019). The Environment Pane is customarily situated in the upper-right panel and is most often used to see brief summaries of objects in the R Environment in the current session, according to (R4epis 2021). The lower-right pane contains the Files, Plots, Packages, Help, and Viewer functions. The R programming incorporates several libraries containing prebuilt functions. Libraries for data science include:

- A caret package is a short form of Classification And Regression Training used for "miscellaneous functions for training and plotting classification and regression models" (Kuhn 2021).

- dplyr is the package which is used for data manipulation by providing different sets of verbs including select (), arrange (), filter (), summarize (), and mutate () (Wickham 2018).

- “ggplot2 is based on the ‘Grammar of Graphics,’ which is a popular data visualization library. Graphs with one variable, two variables, and three variables, along with both categorical and numerical data, can be built” Smith (2020).

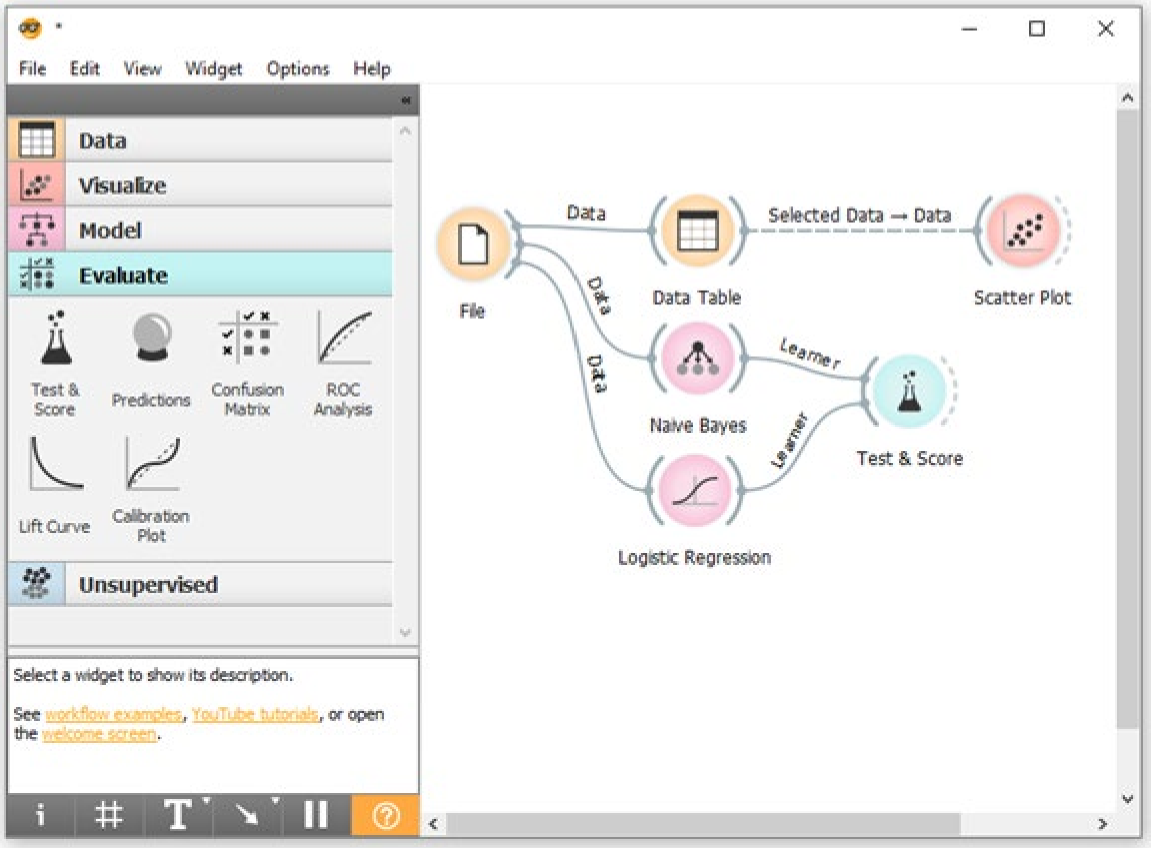
- MLmetrics is a collection of statistical evaluation metrics, including LogLoss, sensitivity, accuracy, RMSE, AUC, ROC score and utility functions, that measure regression, classification, and other performance measures Yan (2016).

- Recursive Partitioning and Regression Trees (rpart) “The rpart code builds classification or regression models of a very general structure using a two-stage procedure; the resulting models can be represented as binary trees.” Atkinson (2019).

b. Orange

Orange is an open-source data visualization and analysis tool that can be used for data mining through visual programming. The underlying concept of Orange is built on Python scripting. Orange provides interactive data analysis workflows with multiple data mining tools supporting both supervised and unsupervised ML approaches. A screenshot of the Orange application user interface is shown in Figure 23 where a subset of the tool available categories are displayed. The categories include Data, Visualize, Model, Evaluate, and Unsupervised. These categories contain tools, also known as widgets, that are selected by the user and placed in the work area on the right.

The Data category contains widgets used to enter data to the system. This is the area allows the user to perform data wrangling. Data wrangling is described as “the process of programmatically transforming data into a format that makes it easier to work with” (Talend SA 2021). The Visualize category is used to view properties of the data using tools such as tree viewer, box plot, violin plot, bar plot, scatter plot, line plot, linear projection, Venn diagram, and silhouette plot. The Model category contains algorithms such as calibrated learner, kNN, decision tree, random forest, gradient boosting, SVM, linear regression, logistic regression, naïve Bayes, neural network, and stochastic gradient descent. The Model category includes the five most popular and effective algorithms listed by Burkov (2019) in Chapter 3—Fundamental Algorithms of *The Hundred-Page Machine Learning Book*. The Evaluate category contains the tools that can be used to assess the quality of the model. This category includes test and score, predictions, confusion matrix, ROC analysis, lift curve, and calibration plot.



Select a widget to show its description.

See workflow examples, YouTube tutorials, or open the welcome screen.

Figure 23. Orange Overview: Source Demšar and Zupan (2013).

The Unsupervised category, as the name implies, contains the algorithms that are used to run unsupervised learning models. Algorithms include distance map, hierarchical clustering, k-Means clustering, DBSCAN, PCA, and manifold learning. Orange executes its tasks by a process called workflow. To start a workflow, widgets are dragged from the left panel and placed in the work area of the user interface. Some widgets have input and output connectors, while others only have input connectors. The connectors are used to transfer information between widgets; the output of one widget is connected to the input of another widget to facilitate the communication and data transfer. An input file is selected in the workspace to initiate the workflow. Orange can handle and process multiple input file types including tab-separated values, *.csv, and Microsoft Excel spreadsheets (*.xlsx). The dataset generated for this thesis was formatted as both **.csv*, and **.xlsx* format, ensuring that the files are readable by Orange.

4. Training/Testing of Model using Simulated Data

Both RStudio and Orange were used in the model development process. RStudio was used in the data wrangling stage of the development, as a first step, to analyze the dataset and extract useful features. Using some of the R programming tools presented earlier in the report, we generated plots of the dataset and capture useful statistics. The plot command was used to generate Figure 24, which shows the relationship between the GPS Time Tag and the estimated position error (EPE) parameter as captured in the ICD-153 message ID 5040. The EPE parameter is one of the metrics used to assess the performance of the FDI model. The GSG-62 GNSS simulator was used to generate the FDI dataset represented in the figure. Figure 24 shows a division in the dataset representing the signal levels before, during, and after the application of FDI to the system.

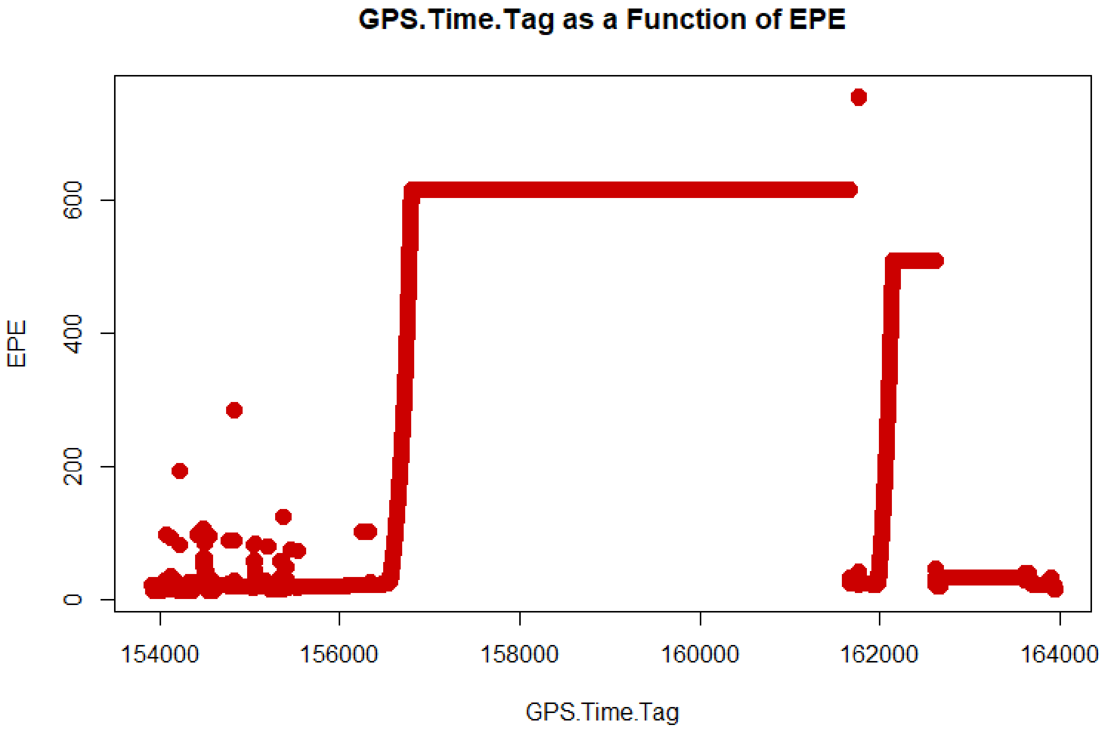


Figure 24. GPS Time Tag and EPE

Analysis of the dataset can be conducted as two separate portions. Figure 25 shows the leftmost portion of the dataset from Figure 24 representing the normal portions of the live sky data. The outliers in this portion of the data were investigated. The spike in EPE occurred at 12:00:14 increasing from 22.49 m to 283.56 m and receding to 27.69 m over 3 consecutive seconds. The EHE and EVE parameters showed significant increases over the same 3 second period as expected. The EPE represents the vector magnitude of EHE and EVE. The jammer-to-signal ratio (J/S) increased from 43 dB to 44 dB, otherwise, all other parameters in the dataset retained their normal state.

$$EPE = \sqrt{EHE^2 + EVE^2}$$

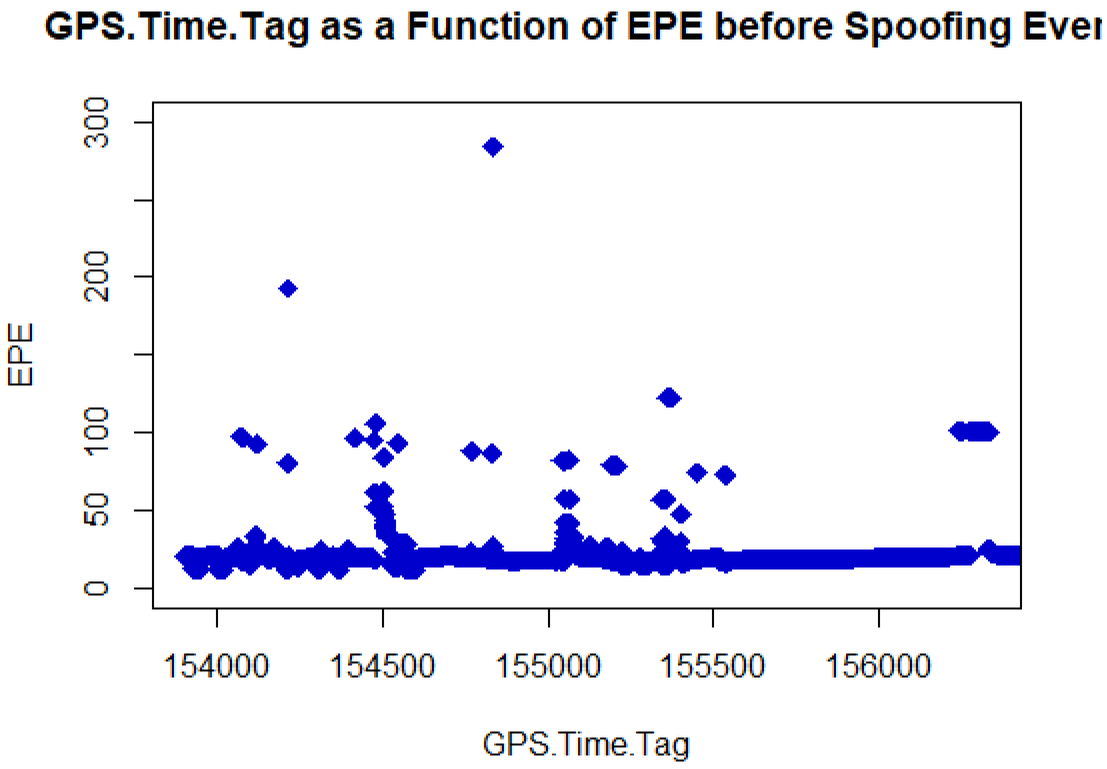


Figure 25. GPS Time and EPE before FDI Application

The second segment of the plot in Figure 24 represents the period in which FDI was being actively applied to the system. There was a gradual increase in the EPE value as the operator increased the transmit power on the GSG-62. The increasing power resulted in the EPE climbing to over 600 m during the active spoofing period. Figure 26 captures the Google Earth view of the result of the FDI application on the system. The scenario was structured to mimic a route driving around in a loop. The green cluster at the northern most portion of the plot represents the normal data (and EPE) captured in the previous figures. With FDI applied to the system and the power levels gradually increasing over time, resulting in the EPE increasing with to the end of the green line shown in Figure 26. The red line on the plot was a line added to measure the distance of the EPE.

