Using Machine Learning for Battle Management Analysis

Anastacia MacAllister, Patrick Rupp, Jason Garrison, George Hellstern
Lockheed Martin Corporation
Fort Worth, TX
anastacia.m.macallister@lmco.com, patrick.h.rupp@lmco.com, jason.t.garrison@lmco.com, george.hellstern@lmco.com

Colonel Daniel Javorsek
Air Force Operational Test and Evaluation Center
Nellis Air Force Base, Nevada
daniel.javorsek@us.af.mil

Phil Chu
CENTRA
Arlington, VA
philip.chu.ctr@darpa.mil

ABSTRACT

As the U.S. military begins to explore using autonomous and artificially intelligent (AI) agents, serious consideration must be given to how these agents will interact with humans. Adding AI teammates will require understanding current warfighter behavior and how AI agents can augment their capabilities. The goal of integrating AI teammates is to allow them to perform low level and dangerous tasks so humans can focus on high-level battle management. However, to conduct this battle management role successfully operators will require high quality data driven insights that help them make sense of an increasingly complex battlespace. This work begins to look at quantifying and measuring behavior in an air combat simulation using machine learning. The goal is to build an understanding of key performance metrics that help drive mission success or failure. From these insights machine learning agents can be created and tuned to properly weight or select the correct behaviors to maximize the warfighters chances of winning. The initial data analysis strategy is based around answering two simple questions: 1) did blue win? 2) if not, why? The analysis specifically looks at determining the importance of different metrics to the outcome of the scenario. Run level metrics like loss exchange ratio are fed into a Random Forest classifier. This classifier makes a win or loss prediction based on scenario metrics. Then, based on the importance of each feature for making the win/loss decision the relative importance of each metric can be gauged. Initial analysis of the data suggests only a handful of traditional performance metrics play a role determining win/loss for the scenario. The final paper will describe model development and the analysis results. Ultimately, the paper will provide insight for the broader community on how to use ML driven methods to develop battle analysis insights for the warfighter.

ABOUT THE AUTHORS

Anastacia MacAllister is a Machine Learning Researcher at Lockheed Martin's Skunk Works®. Her research focuses on developing machine learning algorithms using sparse or imbalanced data sets, exploratory data analytics, and prototyping novel machine learning algorithms. Dr. MacAllister is currently the data analytics lead for Lockheed's Air Combat Evolution (ACE) work with the Defense Advanced Research Project's Agency (DARPA). The work focuses on investigating how including artificial intelligence in the battlefield impacts complex multi-aircraft scenarios. Dr. MacAllister received her Ph.D. from Iowa State University of Science and Technology in Mechanical Engineering and Human-Computer Interaction.

Patrick Rupp is a Data Engineer at Lockheed Martin's Skunk Works®. Mr. Rupp specializes in developing data pipelines particularly for use in analytics problems ranging in size from small prototypes to enterprise solutions. Mr. Rupp is currently developing the data pipeline for Lockheed's Air Combat Evolution (ACE) work with the Defense Advanced Research Project's Agency (DARPA), and has previously supported Lockheed's work in DARPA's Systems of Systems Integration Technology and Experimentation (SoSITE) program. Mr. Rupp received his Masters in Data Science from Johns Hopkins University.

Jason "Vandal" Garrison is a Project Lead at Lockheed Martin Skunk Works. Mr. Garrison is a former USAF pilot with over 3000 hours in USAF F-15, MQ-1, MQ-9s and USMC F-5 Adversaries as well as a graduate of the USN Strike Fighter Course TOPGUN Adversary program. For the previous 4 years he's worked in Skunk Works Operations

Analysis directorate as an F-35 Instructor Pilot and analyst using both virtual and constructive mission effectiveness tools. Most recently Mr. Garrison has focused on leading Reinforcement Learning applications for combat aircraft at the vehicle control and battle management levels.

George Hellstern is a Program Manager at Lockheed Martin Skunk Works. Mr. Hellstern has over 18 years of experience developing Battle Management applications at LM Aero. He has served as a Program Manager on several customer and internally funded research programs in addition to leading several development teams while at Skunk Works. Mr. Hellstern is responsible for Integrating an Alternative Speech Recognition package onto the F-35 Block 2B. He designed and developed analysis software suites for use in the Joint Strike Fighter (JSF) Virtual Simulation environment, managing computer generated forces and providing a white controller capability for both Red and Blue simulation entities.