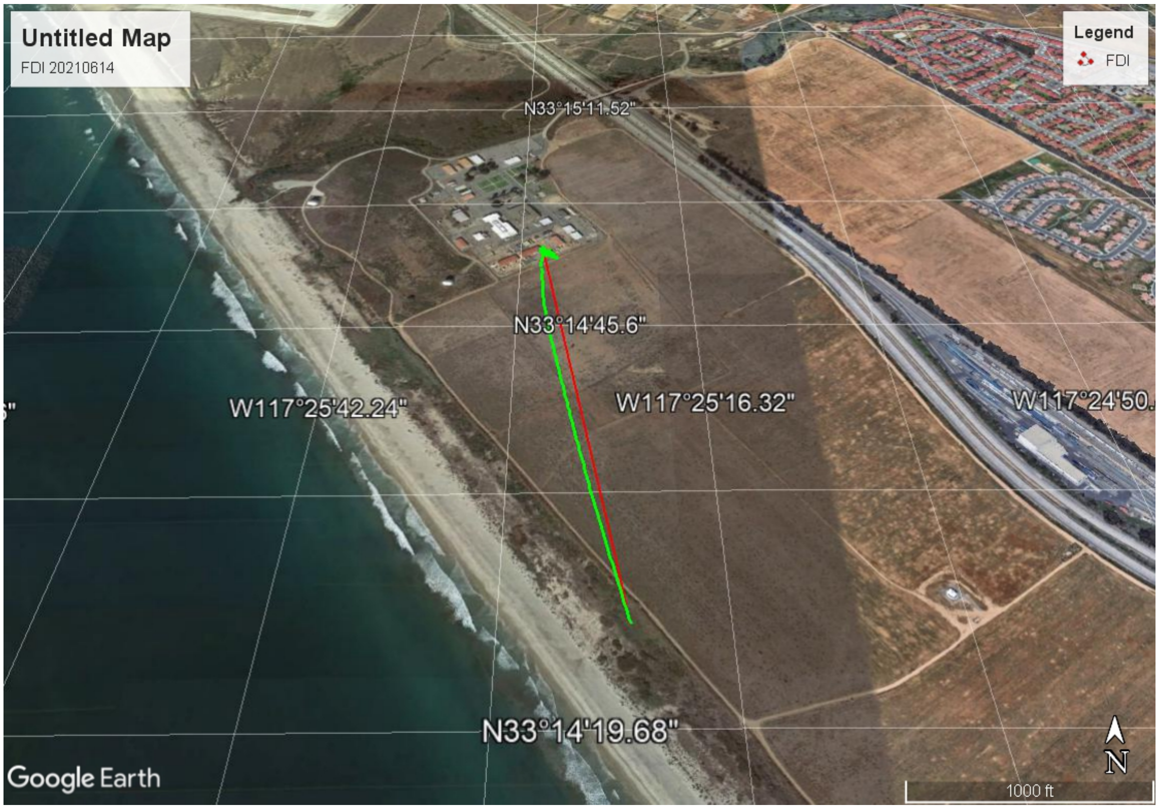


Figure 26. Google Earth View

The dataset used in this study was stored in tables consisting of multiple rows and column data. Entries in the tables were a mix of categorical variables, logical, numeric and integer types. To perform functions in RStudio, the libraries were installed to carry out the functions needed. For instance, to install the *ggplot2* package, type *"library(ggplot2)"* in the R console pane and press return. This principle was followed for the installation of the packages used in the analysis. To get specific measures of the dataset, the appropriate commands would be run. The *names* command, for instance, returns the name of columns in the dataset, while the *dim* command returns the number of rows and columns. The dataset represented in the plot was created from the file *"2021-06-14_1845_unkeyed DAGR_GSG_5040.prn"* that was captured on June 21, 2021. The **.prn* file is a text file generated by the CTS-153 software application which was imported to a **.csv* file format. Executing the (dim) command in RStudio confirms that the file contains 9969 rows and 131 columns of data. Table 3 lists the EPE statistic which ranges between 11.82 m and 757.19 m. The table shows the variation for EPE parameter with FDI applied to the system.

The values in the table were obtained from the results generated from processing the data using the RStudio using the *summary(data)* command. Further analysis of this dataset will be expanded upon in Chapter IV of the report.

Table 3. Simulated Data Statistics

Unkeyed DAGR Parameters					
Min (m)	1st Quartile	Median	Mean	3rd Quartile	Max (m)
11.82	21.37	509.18	344.71	615.66	757.19

Application of FDI on the system was undertaken to pull the end user clandestinely from the preplanned route and possibly their destination. Within the messages used by the DAGR, there are requirements, at the bit level, that need to be set to provide a valid navigation solution. In the ICD 153 message ID 5040, both the *Nav Converged* and the *Vel Valid* bits need to be set to 1. The ICD-153 standard dictates that Nav Converged be set to “1” when the DAGR has achieved both a valid PVT solution and is also providing navigation to the user. The *Status* bit in the RMC NMEA 0183 message should also be set to “A = Data valid” during this time. The discussion of FDI and its impact on the system were centered around these conditions being met. A sample of the distribution of the *Vel Valid* bit in one ICD 153 message ID 5040 is shown in Figure 27. There are 5640 instances of TRUE for this dataset from 9969 rows of data. A similar plot was generated for the *Nav Converged* bit, where there were 3738 instances of TRUE condition.

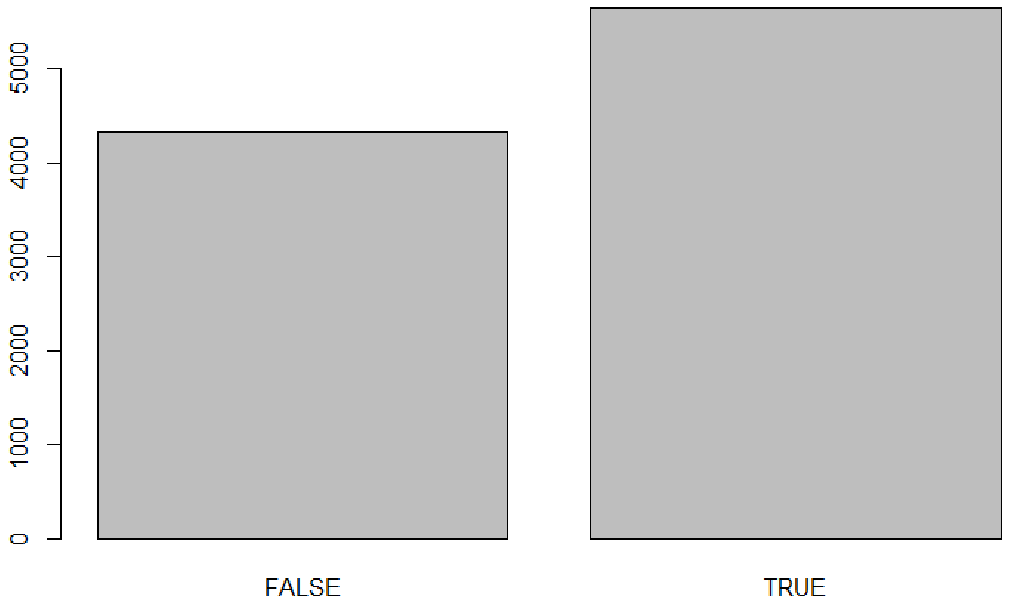
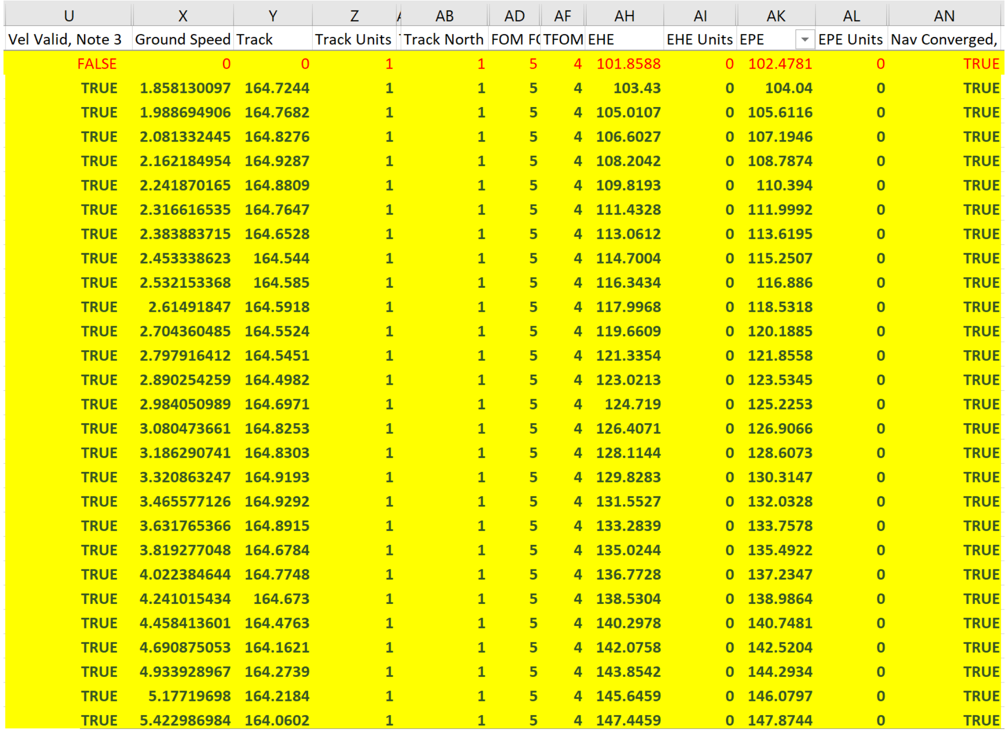


Figure 27. Distribution of Vel Valid Bit

From the previous discussion, not all instances of *Nav Converged* and *Vel Valid* occur together in the dataset. The first instance of the alignment of these two parameters occurred at 12:30:01 and ended 12:32:57. During this time, the *Ground Speed* parameter changed from 0 kilometers per hour (kph) at 12:30:00 to 1.86 kph at 12:30:01 and gradually increased to 30.99 kph at 12:32:57. A sample of the dataset representing the first instance of the convergence of *Nav Converged* and *Vel Valid* occur together is shown in Figure 28. The later portion of the dataset representing the instance when the *Nav Converged* parameter changed to FALSE is not shown here. At the point where the *Nav Converged* parameter changed to FALSE; the navigation solution was no longer valid. The *Vel Valid* parameter maintained the TRUE state until 13:54:02 before it too changed state to FALSE. The *Ground Speed* parameter maintained the 30.99 kph value for the period where the *Vel Valid* parameter maintained the TRUE state. Based on the definition of data mining, the data analysis being discussed here qualifies as such. Han and Kamber (2006, 5) describe data mining as the process of "extracting knowledge from large amounts of data." This process also fits the description of knowledge discovery. The bulk of the preliminary data analysis and data mining tasks were carried out using RStudio. This analysis also served as a reference point in the assessment of the model parameters discussed in Chapter IV.



U	X	Y	Z	AB	AD	AF	AH	AI	AK	AL	AN
Vel Valid, Note 3	Ground Speed	Track	Track Units	Track North	FOM F(	TFOM	EHE	EHE Units	EPE	EPE Units	Nav Converged,
FALSE	0	0	1	1	5	4	101.8588	0	102.4781	0	TRUE
TRUE	1.858130097	164.7244	1	1	5	4	103.43	0	104.04	0	TRUE
TRUE	1.988694906	164.7682	1	1	5	4	105.0107	0	105.6116	0	TRUE
TRUE	2.081332445	164.8276	1	1	5	4	106.6027	0	107.1946	0	TRUE
TRUE	2.162184954	164.9287	1	1	5	4	108.2042	0	108.7874	0	TRUE
TRUE	2.241870165	164.8809	1	1	5	4	109.8193	0	110.394	0	TRUE
TRUE	2.316616535	164.7647	1	1	5	4	111.4328	0	111.9992	0	TRUE
TRUE	2.383883715	164.6528	1	1	5	4	113.0612	0	113.6195	0	TRUE
TRUE	2.453338623	164.544	1	1	5	4	114.7004	0	115.2507	0	TRUE
TRUE	2.532153368	164.585	1	1	5	4	116.3434	0	116.886	0	TRUE
TRUE	2.61491847	164.5918	1	1	5	4	117.9968	0	118.5318	0	TRUE
TRUE	2.704360485	164.5524	1	1	5	4	119.6609	0	120.1885	0	TRUE
TRUE	2.797916412	164.5451	1	1	5	4	121.3354	0	121.8558	0	TRUE
TRUE	2.890254259	164.4982	1	1	5	4	123.0213	0	123.5345	0	TRUE
TRUE	2.984050989	164.6971	1	1	5	4	124.719	0	125.2253	0	TRUE
TRUE	3.080473661	164.8253	1	1	5	4	126.4071	0	126.9066	0	TRUE
TRUE	3.186290741	164.8303	1	1	5	4	128.1144	0	128.6073	0	TRUE
TRUE	3.320863247	164.9193	1	1	5	4	129.8283	0	130.3147	0	TRUE
TRUE	3.465577126	164.9292	1	1	5	4	131.5527	0	132.0328	0	TRUE
TRUE	3.631765366	164.8915	1	1	5	4	133.2839	0	133.7578	0	TRUE
TRUE	3.819277048	164.6784	1	1	5	4	135.0244	0	135.4922	0	TRUE
TRUE	4.022384644	164.7748	1	1	5	4	136.7728	0	137.2347	0	TRUE
TRUE	4.241015434	164.673	1	1	5	4	138.5304	0	138.9864	0	TRUE
TRUE	4.458413601	164.4763	1	1	5	4	140.2978	0	140.7481	0	TRUE
TRUE	4.690875053	164.1621	1	1	5	4	142.0758	0	142.5204	0	TRUE
TRUE	4.933928967	164.2739	1	1	5	4	143.8542	0	144.2934	0	TRUE
TRUE	5.17719698	164.2184	1	1	5	4	145.6459	0	146.0797	0	TRUE
TRUE	5.422986984	164.0602	1	1	5	4	147.4459	0	147.8744	0	TRUE

Figure 28. Sample FDI Dataset

The simulated dataset was combined with the operational dataset captured through the live sky feed to serve as the training dataset for Orange. The file type used in RStudio was also used as the input to Orange to evaluate the model candidates. The algorithm selection, training, evaluation with the test set, and parameter tuning steps as described by Kotsiantis (2007), was applied to the dataset. Unsupervised learning was applied at this stage of the process. Clustering is one of the unsupervised learning algorithms and finds applications in situations where the objects within a dataset contain no known labels. Recall that a cluster groups data items that exhibit similar properties together. Using this concept, data items with dissimilar properties will be placed in other clusters making clustering an ideal tool for outlier detection. Han and Kamber (2006) lists portioning methods, hierarchical methods, and density-based methods as three of the major clustering methods in use. The authors also list *k*-means, *k*-medoids, and their variations, as the most used partition methods (Han and Kamber 2006, 402). According to Witten et al. (2016, 45), "the success of clustering is often measured subjectively in terms of how useful the result appears to be to a human user." Clustering tools have the added benefit of generating great visual representation of the data. Evaluation of the model involved the application of PCA and *k*-means clustering algorithms run on the simulated data from one of the scenarios created to add FDI to the system.