Daniel "Animal" Javorsek Colonel Dan Javorsek is the Commander of Detachment 6, Air Force Operational Test and Evaluation Center, Nellis Air Force Base, Nev., and Director, F-35 U.S. Operational Test Team. AFOTEC's Detachment 6 plans, conducts, and reports on realistic, objective, and impartial operational test and evaluation of fighter aircraft. The detachment evaluates the operational effectiveness, suitability, and mission capability of the A-10, F-16, F-15C/E/EX, F-22, and F-35 aircraft, and reports results in support of major acquisition program milestone decisions and combatant command fielding decisions.

Phil Chu is a Systems Engineering and Technical Assistance (SETA) contractor from CENTRA for the Defense Advanced Research Projects Agency (DARPA).

Using Machine Learning for Battle Management Analysis

Anastacia MacAllister, Patrick Rupp, Jason Garrison, George Hellstern
Lockheed Martin Corporation
Fort Worth, TX
anastacia.m.macallister@lmco.com, patrick.h.rupp@lmco.com, jason.t.garrison@lmco.com, george.hellstern@lmco.com

Colonel Daniel Javorsek
Air Force Operational Test and Evaluation Center
Nellis Air Force Base, Nevada
daniel.javorsek@us.af.mil

Phil Chu
CENTRA
Arlington, VA
philip.chu.ctr@darpa.mil

INTRODUCTION

The United States military is increasingly exploring the use of artificially intelligent (AI) agents. Over the past five years unclassified AI investment by the Department of Defense (DOD) has grown from 600 million in 2016 to over 2.5 billion dollars in 2021 (Congressional Research Service, 2020). This emerging AI technology is a key component in future Mosaic warfare strategies outlined by Defense Advanced Research Projects Agency (DARPA) (DARPA, 2018). The goal of Mosaic warfare is to move away from monolithic specialized systems to a flexible distributed, connected, and reconfigurable warfighting strategy (O'Donougue, 2021; Grana, 2021; Sweetser & Bexfield, 2020; Deptula & Penney, 2019; Jensen & Paschkewitz, 2019). However, to implement this strategy machine control of systems is required since humans will quickly become overloaded by the amount of information and orchestration required for such distribution. As a result, the goal is to mitigate overload by using a human in the loop method of command and control. This strategy lets humans' function as a battle manager, focusing on setting strategy and handling edge cases the AI agent is not adept at (Bryan, Patt, & Schramn, 2020). However, to ensure seamless orchestration of missions, great care needs to be taken to mesh battle managers and AI agents (Grooms, 2019). To do this involves helping battle managers understand how different decisions and strategies impact the system as a whole. This level of understanding will let them make more sound battle management decisions by showing them how to select tactics and strategies that make use of their force's strengths and weaknesses.

Unfortunately, generating this level of insight for complex battle scenarios is challenging due to several factors. For example, using traditional statistical analysis requires historical data. However, collecting real world data on all possible battle scenarios is an infeasible option. In addition, new paradigms like Mosaic warfare will not be operational for some time, creating a data vacuum. Also, by the time a Mosaic system is fielded so data can be collected, it's too late in the process to ensure the system is designed in a human-machine interaction friendly way. To work around this issue, simulation needs to play a critical role in helping conduct the analysis (Gordon, 2018). Through using simulation, different scenarios can be developed to feed analysis. This reliance on simulation to drive the insight mining process will help ensure that any issues in a Mosaic system are identified and mitigated before fielding. Also, most importantly developing simulations of Mosaic type components will provide ample amounts of data that can be mined to provide helpful insights battle managers require to best direct the fight.

The goal of the work is to begin quantifying the importance of different performance metrics. This will allow battle managers to determine what metrics drive performance and help them optimize the team to win. While straight forward in theory, learning how to quantify performance in a military scenario is a challenging task (Gordon, 2018) (Drachen, Seif El-Nasr, & Canossa, 2013). Often analytics focuses on high level metrics that do not tell a nuanced story (Gordon, 2018). In addition, the combinatorial explosion of variables found in many military campaigns makes conducting tradeoff analysis challenging due to the uncertainty in what factors are driving outcomes. This work begins to look at identifying how detailed and nuanced performance indicators impact mission success in a Defensive Counter Air (DCA) scenario built in the Air Force's Advanced Framework for Simulation, Integration and Modeling Software (AFSIM). The analysis focuses on building metrics that help answer the questions: 1) Does blue win? and 2) if not why? To answer these questions the work moves beyond simplistic metrics like win vs loss, kill ratio, and mission

effectiveness found in many traditional military modeling and simulation analysis (Ratnoo & Shima, 2012; Alkire, et al., 2020; Sweetser & Bexfield, 2020). Metrics developed look at aspects like success rates of different weapons on specific platforms and the effectiveness of different control logic in the scenario.