In the text, *Data Mining: Practical Machine Learning Tools and Techniques*, Witten et al., (2016, 93) emphasizes the value of simplicity. The authors believe that "because the structure underlying many real-world datasets is quite rudimentary" simple rules are adequate for achieving the desired results. The simplicity concept was applied to the model development process and was followed throughout the process.

5. Training/Testing of Model using Operational Data

Operational data was captured from the system configured to receive only the live sky data feed. Figure 29 shows a plot of the GPS Time Tag vs EPE for operational message traffic from the dataset. The file *"2021-06-14_Unkeyed DAGR_5040_GSG.prn"* contains the dataset captured over a 36-hour period. During this time, the system was run in a normal operational mode; without FDI applied to the system. The contrast between the data in Figure 24 can be seen in the EPE values in the plot. The dataset displays periodicity representing the movement of the satellites in MEO. The outliers shown in Figure 29 were investigated by examining the dataset and isolating the period in question.

Table 4 shows the statistics of the EPE parameter from the dataset. The EPE ranges between 9.55 m and the maximum of 55.27 m and represents the expected values for this parameter without FDI applied to the system. The values in the table were also obtained from the results generated from processing the data in RStudio using the *summary(data)* command. Discussion of the dataset and the analysis performed will be documented in Chapter IV.

Table 4. Operational Data Statistics

Unkeyed DAGR Parameters					
Min (m)	1st Quartile	Median	Mean	3rd Quartile	Max (m)
9.55	11.48	12.44	13.26	13.85	55.27

The live sky data was aggregated with data from generated from the GSG-62 and applied to the model as a training data. This method ensures that the model is trained with a representative sample of both types of data.

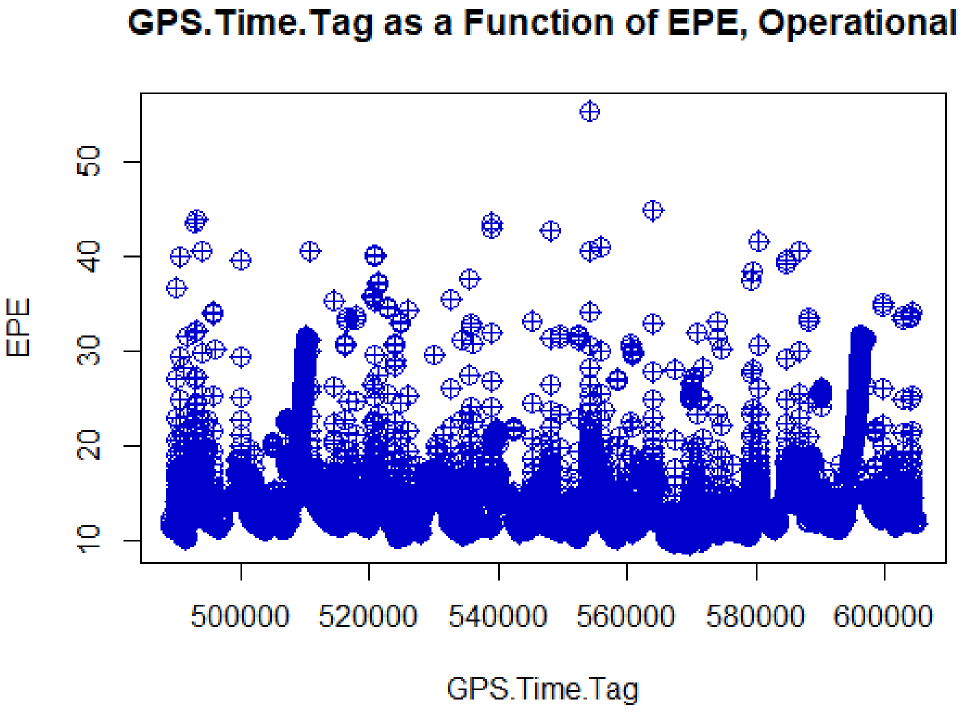


Figure 29. GPS Time Tag and EPE

Figure 30 shows the maximum value from the dataset over the period 02:54:00 to 02:56:59 on May 29, 2021. The spike occurred over a three second period, 02:55:27 to 02:55:29. The EPE increased from 17.11 m at 02:55:26 to 55.27 m, before declining 40.45

m at 02:55:30. Other parameters within message ID 5040, including PRN Code Tracked, Tracking State, Chan Status A/B Valid, jammer-to-signal ratio (J/S), and carrier-to-noise ratio (C/No) were all examined and found to be within the range of values for the period preceding and immediately following. The fact that the sharp rise and correspondingly sharp decline occurred within three seconds assures us that, indeed, this was not a FDI event. As shown previously, a FDI event persists in the system for a much longer than the time shown here. This spike event can be attributed to environmental factors, and not a FDI attempt.

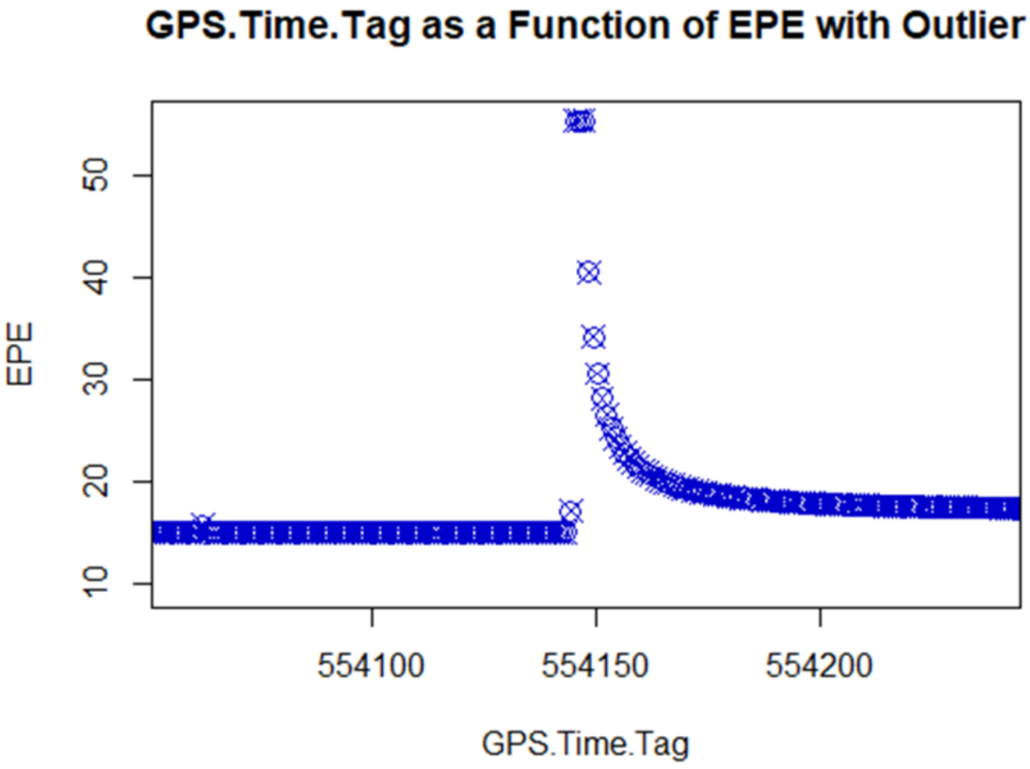


Figure 30. GPS Time Tag and EPE with Outlier

The dataset from which the figures were obtained, was used as an input to the model to validate the inferences gained from the analysis done using RStudio and the application of statistical tools. Orange was recommended as the tool to execute this stage of the FDI detection process. The same operational data that was used in the analysis described here, was used in Orange to assess the model performance. The model process employed by Orange provides more than ten model types. The approach used in this stage of the development focused on applying multiple models from the available options and evaluate the scores to determine the best. Evaluating the performance of the models is a fundamental step in the model development process. It is imperative to find the model that best represents the data as this effort forecast how effectively the model will work. This roughly follows the process described by Kotsiantis (2007, 250) in Figure 20.

E. SUMMARY

This chapter presented the taxonomy of the GPS spoofing along with a summary of the effectiveness and complexity of each method. The systems engineering process was followed to conceive and develop the system through the MBSE construct. The concept of operations applied to the system exposed some of the vulnerabilities inherent in the GPS structure. The tools that were applied to the results to show the nature of the GPS spoofing problem and the effectiveness of the toolset that will aid in ameliorating the FDI threat detection.

IV. ANALYSIS RESULTS

This chapter provides an analysis of the findings of the study, expanding on the preliminary results presented in the previous chapter. The analysis was focused on the message traffic from both an operational concept as well as message traffic generated by a GNSS simulator. Both unsupervised and supervised learning model types were explored and applied to the datasets to assess performance measures. The results are presented with plots to visualize the dataset and the Orange dataflow used to generate the results.

A. BENEFITS AND SHORTFALLS OF DATA TYPES

The engineered system used for this study was set up to capture data traffic that conforms to commercial and military standards. Gakstatter (2015) describes NMEA as "a standard data format supported by all GPS manufacturers, much like ASCII is the standard for digital computer characters in the computer world." This standard allows manufacturers in diverse operational areas to develop and manufacture commercial systems that allows interoperability across platforms and companies. The NMEA 0183 standard forms the basis of the message structure applied to most commercial GPS user devices. The ICD-153 standard is applied to the development for application to military navigation systems. Military navigation systems are compliant with the NMEA 0183 standards while also implementing the DOD specific message structure. While both the NMEA 0183 and the ICD-153 messages contain the same navigation information, this information is distributed across multiple messages in the NMEA 0183 messages. The ICD-153 messages contain a more comprehensive list of attributes using fewer message types than NMEA 0183. This difference allowed for the verification of message information by consulting both message types for concurrence and consistency.

The dataset for this thesis consisted of messages conforming to both the NMEA 0183 standard and the ICD-153 standard. NMEA messages are sent from talker to listener. The NMEA 0183 standard defines a talker as any device that sends data to other devices within this standard, while a listener is any device that receives data from another device (NMEA 2012). NMEA 0183 messages contain a two-letter prefix that represents the system on which they are being used. The U.S. system affixes GP to the three letter codes listed. Seven NMEA 0183 messages were monitored for this study:

- GPGGA—GPS fix data message
- GPGSA—GNSS DOP and active satellites message
- GPGSV—GNSS satellites in view message
- GPRMC—recommended minimum specific GNSS data
- GPZDA—time and date message
- GPZLZ—time of day message

The ICD-153 messages consist of data messages and command messages. are sent. Data messages are used to transmit from intelligent terminals (INTER) and receivers (RCVR). Command messages facilitate the request of information by the INTER from the RCVR. Message ID 253—buffer box status request message is sent automatically every 6 seconds in compliance with the standard. Likewise, message ID 5044—warning message is generated to capture and report detected errors. Message ID 9—receiver capabilities message, is a command message that returns system information such as hardware and software version from the DAGR. Four ICD-153 messages were actively monitored for this study:

- Message ID 3—time mark data message
- Message ID 4—24-channel time mark data message
- Message ID 5040—current status message
- Message ID 5042—extended status message

The files *"2021-06-14_1845_unkeyed DAGR_GSG_5040"* and *"2021-06-14_11-41-26_DAGR_NMEA_Data"* both contain the message traffic that captured the scenario that was analyzed in Chapter III. In all the simulated scenarios used in this experiment, at the start of the application of FDI, the message ID 5040 showed the Vel Valid and Nav

Converged parameters both being set to TRUE. The equivalent validation in the NMEA 0183 message was indicated by parameters within the GPRMC message. With the *Status* bit set to *A = Data valid* and the Mode Indicator bit was set to *A = Autonomous*, a valid navigation was being provided to the user. Data valid is also expressed in message ID 4 which contains the *Nav Data Not Valid* parameter. When set to 0, the navigation solution is valid. The parameters were crosschecked across the three messages to ensure alignment, thereby verifying the validity of the solution. *Ground Speed* from the ICD 153 message ID 5040 is represented by *speed over ground* in units of knots in the GPRMC message. Applying the conversion factor of 1.852 translates knots to kph. The equivalency was verified by multiplying the speed value from the GPRMC message. The parameter *Track* in message ID 5040 and *course over ground* in the GPRMC message both represent the direction of travel in degrees. The combination of the GPZDA and GPZLZ messages capture the time and date as represented in message ID 5040.

Message ID 5040 also captures the satellites in the constellation that are available through the *Number of Visible SVs* and *Number of Healthy SVs* parameters. The GPGSV message shows the number of SVs in view, SV PRN, as well as the elevation, azimuth, and SNR value of the SVs. The SNR in the GPGSV message is represented by the C/No parameter in message ID 5040. The GPGSA message displays the number of SVs in the active solution along with operating characteristics of said SVs. Two *Mode* parameters are included in the GPGSA message. In normal operations, it is customary to see one parameter set to A indicating that the system is allowed to switch automatically between a 2D and 3D fix. The second *Mode* parameter is enumerated with 1 representing “Fix not available,” 2 representing 2D fix, and 3 representing 3D fix. The PDOP, HDOP and VDOP values are also shown in the GPGSA message. Message ID 5040 shows the GDOP parameter which is equivalent to the PDOP measure according to the research. GIS Geography (2021) confirms the relatonship between GDOP and PDOP with the statement that “GDOP or PDOP describes the error caused by the relative position of the GPS satellites.” Current DOP value is captured in message ID 4. Current DOP is “calculated using only SVs used in the PVT solution, represented as a scaled integer having a value from 0.0 to 6553.5, where 0 means DOP < 0.1 and 6553.5 means DOP ≥ 6553.5.” GPS Wing (2007, B-29).

The GPGGA message captures values related to the "time, position, and fix for a GPS receiver" (National Marine Electronics Association 2012). A GPS *quality indicator* of 1 represents a valid fix in the GPS SPS Mode in the GPGGA. The equivalent of the quality indicator is found in the *Operating Mode* parameter in message ID 5040. This discussion helps to verify that both ICD 153 and NMEA 0183 messages contain the same information, only that they are not all represented in the same way or in the same message.

B. RESULTS OF MACHINE LEARNING EXPERIMENTATION

This analysis created and experimented with unsupervised and supervised machine learning models both in RStudio and Orange using the hybrid live stream/simulated datasets with FDI applied to the system.

1. Unsupervised Learning Model

This experiment demonstrated the ability of an unsupervised ML algorithm to cluster data according to similar characteristics. In the experiment, the ML algorithm correctly grouped the live stream/simulated data into four clusters: one representing the real GPS data, two representing two separate GPS spoofing attempts, and the fourth representing the spoofed state. The unsupervised model identified the portion of the dataset representing the onset of spoofing. This relationship was captured by the *Ground Speed* and *Track* parameters that registered non-zero values at the onset of the spoofing activity. The application of FDI on the engineered resulted in a change of the Ground Speed parameter from 0 and to 30.99 kph. This is another significant finding as it once again demonstrated the utility of the algorithms being applied to the dataset. Although spoofing was being actively applied to the system for upwards of 60 minutes in each scenario, the data shows that active spoofing triggers a change in position in as little as 40 seconds. From that point forward, the system remains in a spoofed state at the newly established position. Parameters such as J/S and C/No in message ID 5040 did not reflect the status change from the "normal state" to the "spoofing detected" state; however, the clustering algorithm detected spoofing activity. Further inspection and evaluation of the rate of change and second derivatives did not clearly identify a relationship with the spoofing activities. These findings merit further research and should be pursued as part of the future work associated with the current study.

Datasets generated from the engineered system are unlabeled and therefore do not contain any labeled parameter that defines spoofing. Armed with that information, we concluded that an unsupervised learning algorithm is an appropriate choice. Witten et al., (2016, 45) in discussing clustering, states "When there is no specified class, clustering is used to group items that seem to fall naturally together." Clustering is one of the tools in the unsupervised learning model paradigm. Its use allowed us to capture characteristics of the dataset. Using this idea as a starting point we applied the *k*-means algorithm in keeping with the belief that *k*-means is one of "the most well-known and commonly used partition methods" (Han and Kamber 2006, 402). The Orange workflow shown in Figure 31 represents the unsupervised learning model used to conduct the clustering operation on the dataset. The *k*-means clustering method is a centroid-based technique, where the similarity measure is based on the mean value of the other members of the cluster. Clustering methods such as DBSCAN are regarded as density-based algorithms because the measure is based on a density factor and not a mean value as in *k*-means. Lecture notes state that "DBSCAN focuses on clustering based on density: cluster observations in high density regions" (Michael Atkinson, lecture notes, May 5, 2021).

The unsupervised learning workflow also contains Manifold learning, and PCA algorithms. The PCA results from the model resulted in two principal components with PC1 showing a value of 0.936. This value means that 93.6% of the variance of our dataset can be explained by PC1. The results of PCA were examined and showed the relationship between the *North Ref Correction* with *EPE* and its vector components. This was another previously unidentified relationship. The DBSCAN model showed a Euclidean distance of 0.010 representing the neighborhood distance. The cluster from DBSCAN identified the same characteristics as the *k*-means results. The Manifold Learning results were also consistent with the others within the workflow. The model results from the other methods further amplified the results from the *k*-means clustering algorithm in identifying the clusters created by the application of FDI. Applying multiple models to the dataset provided the same answer to the question; can the AI model detect cyber attacks on the military GPS? The Select Columns widget in Figure 31 was used to segregate the variables within the workflow. Obvious variables such as GPS Time Tag and Ground Speed were put in the Metas section of the widget, thereby isolating those parameters from the solution, yet leaving all other parameters in play. Ignoring the 18 data columns designated as "*Spare*" the model found and used all other 100+ parameters to determine the relationships that identified the spoofing threat.

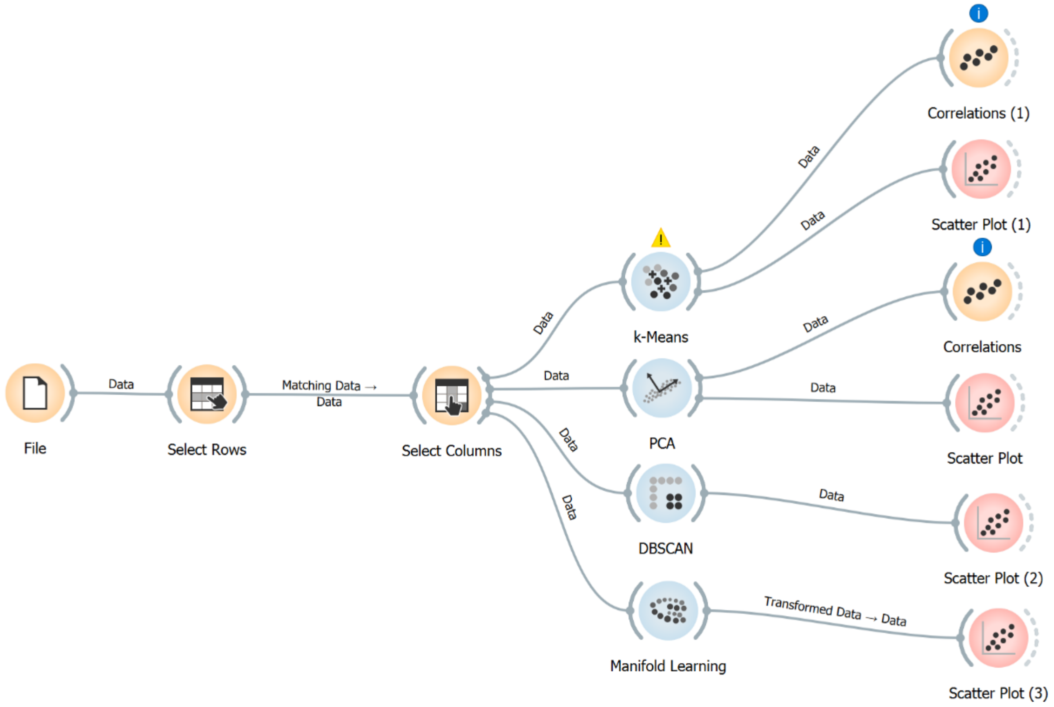


Figure 31. Orange Workflow

Running the *k*-means clustering algorithm with *k=4*, identified the four clusters C1, C2, C3, and C4 shown in Figure 32 This plot exhibits a similar structure to the RStudio plot discussed in Chapter III (Figure 24) for the same dataset. We experimented with the *k* values from three to seven. In every instance, the clusters were represented as distinct clusters. During the period of actively applying FDI to the system (both *Vel Valid* and *Nav Converged* parameters = *TRUE*), the EPE recorded a gradual increase. The conditions for a valid solution were sustained for a period of 176 seconds (12:30:01 – 12:32:57). The four identified clusters are in keeping with the analysis and discussion in Chapter III. The normal data flow discussed in Chapter III is represented by cluster C1 as shown in Figure 32. Cluster C4, shown in gold represents the two active spoofing portions of the exercise. The EPE experienced a gradual increase from 103 meters climbing to 610 meters during this period. Correspondingly, the *Ground Speed* and *Track* parameters started to register non-zero values starting at 1.85 kilometers per hour (kph) up to 30.99 kph. The change in ground speed was an unexpected, but nonetheless important finding. This observation gave us the indication that active interference was occurring on the engineered system.

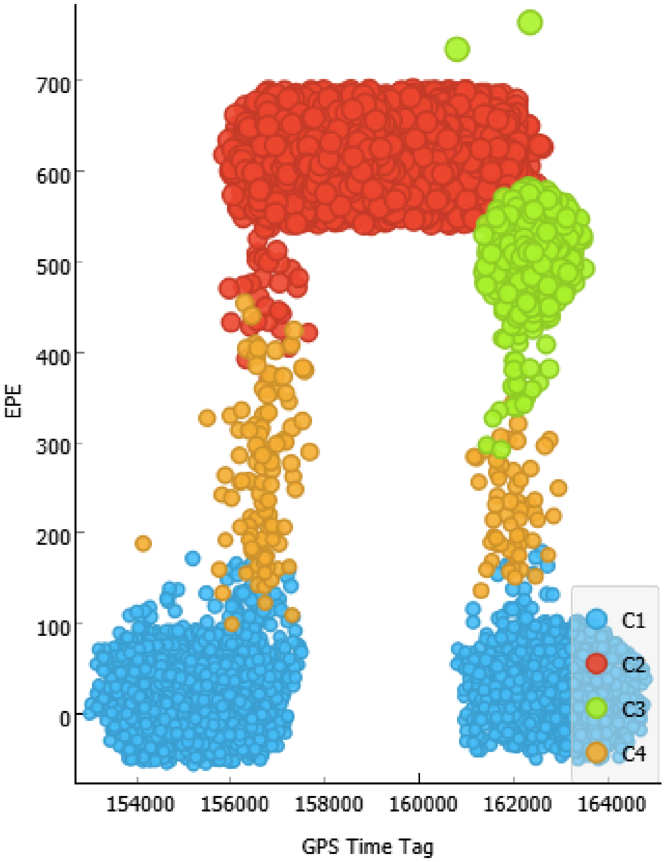
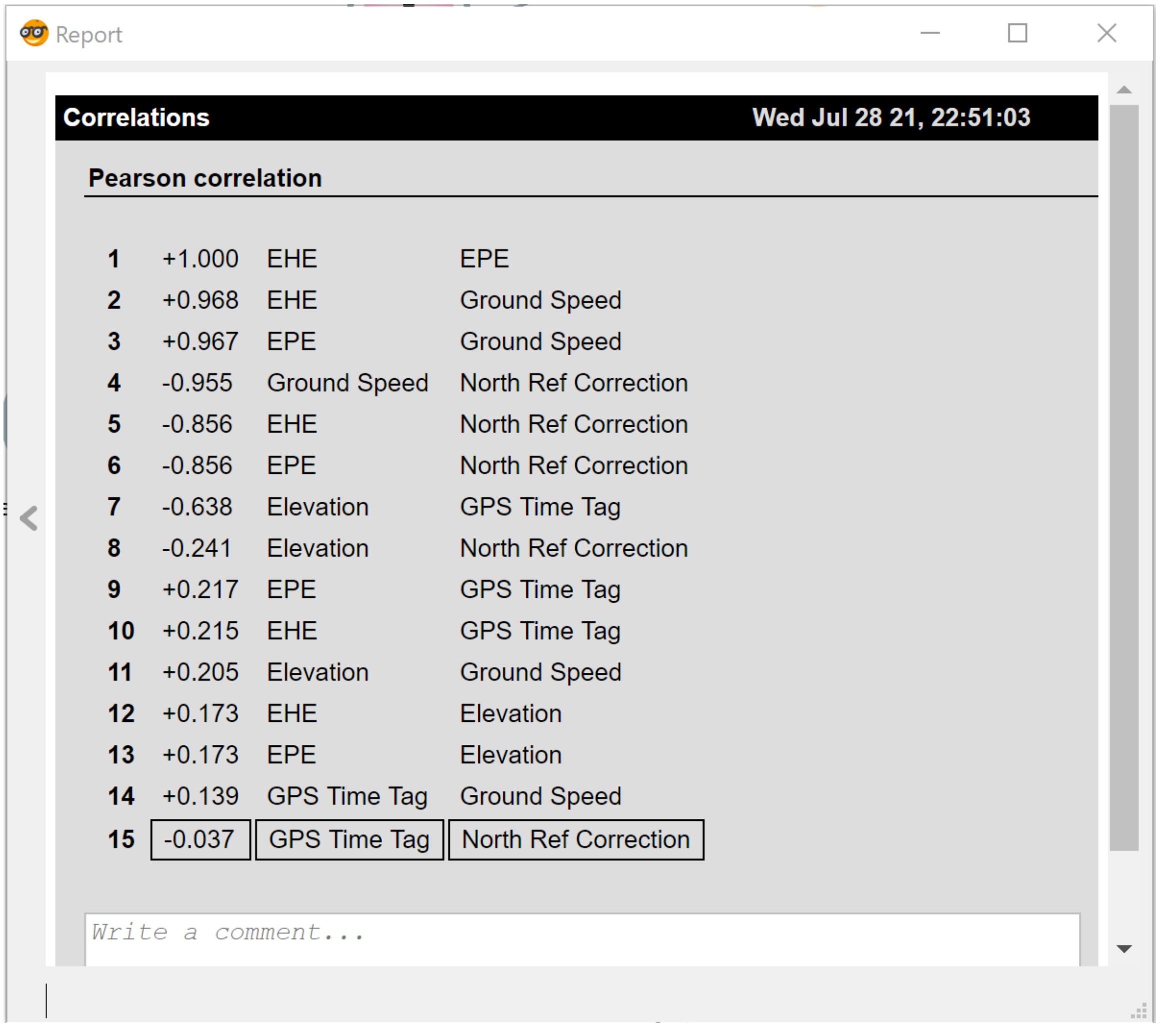


Figure 32. K-means Clustering

The highest EPE recorded in C4 was 409.8 meters at 12:32:09 which, by then the engineered system was displaying characteristics of being spoofed. Cluster C2, in red, represents the ‘end state’ where the system is already spoofed by the application of FDI.

The data represented by C2 could also be considered as normal data to the naked eye as it displays the characteristics of EPE discussed in Chapter III. The difference between C1 and C2 is the distance measure of EPE. This finding is the result of a change in position brought on by active spoofing. Cluster C3 in green represents a second active FDI event that lasted 113 seconds from 14:00:10 to 14:02:03. The EPE reached a maximum of 509 m during this episode as can be gathered from the figure. The two high measures in C3 represent EPE values of 753 meters and 757 meters; both spikes occurred within two seconds of each other. The navigation solution was not valid at this point as only the *Nav Converged* parameter was set to *TRUE*. This observation is quite similar to the discussion of the anomalous data points in the operational dataset discussed in Chapter III and Figure 30. Nothing in the dataset gave us any clear indication of what triggered these two spikes. All other variables in the dataset remained in their normal state during this period. Even though the conditions for a valid solution ended at 12:32:57, the EPE of 615 m remained until 13:54:02 resulting in the extended breadth of C2. Cluster C2 and C3 represent the engineered system in a spoofed state. In trying to identify relationships between and among the parameters in the datasets, a correlation method was applied. The Pearson correlation coefficient was chosen for this task. The Pearson correlation coefficient measures the linear dependence between two variables and is shown in Figure 33 for the predictors. The correlations shown in Figure 33 confirm that Ground Speed and North Ref Correction strongly impacts EPE and its vector components EHE and EVE. Correlation values above 0.7 cover the range of high to very high correlation. Choosing EPE as the target yielded great dividends as this concept was then applied to a supervised learning tree. Discussion of the supervised concepts and applications are presented in the Supervised Learning Model section of this report.



Correlations Wed Jul 28 21, 22:51:03

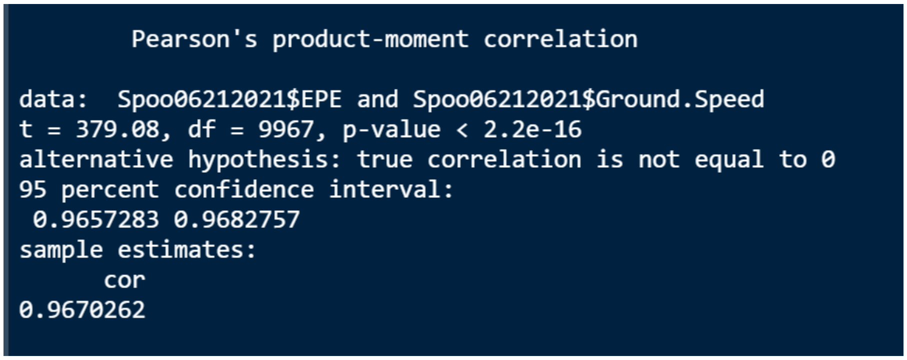
Pearson correlation

1	+1.000	EHE	EPE
2	+0.968	EHE	Ground Speed
3	+0.967	EPE	Ground Speed
4	-0.955	Ground Speed	North Ref Correction
5	-0.856	EHE	North Ref Correction
6	-0.856	EPE	North Ref Correction
7	-0.638	Elevation	GPS Time Tag
8	-0.241	Elevation	North Ref Correction
9	+0.217	EPE	GPS Time Tag
10	+0.215	EHE	GPS Time Tag
11	+0.205	Elevation	Ground Speed
12	+0.173	EHE	Elevation
13	+0.173	EPE	Elevation
14	+0.139	GPS Time Tag	Ground Speed
15	-0.037	GPS Time Tag	North Ref Correction

Figure 33. Pearson Correlations

Although GPS Time Tag and EPE were chosen as the two parameters to show on a plot, they were just two of the many potential combinations of data parameters that could be selected from the dataset. The selection was made based on the interesting characteristics observed during the data wrangling phase. Other combinations were examined and shown to give an equally vivid representation of the characteristics of the spoofing threat. The relationship between *EPE* and *Ground Speed* is shown from the RStudio window in Figure 34 and at 95% confidence interval, shows a very high correlation of 0.9670262. Correlation coefficients range from –1 to 1 where a value of 1.0 represents a positive relationship between the two variables. For the variables *EPE* and

Ground Speed, a positive increase in the *EPE* variable, results in a positive increase in the *Ground Speed* variable



```
        Pearson's product-moment correlation

data:  Spoo06212021$EPE and Spoo06212021$Ground.Speed
t = 379.08, df = 9967, p-value < 2.2e-16
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 0.9657283 0.9682757
sample estimates:
      cor
0.9670262
```

Figure 34. Pearson Correlation of *Ground Speed* and *EPE*

From statistics, recall than a p-value ≤ 0.05 is considered statistically significant. A low p-value "indicates strong evidence against the null hypothesis, as there is less than a 5% probability the null is correct" (McLeod 2019). Therefore, we reject the null hypothesis, and accept the alternative hypothesis. According to McLeod (2019), "the null hypothesis states that there is no relationship between the two variables being studied." We have demonstrated that increases in *EPE* results in increases in the *Ground Speed* parameter, thereby rejecting the null hypothesis. The dataset from the other simulated spoofing scenarios run during the experiment were also evaluated through the unsupervised learning model. Overall, the characteristics of the other scenarios and results were just like the scenario discussed in this chapter. No additional insights were gained from these datasets, so the results and analysis discussed for *k*-means equally applies to those datasets.

2. Supervised Learning Model

The unsupervised learning exercise provided additional insights to the dataset and created the desire to do more exploration of the data. The data wrangling process was once again initiated, this time to create a labelled dataset with the goal of training a supervised model. Two approaches were taken in this endeavor. First, the unsupervised model was used to identify clusters that served as the input dataset to train the supervised model. The unlabeled test set was used to check the trained model and check the performance metrics to see how well the test set fit in the clusters. The second approach used a form of association rules to create a labelled column in the training dataset to serve as the supervised set.

As a visual tool, the unsupervised learning algorithm proved their worth in the identification of characteristics of the datasets. This method aided in determining the binary classification of the data by identifying the both the spoofing and spoofed state of the system. Clustering algorithms have no tangible numerical measures in identifying spoofing. Enter the supervised learning algorithms which have several performance measures that can be applied to regression or classification problems. Spoofing detection falls in the category of binary or binomial classification problems. Burkov (2019, 19) states that "in a classification problem, a label is a member of a finite set of classes." The spoofing detection problem is a binomial classification used to identify "spoofing" and "not spoofing" on the dataset. The supervised learning algorithms applied to the datasets gives confirmation of the procedures and results of the unsupervised algorithms.

Recall that the message ID 5040 dictates that both the *Vel Valid* and *Nav Converged* parameters must be in a *TRUE* state. Using this knowledge, we used an *AND* function in Excel *(=AND(U2,AN2)*) and created a new column labelled '*Spoofing'.* For reference, in the Excel spreadsheet column *"U"* represents *Vel Valid* and column *"AN"* represents *Nav Converged* values. The condition where both *Vel Valid* and *Nav Converged* parameters are TRUE populates a value of TRUE in the Spoofing column. Conversely, in the case where either of Vel Valid or Nav Converged parameter is FALSE, a value of FALSE is populated in the Spoofing column. Witten et al., (2016, 11) describe association rules as "rules that strongly associate different attribute values." The dataset was first run through RStudio to create a classification and regression tree (CART) model. The *'dim'* command in RStudio indicated 4951 rows and 132 columns in the dataset. Querying the dataset with the '*summary'* command in RStudio revealed that the Spoofing column was of data type "Logical" and contained 4774 FALSE and 177 TRUE values. After successfully developing the CART model in RStudio, the workflow was created in Orange to verify the methods. Figure 35 shows the CART model using Orange. The CART model correctly identified the labelled data in both RStudio and Orange tools. The CART algorithm poses a series of questions, and the answers result in either a terminal node or another question. CART models "are visually appealing, are computationally fast, and automatically includes interactions and non-linear relationships" (Michael Atkinson, lecture notes, May 5, 2021). Figure 35 shows that for Ground Speed > 2.43927, the model is 100% accurate. For the implementation of the supervised learning tree, the select columns widget was once again applied with the Spoofing variable as the target. Metas consisted of GPS Time Tag, Nav Converged, and Vel Valid. The input features/variables included EPE, Current DOP, North Ref Correction, Channel J/S, Channel C/No, Ground Speed, and 100+ other features/ variables. The decision trees automatically selected the decision variables/features as shown in Figure 35. The automatic variables/features selection is one of the advantages of decision trees.

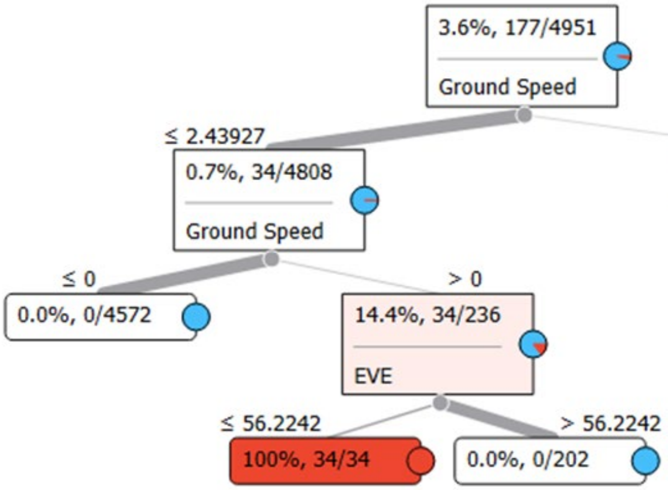


Figure 35. CART Model Results

A second workflow was created in Orange to test the model design concept. Figure 36 displays the workflow used for the supervised learning model. The workflow includes random forest, neural network, logistic regression, and kNN models. Multiple supervised models were selected for this workflow to decide on the best model based on performance measures. The workflow provides the option to run the models in several different configurations. For this stage, the models were run using 10 folds cross validation, test on train data, and test on test data configurations. It is customary for a dataset to be split into three portions, a training set, a validation set, and a test set. Running cross-validation with 10 folds eliminates the need for a validation set in the development and testing of the model. For the test on train set configuration, Orange sets the spilt ratio for the training and test sets. The test and score widget along with the lift curve and the confusion matrix widgets captured the performance metrics of the models. The scatter plot widget displays the characteristics of the output data.

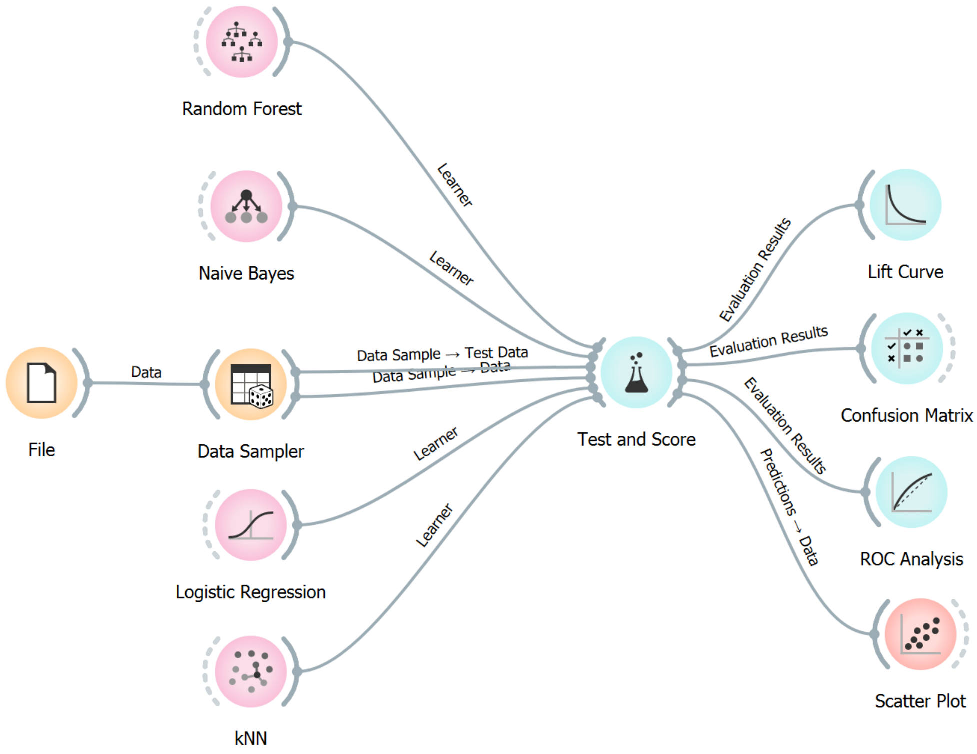
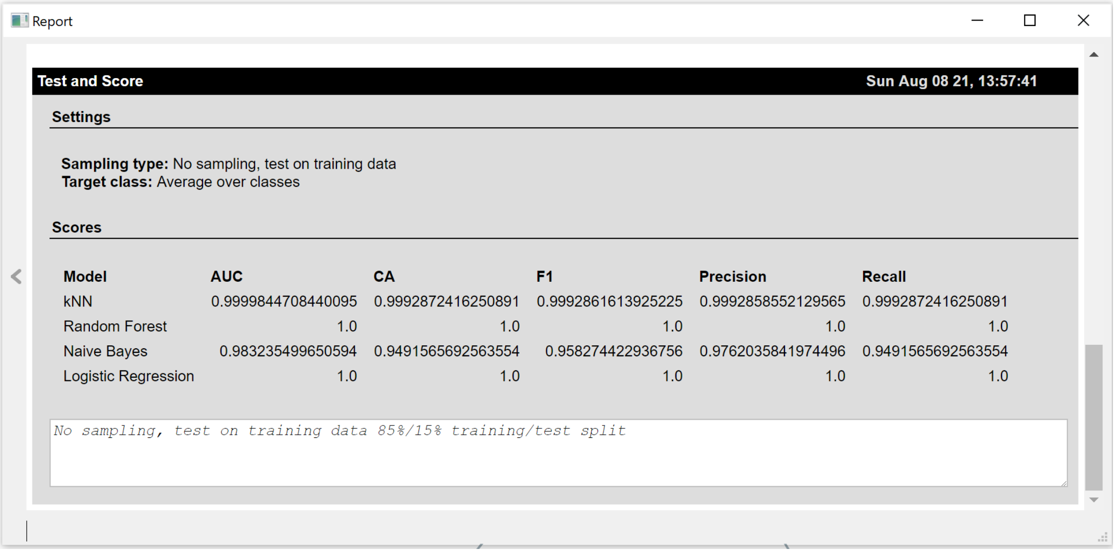


Figure 36. Supervised Learning Orange Workflow

3. Model Performance

After loading the dataset and running the workflow, the results were examined to determine the best model. Burkov (2019, 54) suggests that confusion matrix, accuracy, cost-sensitive accuracy, precision/recall, and the area under the receiver operating characteristic (ROC) curve (AUC) are “the most widely used metrics and tools to assess the classification model.” Figure 37 shows the performance metrics of each model from testing on the training dataset. All models performed admirably on the training data. The AUC measures the area underneath the entire ROC curve from (0,0) to (1,1) and just as integral calculus, the AUC measures the area under a curve. A ROC curve is a graph that shows the performance of a classification model at all classification thresholds (Google Developers n.d.). The ROC curve represents the true positive rate (TPR) and the false positive rate (FPR) parameters.



Test and Score Sun Aug 08 21, 13:57:41

Settings

Sampling type: No sampling, test on training data
Target class: Average over classes

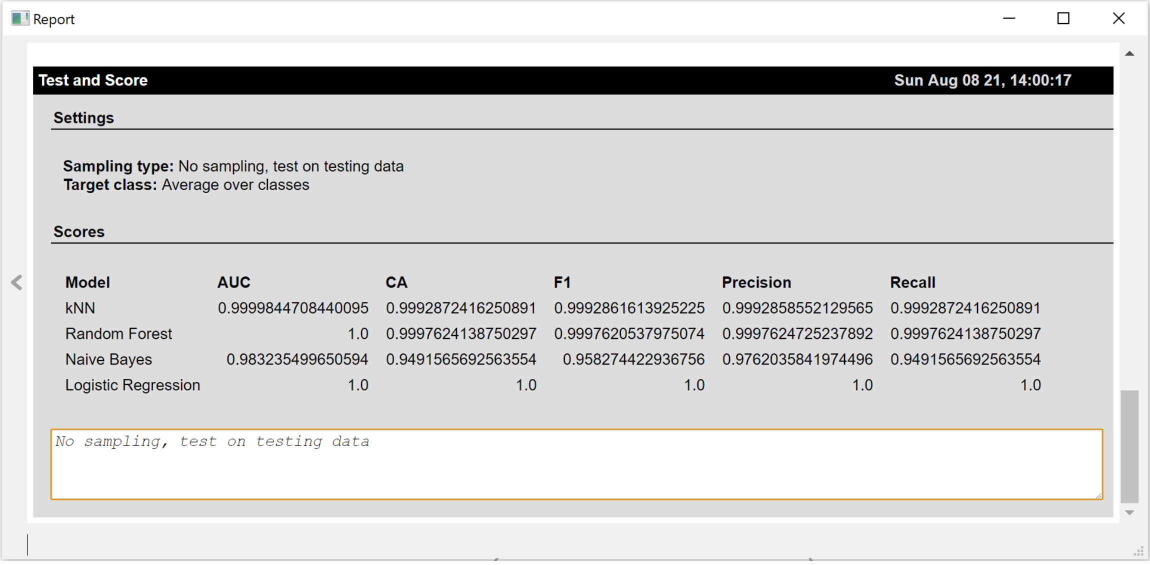
Scores

Model	AUC	CA	F1	Precision	Recall
kNN	0.9999844708440095	0.9992872416250891	0.9992861613925225	0.9992858552129565	0.9992872416250891
Random Forest	1.0	1.0	1.0	1.0	1.0
Naive Bayes	0.983235499650594	0.9491565692563554	0.958274422936756	0.9762035841974496	0.9491565692563554
Logistic Regression	1.0	1.0	1.0	1.0	1.0

No sampling, test on training data 85%/15% training/test split

Figure 37. Model Performance Test and Scores on Training Data

Cross validation was used displayed similarly efficient results as on the training data with the random forest model having the best overall scores. The fact that random forest had the best result is in keeping with the instructions from the OS4106 Advanced Data Analysis lecture notes where it was reported that “random forests are very easy to train and tend to work very well in many situations" (Michael Atkinson, lecture notes, May 26, 2021). Because of the ease of use and benefits of random forest, the model should be one of the first go-to models according to the lecture notes. The results shown in Figure 38 represent the scores from the models running the test set. Naïve Bayes, as in the training set results, shows the lowest ratings of the models. The other three models once again performed at close to 100% with the logistic regression model topping the list.



Test and Score Sun Aug 08 21, 14:00:17

Settings

Sampling type: No sampling, test on testing data
Target class: Average over classes

Scores

Model	AUC	CA	F1	Precision	Recall
kNN	0.9999844708440095	0.9992872416250891	0.9992861613925225	0.9992858552129565	0.9992872416250891
Random Forest	1.0	0.9997624138750297	0.9997620537975074	0.9997624725237892	0.9997624138750297
Naive Bayes	0.983235499650594	0.9491565692563554	0.958274422936756	0.9762035841974496	0.9491565692563554
Logistic Regression	1.0	1.0	1.0	1.0	1.0

```
No sampling, test on testing data
```

Figure 38. Model Performance Scores on Test Data

Elaboration of the performance measures applied to the models are presented here. Precision is represented mathematically as "the ratio of correct positive predictions on the overall number of positive predictions" (Burkov 2019, 55):

$$precision = \frac{TP}{TP+FP},$$

According to Burkov (2019, 55), recall captures "the ratio of correct positive predictions to the overall number of positive examples in the dataset."

$$recall = \frac{TP}{TP+FN},$$

Classification accuracy (CA) “is the ratio of the number of correct predictions to the total number of input samples” (Mishra 2018). In some instances, the CA value is represented by accuracy. Burkov (2019, 56) states that “accuracy is a useful metric when errors in predicting all classes are equally important.”

$$accuracy = \frac{TP + TN}{TP + TN + FP + FN},$$

The F1 score ranges between 0 and 1 and is seen as a measure of the precision of the classifier. Mishra (2018) defines the F1 score as “the harmonic mean between precision and recall.” A confusion matrix is another metric used calculate precision and recall. Examining the test and score results from the Test and Score widget, all models had similar overall scores. The overall scores from the supervised learning models demonstrate that both ML tools can be applied to the dataset from the engineered system to effectively used to detect the spoofing threat on the DAGR.

V. CONCLUSIONS

The research documented in this thesis explored the use of machine learning to detect GPS spoofing attempts. The thesis developed an experimental methodology to analyze the utility of machine learning methods to this problem and conducted a set of experiments to demonstrate the utility of this approach. The thesis used a prototype military GPS system and created hybrid livestream/simulated GPS datasets for training and experimentation. The thesis concluded that both unsupervised and supervised machine learning methods show applicability for detecting spoofing attempts in military GPS systems. The thesis recommends further research and development with the intent to lead to an operational solution to address this critically important threat.

A. THE UTILITY OF MACHINE LEARNING FOR DETECTING GPS SPOOFING ATTEMPTS

Throughout Chapters III and IV we demonstrated that ML algorithms can be used to effectively detect GPS spoofing on the DAGR. In evaluating the ML tools, we applied concepts from both the unsupervised learning and the supervised learning arenas. For the unsupervised learning concepts several clustering algorithms were applied to the dataset with the goal of identifying and classifying the spoofing activities. Distance parameters were used as one measure in the initial stages of using clustering schemas. The relationship between EPE and GPS time was one of the predictors from the dataset that demonstrated the result of the ML tools. The grouping of the data formed by the clusters provided a clear demarcation between the normal data from the live sky feed and the spoofing data generated from the GSG-62 GNSS simulator. The clusters provided visual evidence of the effectiveness of the ML tools.

The unsupervised methods applied to the dataset revealed the characteristics of the data while clearly showing the effects of the application of the spoofing signal. The clusters were identified and verified with the timeline of active spoofing. The study identified peculiarities and previously unidentified correlations within the data parameters that proved enlightening to the study. Applying data reduction methods such as PCA revealed that even with upwards of 100 predictors in the dataset, approximately 94% of the data can be represented by the first principal component. This points to a dataset that is highly correlated as evidenced by the Pearson correlations captured throughout the study.

The next level of the evaluation applied methods and tools from the supervised learning ML arena. As the dataset generated from the engineered system did not include a labeled parameter to capture the spoofing threat, we applied data mining tools combined with requirements from the NMEA 0183 and ICD-153 standards to define spoofing. The supervised learning tools were then applied to the new dataset to train and test the newly developed models. The effectiveness of supervised learning models was evaluated using a suite of performance measures that are generally applied to ML, data science, and statistics problems. The training of the models presented excellent results with perfect recall and precision for all models. Recall is an important metric used in assessing the effectives of a tool in the detection of malicious activity such as spoofing attempts on the DAGR. Lecture notes from the *CS4315 Introduction to Machine Learning* class, state that,

- In detecting malicious behavior on a digital system, recall is more important than precision.
- You don't want to miss attacks, and you are willing to accept a lot of false positives. (Neil Rowe, lecture notes, July 21, 2021)

The approach to this problem offered an opportunity to apply tools from both unsupervised learning and supervised learning algorithms to identify the spoofing threat during the application of the interference on the system.

B. AN EXPERIMENTAL METHODOLOGY FOR ANALYZING MACHINE LEARNING METHODS FOR GPS SPOOFING DETECTION

To begin the experiment of spoofing detection, it was imperative that an engineered system be developed to create a prototype to simulate a representative operational system. The design criteria of the engineered system included functional and operational requirements. The input to the system comprised live sky GPS traffic and spoofing signal traffic emanating from a GNSS simulator. This concept resulted in the development of a hybrid live-stream/simulated data sets to both train and experiment with the prototype GPS model. The results of the data analysis and model development revealed that the engineered system can indeed be spoofed effectively. The application of a spoofing signal successfully moved the target, placing the user in a different location. The spoofing method used in this study is described as overt spoofing in the publications. A feature of overt spoofing is the jam-then-spoof strategy. This strategy is described as one where "the counterfeit GPS signals are simply broadcast at a significantly higher power level than the authentic satellite signals." Chapman (2017, 1). Evidence of overt spoofing was identified as reported by the GPGSA message. The reduced number of active satellites in the solution were some of the first indicators of the impact of application of the spoofing signal on our system. While the number of visible SVs and number of healthy SVs recorded in message ID 5040 did not change from a normal operational scenario to a spoofing scenario, the number of active satellites reported in the GPGSA message were reduced. The equivalent DOP measures in the GPGSA messages correspondingly increased to values higher than the normal operating range. Overt spoofing methods were also described by (Kerns, et al. 2014), where the point was made that the adversary is not interested in being inconspicuous; their singular goal is to overpower the target.

The requirements for the engineered system were tailored to both the operational and functional contexts of the experiment. For this study, the Orolia GSG-62 Multiband GNSS simulator was used to generate GPS spoofing signals by applying FDI onto the engineered system. Once again following a scaled version of the systems engineering construct, the results of the system would undergo test and evaluation. The test and evaluation (T&E) process is employed to verify the requirements of the system. Validation and verification are two steps used in systems engineering to assess the T&E phase of the systems engineering process. The AcqNotes portal describes verification as "a quality control process that determines if a system meets its system-level requirements" (AcqNotes 2021). System verification methods are customarily listed as inspection, test, analysis, and demonstration. Demonstration and test were the verification testing methods used in the verification of the engineered system requirements.

The results of requirements verification include:

- The system operated as a software application as discussed throughout the report.

- The system received NMEA 0183 navigation messages.
- The system received ICD 153 navigation messages.
- The system received GPS synchronization timing information.
- The system received live sky GPS input signals through the antenna port.
- The system identified spoofing attacks using AI methods.
- The system accepted navigation input from the user through the spoofing scenarios.
- The system operated in accordance with USMC CONOPS.

The results of the data analysis and model development revealed that the engineered system can indeed be spoofed effectively. The application of a spoofing signal successfully moved the target, placing the user in a different location. The spoofing method used in this study is described as overt spoofing in various publications. A feature of overt spoofing is the jam-then-spoof strategy. This strategy is described as one where "the counterfeit GPS signals are simply broadcast at a significantly higher power level than the authentic satellite signals" Chapman (2017, 1). Evidence of overt spoofing was identified as reported by the GPGSA message. The reduced number of active satellites in the solution were some of the earliest indicators of the impact of application of the spoofing signal on our system. While the number of visible SVs and number of healthy SVs recorded in message ID 5040 did not change from a normal scenario to a spoofing scenario, the number of active satellites reported in the GPGSA message were reduced. The corresponding DOP measures in the GPGSA messages similarly increased to values higher than the normal operating range. Overt spoofing methods were also described by Kerns et al. (2014), where the point was made that the adversary is not interested in being inconspicuous; their singular goal is to overpower the target.

The unsupervised methods applied to the dataset revealed the characteristics of the data while clearly showing the effects of the application of the spoofing signal. The clusters were identified and verified with the timeline of active spoofing. The study identified peculiarities and previously unidentified correlations within the data parameters that proved enlightening to the study. The decision to create a labeled dataset for supervised learning model has proven to be an effective strategy as evidenced by the results of the application of the training data. The supervised learning model was created and verified using CART in RStudio and further evaluated by the supervised models in Orange. As shown in this study, given the appropriate tools and access, an adversary can effectively spoof a military GPS device. The tools developed and demonstrated throughout this thesis show that we can use AI methods to detect spoofing attacks on the military GPS infrastructure.

C. FUTURE WORK

The study and results presented in this thesis are but a first step in the creation of an operational spoofing detection system. Additional research should be conducted to advance the methods used herein. The engineered system used in this study applied the theories described as simplistic spoofing and overt spoofing. As the names imply, there are additional levels of sophistication that can be used to apply to the spoofing threat. More research and development (R&D) is needed to move the study forward. Future work should attempt to further study the impact of more sophisticated measures on the GPS infrastructure.

1. Training and Testing Dataset

The datasets used in this study were generated from one operational scenario and three separate spoofing scenarios created and executed from the GSG-62 GNSS simulator. Each of the spoofing scenarios were executed for a period of a few hours with each scenario representing a different spoofing operation. Having this limited dataset impeded the development of adequate repository of training and testing data. Having access to a larger repository of training data representing multiple spoofing scenarios would offer a more comprehensive training dataset. The rate of change of the parameters within the dataset is a valuable indicator of anomalous activities. Exploration of the impact of the first and second derivatives should be undertaken as part of the dataset used in training and testing as they usually provide valuable insight for anomaly detection. Future work should also study the impact of more sophisticated measures on the GPS infrastructure.

2. Tuning Model Parameters

The algorithm exploration for this study used RStudio. Using the RStudio toolset allowed us to experiment with individual parameters of the dataset and the options available in the application. Orange does not offer the same level of flexibility as most parameters are set "under the hood." Further exploration of the capabilities of RStudio along with tools from providers such as TensorFlow, MATLAB, and Python would allow for customizing the model for spoofing detection. Tuning of the hyperparameters would also allow for better utility of the spoofing detection applied to the problem space.

3. Testing on Keyed Systems

Some of the vulnerabilities seen in this study can be addressed by applying keys to the GPS device used in the study. Can an adversary spoof a keyed system? The limited capability of the GNSS simulator used in the study prevented the testing of the methods on an encrypted system. Future work should include the exploration of means to test the algorithms developed in this study on an encrypted DAGR or a Military GPS User Equipment (MGUE) system.

4. Creating Software Application

The usefulness of this study will ultimately be determined by the application of the spoofing detection system in an operational environment. The engineered system used in this study used a combination of CTS-153 software, a LabVIEW program, and visually monitoring the DAGR to detect evidence of spoofing activities. Development of a user interface that can be installed in an operations center is critical for practical use. The user interface would allow the user to receive auditory and visual cues alerting the user to the detection of anomalous activity. The spoofing detection was developed as a supplement to the human element.

LIST OF REFERENCES

Abramovich, Giselle. 2021. “Fifteen Mind-Blowing Stats about Artificial Intelligence.” *Adobe* (blog). Accessed June 26, 2020. https://www.adobe.com/cn/insights/15-stats-about-artificial-intelligence.html.

AcqNotes. 2021. “Systems Engineering: Verification and Validation.” https://acqnotes.com/acqnote/careerfields/validation-and-verification.

Ao, Sio-Iong. 2010. “Hybrid Machine Learning Model for Continuous Microarray Time Series.” In *Advances in Machine Learning and Data Analysis*, Sio-Iong Ao, Burghard Reiger and Mahyar Amouzegar, 57 - 77. New York: Springer.

Atkinson, Beth. 2019. “The ‘rpart’ Package.” RDocumentation. April 12. https://www.rdocumentation.org/packages/rpart/versions/4.1-15.

BAE Systems. 2020. “Defense Advanced GPS Receiver.” https://www.baesystems.com/en-uk/product/defense-advanced-gps-receiver.

Barker, Brian C, John W Betz, John E Clark, Jeffrey T Correia, James T Gillis, Steven Lazar, Kaysi A Rehborn, and John R. Stratton III. 2005. “Overview of the GPS M Code Signal. “ “The MITRE Corporation.” https://www.mitre.org/sites/default/files/pdf/betz_overview.pdf.

Bhamidipati, Sriramya, Kyeong, J Kim, Hongbo Sun, Philip Orlik, and Jinyun Zhang. 2019. “Joint BP and RNN for Resilient GPS Timing Against Spoofing Attacks.” *PeerJ Computer Science*, 1–20.

Blewit, Geoffrey. 1997. “Basics of the GPS Technique: Observation Equations.” In *Geodetic Applications of GPS*, 1–46. Swedish Land Survey.

Bonebrake, Chris, and Lori Ross O’Neil. 2014. “Attacks on GPS Time Reliability.” *IEEE Computer and Reliability Societies* 82–84.

Browne, Mathuel. 2016. “Battlefield Appears With The Touch of A Finger.” U.S. Marines. June 14, 2016. https://www.marcorsyscom.marines.mil/News/News-Article-Display/Article/798476/battlefield-appears-with-the-touch-of-a-finger/.

Burkov, Andriy. 2019. *The Hundred-Page Machine Learning Book.* San Bernardino: Andriy Burkov.

Chen, Zhiyuan, and Bing Liu. 2018. *Lifelong Machine Learning.* San Rafael: Morgan & Claypool Publishers.

Choudhary, Mahashreveta. 2019.”What Are the Various GNSS Systems?” *Geospatial World* (blog). November 20, 2019. https://www.geospatialworld.net/blogs/what-are-the-various-gnss-systems/#:~:text=But%20did%20you%20know%20GPS,IRNSS%20or%20NavIC%20(India).

Cole, Sally. 2015. “Securing Military GPS from Spoofing and Jamming Vulnerabilities.” Military Embedded Systems. November 30, 2015. https://militaryembedded.com/comms/encryption/securing-military-gps-spoofing-jamming-vulnerabilities.

Crockett, Christopher. 2012. *EarthSky.* June 10. https://earthsky.org/astronomy-essentials/what-is-sidereal-time.

Cuntz, Manuel, Konovaltsev, Achim Dreher, and Michael Meurer. 2012. “Jamming and Spoofing in GPS/GNSS Based Applications and Services - Threats and Countermeasures.” *7th Security Research Conference Future Security.* Bonn: Springer-Verlag. 196–199.

Cybersecurity & Infrastructure Security Agency. 2019. *Security Tip (ST04-015) Understanding Denial-of-Service Attacks.* November 20. https://us-cert.cisa.gov/ncas/tips/ST04-015.

Darwish, Muhammad. 2017. *Did Russia Make This Ship Disappear?* CNN. November 3. https://money.cnn.com/2017/11/03/technology/gps-spoofing-russia/index.html.

Delua, Julianna. 2021. *Supervised vs. Unsupervised Learning: What’s the Difference?* March 12. https://www.ibm.com/cloud/blog/supervised-vs-unsupervised-learning.

Demšar, Janez, and Blaž Zupan. 2013. *Orange: Data Mining Fruitful and Fun - A Historical Perspective.* https://www.semanticscholar.org/paper/Orange%3A-Data-Mining-Fruitful-and-Fun-A-Historical-Demsar-Zupan/2a52478be9b4055aaae729090846e8dc318f7672.

Dietterich, Thomas G. 2009. “Machine Learning and Ecosystem Informatics: Challenges and Opportunities.” In *Advances in Machine Learning*, by Zhi-Hua Zhou and Takashi Washio, 1 - 5. Nanjing: Springer.

Donges, Niklas. 2019. *A Guide To RNN: Understanding Recurrent Neural Networks and LSTM.* June 16. https://builtin.com/data-science/recurrent-neural-networks-and-lstm.

Enge, Per K. 1994. “The Global Positioning System: Signals, measurements, and performance.” *International Journal of Wireless Information Networks* 83–105.

Eremenko, Kirill, and Hadelin de Ponteves. 2021. “Machine Learning A-Z™: Hands-On Python & R In Data Science.” *udemy.com.* https://www.udemy.com/course/machinelearning/learn/lecture/11692666#overview.

Faisandier, Alan, Garry Roedler, Richard Turner, Rick Adcock, and Ariela Sofer. 2021. *System Requirements.* May 19. https://www.sebokwiki.org/wiki/System_Requirements.

Farivar, Cyrus. 2013. *Professor Fools $80M Superyacht's GPS Receiver on the High Seas.* July 29. https://arstechnica.com/information-technology/2013/07/professor-spoofs-80m-superyachts-gps-receiver-on-the-high-seas/.

FCC. 2020. *Jammer Enforcement.* April. https://www.fcc.gov/general/jammer-enforcement.

FCC. 2013. *Notice of Apparent Liability for Forfeiture (NAL).* Notice of apparent liability, Washington, DC: Federal Communications Commission.

Ferre, Ruben. 2018. "Analysis of GNSS Replay-Attack Detectors Exploiting Unpredictable Symbols." PhD Thesis.

Frankenfield, Jake. 2021. *Data Analytics.* April 16. https://www.investopedia.com/terms/d/data-analytics.asp.

Fujimaki, Ryohei, Satoshi Morinaga, Michinari Momma, Kenji Aoki, and Takayuki Nakata. 2009. "Linear Time Model Selection for Mixture of Heterogeneous Components." In *Advances in Machine Learning*, by Zhi-Hau Zhou and Takashi Washio, 82 - 97. Nanjing: Springer.

Gakstatter, Eric. 2015. *GPS World.* February 4. https://www.gpsworld.com/what-exactly-is-gps-nmea-data/.

Garamone, Jim. 2018. "National Defense Strategy a 'Good Fit for Our Times,' Mattis Says." Department of Defense. January 19. https://www.defense.gov/Explore/News/Article/Article/1419671/national-defense-strategy-a-good-fit-for-our-times-mattis-says/.

General Dynamics Mission Systems. 2021. *C21 Combiner.* https://www.gpssource.com/products/rf-gps-signal-combiner.

GIS Geography. 2020. *GPS Accuracy: HDOP, PDOP, GDOP, Multipath & the Atmosphere.* December 26. https://gisgeography.com/gps-accuracy-hdop-pdop-gdop-multipath/.

Gonzales, Matt. 2019a. *Handheld Digital Targeting System Provides Fire And Air Support To Marines.* September 12. https://www.marines.mil/News/News-Display/Article/1958760/handheld-digital-targeting-system-provides-fire-and-air-support-to-marines/.

———. 2019b. *Handheld Tablet Improves Situational Awareness.* September 5. https://www.marcorsyscom.marines.mil/News/News-Article-Display/Article/1952936/handheld-tablet-improves-situational-awareness/.

Google Developers. n.d. *Classification: ROC Curve and AUC.* https://developers.google.com/machine-learning/crash-course/classification/roc-and-auc.

GPS.gov. 2014. "GPS Applications." November 25, 2014. https://www.gps.gov/applications/.

———. 2020. *GPS Accuracy.* April 22. https://www.gps.gov/systems/gps/performance/accuracy/.

———. 2018. "Control Segment." November 8. https://www.gps.gov/systems/gps/control/.

———. 2021a. *Interface Control Documents.* January 19. https://www.gps.gov/technical/icwg/.

———. 2021b. *Pseudorandom Noise Code Assignments.* January 26. https://www.gps.gov/technical/prn-codes/.

Grady, John. 2020. *Panel Details Global Artificial Intelligence Arms Race.* December 9. https://news.usni.org/2020/12/09/panel-details-global-artificial-intelligence-arms-race.

Haider, Zeeshan, and Shehzad Khalid. 2016. "Survey on Effective GPS Spoofing Countermeasures." *Sixth International Conference on Innovative Computing Technology.* INTECH. 573–577.

Hall, Patrick, and Navdeep Gill. 2018. *An Introduction to Machine Learning Interpretability.* Sebastopol: O'Reilly Media, Inc.

Han, Jiawei, and Micheline Kamber. 2006. *Data Mining Concepts and Techniques.* San Francisco: Morgan Kaufmann.

Hargrave, Mardshall, and Somer Anderson. 2021. *Deep Learning.* May 17. https://www.investopedia.com/terms/d/deep-learning.asp.

Hastie, Trevor, Robert Tibshirani, and Jerome Friedman. 2008. *The Elements of Statistical Learning Data Mining, Inference, and Prediction.* Stanford: Springer.

Hohman, Tyler. 2020. "Orolia" (blog), May 19. https://www.orolia.com/resources/blog/tyler-hohman/2020/inside-scoop-gps-spoofing.

Howell, Elizabeth. 2018. *Space.com.* April 27. https://www.space.com/19794-navstar.html.

Huber, Mark. 2018. *AIN online.* May 22. https://www.ainonline.com/aviation-news/business-aviation/2018-05-22/gps-jamming-and-spoofing-rise.

Huddleston, Samuel H, and Gerald G Brown. 2019. "Machine Learning." In *INFORMS Analytics Body of Knowledge*, by James J Cochran, 231–274. Hoboken: John Wiley & Sons, Inc.

Humphreys, Todd E, Brent M Ledvina, Mark L Psiaki, Brady W O'Hanlon, and Paul M Kintner Jr. 2008. *Assessing the Spoofing Threat: Development of a Portable GPS Civilain Spoofer.* Conference, Savanna: ION GNSS Conference.

International Civil Aviation Organization.. 2019. *Guidance on GNSS Implementation in the MID Region.* CNS planning and implementation in the MID Region, Cairo: International Civil Aviation Organization.

INCOSE. n.d. *What is Systems Engineering?* Accessed June 6, 2021. https://www.incose.org/about-systems-engineering.

Innoslate. 2021. *Innoslate 101.* https://www.innoslate.com/help-center/innoslate-101/.

IPSES. 2015. *IPSES Scientific Electronics.* https://www.ipses.com/eng/In-depth-analysis/Standard-of-time-definition.

Jafarnia Jahromi, Ali. 2013. *GNSS Signal Authenticity Verification in the Presence of Structural Interference.* Unpublished doctoral thesis, Calgary, AB: University of Calgary.

Jafarnia-Jahromi, Ali, Ali Broumandan, John Nielsen, and Gerard Lachapelle. 2012. "GPS Vulnerability to Spoofing Threats and a Review of Antispoofing Techniques." *International Journal of Navigation and Observation* 16.

James, Gareth, Daniela Witten, Trevor Hastie, and Tibshirani. 2017. *An Introduction to Statistical Learning with Applications in R.* New York: Springer.

Johari, Aayushi. 2020. *AI Applications: Top 10 Real World Artificial Intelligence Applications.* September 18. https://www.edureka.co/blog/artificial-intelligence-applications/.

Joint Program Office. 2002. *GPS User Equipment (Phase III) Interface Control Document for the Precise Time and Time Interval (PTTI) Interface.* DOC. NO. ICD-GPS-060. Accessed September 25, 2020. https://www.navcen.uscg.gov/pdf/gps/ICD-GPS-060B.pdf.

Jones, Michael. 2019. “New Military Code about to Board 700 Platforms.” GPS World. April 9. Accessed December 26, 2020. https://www.gpsworld.com/new-military-code-about-to-board-700-platforms/.

Kaplan, Elliott D, and Christopher J Hegarty. 2017. *Understanding GPS/GNSS Principles and Applications.* Boston: Artech House.

———. 2006. Understanding GPS: Principles and Applications. Norwood: Artech House, Inc.

Kasparov, Garry. 2018. *Intelligent Machines Will Teach Us—Not Replace Us | WSJ | May 7th, 2018.* May 7. Accessed July 13, 2021. https://www.kasparov.com/intelligent-machines-will-teach-us-not-replace-us-wsj-may-7th-2018/.

Kent State University. 2021. *Statistical & Qualitative Data Analysis Software: About R and RStudio.* July 7. https://libguides.library.kent.edu/statconsulting/r.

Kerns, Andrew, Daniel P Shepard, Jahshan A Bhatti, and Todd E Humphreys. 2014. “Unmanned Aircraft Capture and Control Via GPS Spoofing.” *Journal of Field Robotics* 617–636.

Khan, Shah Z, Mujahid Moshin, and Waseem Iqbal. 2021. “On GPS Spoofing of Aerial Platforms: A Review of Threats, Challenges, Methodologies, and Future Research Directions.” *PeerJ Computer Science.*

Knapp, Brandon. 2018. *The Army Wants to Give Soldiers a Netflix-Like Recommendation on the Battlefield.* May 2. https://www.c4isrnet.com/it-networks/2018/05/02/the-army-wants-to-give-soldiers-a-netflix-like-recommendation-on-the-battlefield/.

Korolov, Maria. 2019. *What is GPS Spoofing? And How You Can Defend Against It. csoonline.com.* May 17. https://www.csoonline.com/article/3393462/what-is-gps-spoofing-and-how-you-can-defend-against-it.html.

Kotsiantis, S. B. 2007. “Supervised Machine Learning: A Review of Classification Techniques.” *Informatica 31* 249–268.

Kuhn, Max. 2021. *Package ‘caret’.* May 15. https://cran.r-project.org/web/packages/caret/caret.pdf.

Lakshminarayana, Subhash, Ablla Kammoun, Meroune Debbah, and H Vincent Poor. 2020. *Data-Driven False Data Injection Attacks Against Power Grids: A Random Matrix Approach.* July 21. https://arxiv.org/abs/2002.02519.

lawinsider.com. 2021. *Mission Requirements Definition.* https://www.lawinsider.com/dictionary/mission-requirements.

Ledvina, B. M, W. J Bencze, B Galusha, and Miller. 2010. "An In-Line Anti-Spoofing Device for Legacy Civil GPS Receivers." *Proceedings of the Institute of Navigation—International Technical Meeting.* San Diego: Institute of Navigation. 698–712.

Lee, Connie. 2018. *National Defense.* January 4. https://www.nationaldefensemagazine.org/articles/2018/1/4/spoofing-risks-prompt-military-to-update-gps-devices#:~:text=In%202011%2C%20Iranian%20citizens%20turned,GPS%20coordinates%20of%20the%20system.

logsa.army.mil. n.d. *Keying Is the Key to GPS Protection!* Accessed January 29, 2021. https://www.logsa.army.mil/psmag/archives/PS2016/761/761-49-51.pdf.

Magnuson, Stew. 2010. "National Defense." *National Defense.* January 1 https://www.nationaldefensemagazine.org/articles/2009/12/31/2010january-military-swimming-in-sensors-and-drowning-in-data.

MathWorks. 2020. *Machine Learning with MATLAB.* Natick: The MathWorks, Inc.

Mattis, James N. 2018. "Defense.gov." *U.S. Department of Defense.* January 19. https://www.defense.gov/Newsroom/Transcripts/Transcript/Article/1420042/remarks-by-secretary-mattis-on-the-national-defense-strategy/#:~:text=Success%20does%20not%20go%20to,department%27s%20needs%20for%20new%20equipment.

McAfee. 2020. *What is GPS Spoofing?* August 25. https://www.mcafee.com/blogs/consumer/what-is-gps-spoofing/.

McLaughlin, Matt. 2020. *Protecting the U.S. Military from GPS Jamming and Spoofing.* C4ISRNET Whitepaper, Tysons: C4ISRNET.

McLeod, Saul. 2019. *Constructivism As a Theory for Teaching and Learning.* https://www.simplypsychology.org/constructivism.html.

———. 2019. *What a P-Value Tells You About Statistical Significance.* May 20. https://www.simplypsychology.org/p-value.html.

Merriam-Webster. 2021. "*perspicacious.*" https://www.merriam-webster.com/dictionary/perspicacious#learn-more.

Merriam-Webster. 2021. S.v. *"ephemeris."* https://www.merriam-webster.com/dictionary/ephemeris.

Microsoft. 2021. *Reinforcement Learning.* https://www.microsoft.com/en-us/research/theme/reinforcement-learning-group/.

Mishra, Aditya. 2018. *Metrics to Evaluate your Machine Learning Algorithm.* February 24. https://towardsdatascience.com/metrics-to-evaluate-your-machine-learning-algorithm-f10ba6e38234.

Montgomery, Paul Y, Todd E Humphreys, and Brent M Ledvina. 2009. "Receiver-Autonomous Spoofing Detection: Experimental Results of a Multi-antenna Receiver Defense Against a Portable Civil GPS Spoofer." *ION 2009 International Technical Meeting.* Anaheim: Institute of Navigation. 124–130.

Moody, Oliver. 2018. *Galileo Satellite Row Endangers Military.* May 18. https://www.thetimes.co.uk/article/galileo-satellite-row-endangers-military-cjcvsr8lm.

Morgan, Forrest E, Benjamin, Lohn, Andrew J Boudreaux, Mark Ashby, Christian Curriden, Kelly Klima, and Derek Grossman. 2020. *Military Applications of Artificial Intelligence: Ethical Concerns in an Uncertain World.* Santa Monica: RAND Corporation.

Mosavi, M. R, and Maryam Moazedi. 2017. "Detection of Spoofing Attack Using Machine Learning Based on Multi-Layer Neural Network in Single-Frequency GPS Receivers." *The Journal of Navigation* 1–20.

Multi-Physics Technologies. 2021. *Artifical Intelligence.* https://multi-physics.com/machine-learning/.

National Cybersecurity & Communications Integration Center. 2016. *Improving the Operation and Development of Global Positioning System (GPS) Equipment Used by Critical Infrastructure.* Best Practices Guide, Department of Homeland Security Science & Technology Directorate.

National Marine Electronics Association. 2012. "NMEA 0183 - Standard for Interfacing Marine Electronic Devices." *END-USER LICENSE AGREEMENT FOR THE NMEA 0183® STANDARD.* Severna Park: NMEA, August 1.

National Research Council. 1997. The Global Positioning System for the Geosciences: Summary and Proceedings of a Workshop on Improving the GPS Reference Station Infrastructure for Earth, Oceanic, and Atmospheric Science Applications. Washington, DC: he National Academies Press.

Newberry, Marshall E. 2014. U.S. *Coast Guard Proceedings.* www.uscg.mil/proceedings.

National Institute of Standards and Technology. 2021. *communications security (COMSEC).* https://csrc.nist.gov/glossary/term/communications_security.

———. 2019. *National Institute of Standards and Technology.* October 24. https://www.nist.gov/pml/time-and-frequency-division/nist-time-frequently-asked-questions-faq.

NMEA. 2012. NMEA 0183 - Standard for Interfacing Marine Electronic Devices. NA: National Marine Electronics Association.

Northern NSW Local Health District. 2014. *r-studio-1.* https://nnswlhd.health.nsw.gov.au/research-support-resources/software-for-research/r-studio-1/.

NovAtel. 2012. *Mitigating the Threat of GPS Jamming.* White Paper, Calgary: NovAtel.

Office of the Department of Defense. 2020. *Global Positioning System Standard Positioning Service Performance Standard.* Performance Standardd, Washington, DC: Office of the Department of Defense.

Ogaja, Clement O. 2011. Applied GPS for Engineers and Project Managers. Reston: ASCE Press.

Oozer, Mohammad Irshaad, and Simon Haykin. 2019. "Cognitive Dynamic System for Control and Cyber-Attack Detection in Smart Grid." *IEEE Access* 78320-78335.

Orolia. 2018. *What is a GPS Simulator?* December 31. https://www.orolia.com/sites/default/files/document-files/What_is_a_GPS_Simulator_WP08-101_revB_v3-02-04-19.pdf.

Pearson, Ronald K. 2005. *Mining Imperfect Data: Dealing with Contamination and Incomplete Records.* Philadelphia: Society for Industrial and Applied Mathematics.

Perdue, Lisa. 2017. *What Is a GPS Simulator?* (blog) December 13. https://www.orolia.com/what-is-a-gps-simulator-2/

Phillips, Nathaniel D. 2019. *YaRrr! The Pirate's Guide to R.* https://bookdown.org/ndphillips/YaRrr/the-four-rstudio-windows.html.

PM PNT. 2021. *General FAQs.* https://pm-pnt.army.mil/faq.

Program Executive Office Soldier. 2019. *Nett Warrior (NW).* https://www.peosoldier.army.mil/Equipment/Equipment-Portfolio/Project-Manager-Close-Combat-Squad-Portfolio/Nett-Warrior/.

Psiaki, Mark L, Todd E Humphreys, and Brian Stauffer. 2016. "Attackers Can Spoof Navigation Signals Without Our Knowledge. Here's How to Fight Back GPS Lies." *IEEE Spectrum* 26–53. https://spectrum.ieee.org/telecom/security/protecting-gps-from-spoofers-is-critical-to-the-future-of-navigation.

R4epis. 2021. *Orientation to RStudio.* https://r4epis.netlify.app/training/r_basics/rstudio_overview/.

Raytheon. 2021. *Advanced Field Artillery Tactical Data System (AFATDS).* https://www.raytheon.com/capabilities/products/afatds.

Rooker, J.W. Captain. 2008. Satellite Vulnerabilities EWS Contemporary Issue Paper. Report, Monterey: NPS.

RStudio. 2021. *Take Control of Your R Code.* https://www.rstudio.com/products/rstudio/.

Russell, Stuart, and Peter Norvig. 2003. *Artificail Intelligence: A Modern Approach.* Upper Saddle River: Prentice Hall.

Shevchenko, Nataliya. 2020. *An Introduction to Model-Based Systems Engineering (MBSE).* December 21. https://insights.sei.cmu.edu/blog/introduction-model-based-systems-engineering-mbse/#:~:text=MBSE%20brings%20together%20three%20concepts%3A%20model%2C%20systems%20thinking%2C,formality%20or%20rules%20in%20simplifying%2C%20representing%2C%20or%20abstractin.

Shultz, Richard H, and Richard D Clarke. 2020. *Big Data at War: Special Operations Forces, Project Maven, and Twenty-First-Century Warfare.* August 25. https://mwi.usma.edu/big-data-at-war-special-operations-forces-project-maven-and-twenty-first-century-warfare/.

Simeone, Osvaldo. 2018. A Brief Introduction to Machine Learning for Engineers. Hanover: now Publishers Inc.

Smith, Olivia. 2020. *The 10 Most Important Packages in R for Data Science.* August 30. https://www.datacamp.com/community/tutorials/top-ten-most-important-packages-in-r-for-data-science.

Sutton, Richard S, and Andrew G Barto. 2015. *Reinforcement Learning: An Introduction.* Cambridge: The MIT Press.

Swarts, Philip. 2017. *Congress Warned of GPS Vulnerabilities.* March 30. https://spacenews.com/congress-warned-of-gps-vulnerabilities/.

Talbot, Partick J, and Dennis Ellis. 2015. *Applications of Artificial Interference for Decision-Making.* CreateSpace Independent Publishing Platform.

Talend SA. 2021. *Data Wrangling: Definition and Examples.* https://www.stitchdata.com/resources/glossary/data-wrangling/.

Tamazin, Mohamed, Malek Karaim, and Aboelmagd Noureldin. 2018. "GNSSs, Signals, and Receivers." In *Multifunctional Operation and Application of GPS*, by Rustam B Rustamov and Arif M Hashimov. https://www.intechopen.com/books/multifunctional-operation-and-application-of-gps/gnsss-signals-and-receivers: IntechOpen.

Techopedia. 2021. *Data Preprocessing.* https://www.techopedia.com/definition/14650/data-preprocessing.

TensorFlow. 2021. *Why TensorFlow.* https://www.tensorflow.org/.

The R Foundation. 2021. *What is R?* https://www.r-project.org/about.html.

Tippenhauer, Nils Ole, Christina Popper, Kasper B Rasmussen, and Srdjan Capkun. 2011. "On the Requirements for Successful GPS Spoofing Attacks." *Proceedings of the 18th ACM Conference on Computer and Communications Security.* Chicago: ACM. 75–85.

Tsui, James Bao-Yen. 2000. Fundamentals of Global Positioning Receivers A Software Approach. New York: John Wiley & Sons, Inc.

U.S. DoT. 2020. *2021–004-Various-GPS Interference.* https://www.maritime.dot.gov/msci/2020-016-various-gps-interference.

UT News. 2013. *UT Austin Researchers Successfully Spoof an $80 million Yacht at Sea.* July 29. https://news.utexas.edu/2013/07/29/ut-austin-researchers-successfully-spoof-an-80-million-yacht-at-sea/#:~:text=Jul%2029%2C%202013-,UT%20Austin%20Researchers%20Successfully%20Spoof%20an%20%2480%20million%20Yacht%20at,a%20custom%2Dmade%20GPS%20device.

Volpe, John A. 2001. Vulnerability Assessment of the Transportation Infrastructure Relying on the Global Positioning System. National Transportation Systems Center.

Vorobeychik, Yevgeniy, and Murat Kantarcioglu. 2018. *Adversarial Machine Learning.* Morgan & Claypool.

Warner, Jon S, and Roger G. Johnston. 2012. "A Simple Demonstration that the Global Positioning System (GPS) is Vulnerable to Spoofing." *The Journal of Security Administration* 9.

Wickham, Hadley. 2018. *dplyr.* November 9. https://rdocumentation.org/packages/dplyr/versions/0.7.8.

Wickham, Hadley, and Garrett Grolemund. 2017. R for Data Science: Import, Tidy, Transform, Visualize, and Model Data. Boston: O'Reilly Media.

Witten, Ian, Eibe Frank, Mark Hall, and Christopher Pal. 2016. *Data Mining Practical Machine Learning Tools and Techniques, Fourth Edition.* Cambridge: Morgan Kaufmann.

Yan, Yachen. 2016. *RDocumentation.* May 13. https://www.rdocumentation.org/packages/MLmetrics/versions/1.1.1

Yegulalp, Serdar. 2019. *What is TensorFlow? The Machine Learning Library Explained.* June 18. https://www.infoworld.com/article/3278008/what-is-tensorflow-the-machine-learning-library-explained.html.

INITIAL DISTRIBUTION LIST

1. Defense Technical Information Center
   Ft. Belvoir, Virginia

2. Dudley Knox Library
   Naval Postgraduate School
   Monterey, California