While developing the metrics themselves is challenging, they are of limited utility unless they can be tied back to the overall mission goals. To accomplish this, a way needs to be developed that identifies metrics significantly driving performance. To do this the team developed a machine learning based Random Forest classifier. The classifier was trained on the developed run level metrics and then feature importance's were used to help determine the most impactful metrics predicting mission success. After driving features were identified, they were then visualized to help determine tradeoffs and what strategies the battle manager should employ to maximize mission success. The following sections start by describing previous work, development of the analysis, results from the DCA scenario, and lastly conclusions from the work. Ultimately, the insights provided by this work will take steps towards helping manage the increasing complexity introduced by the Mosaic warfare concept. Thus, ensuring battle managers better understand their heterogeneous force, therefore allowing them to make more informed and sound decisions.

BACKGROUND

Much research in the military modeling and simulation space looks at isolated deconstructed tasks and scenarios (Jung, et al., 2019). Ratnoo and Shima looked at how different missile pursuit strategies impacted performance in a simulation (Ratnoo & Shima, 2012). They formulated the model as a traditional controls problem. While helpful for evaluating the differences in high level pursuit strategies, the work has limited potential to offer any insight into high level tactic selection. Hanlon et al. looks at forming an analytical solution to the target-defender tactics problem (Hanlon, Garcia, Casbeer, & Pachter, 2018). However, they again approach a very small piece of the puzzle and solve a low level problem. Heinze et al. look at using machine learning to recognize different pilot maneuvers (Heinze, Gross, & Pearce, 1999). The underlying model is simplistic but shows the first steps towards integrating machine learning pattern recognition into military modeling and simulation work. Connors looks at using AFSIM for determining the effectiveness of a new missile design (Connors, 2015). They use the AFSIM engine to design a Sweep mission. They evaluate the performance of the new missile against previous versions using standard high level measures of effectiveness (MOE) and measures of performance (MOP). They find the new missile drives down completion time and increases the proportion of targets destroyed.

While there is a body of research looking at more simplistic and isolated metrics, work dealing at the campaign level for military modeling and simulation is limited (Jung, et al., 2019). Work by Gordon uses engineering design principals to make a meta-model of a reconnaissance scenario (Gordon, 2018). The goal is to build a model answering more complex questions about scenario performance and capture nuance in tradeoffs. In their work, they describe the importance of identifying features capturing scenario variability and performance when building machine learning models. Their work demonstrates the need for more complex performance evaluation methods in the defense area and the importance of understanding what features drive performance. Gulden et al. looks to evaluate the utility of Mosaic warfare compared to traditional monolithic methods (Gulden, Lamb, Hagen, & O'Donoughue, 2021). They develop a low fidelity simulation in NetLogo to model mission completion and effectiveness. Results show that orchestration between different entities in the simulation is highly important for the Mosaic method. Their results suggest that poor orchestration and coordination between entities can reduce effectiveness of the Mosaic method by as much at fifty percent. Their work, while admittedly low fidelity, shows the necessity of running simulations to compare performance. This necessity is especially important for new and emerging concepts where there is little domain knowledge or experience to draw from. Hodicky et al. looks at developing a war game to model the impact of different defense funding decisions (Hodickỳ, et al., 2020). They then perform a case study describing the use of the simulation and how the human element add biases to results and decisions. They assert that most simulation analysis are currently too simplistic and do not take into account the human element. Alkire et al. looks at the need for understanding how space combat will unfold (Alkire, et al., 2020). They make the case that for unknown and new areas of combat, simulation based analysis is critical to understanding key discriminating factors. Rushing et al. begins to look at the roll of AI in wargames (Rushing, Tiller, Tanner, & Mcdowell, 2004). They assert that AI can be a helpful tool but to design a robust system, features used need to match the problem being solved. They use a clustering algorithm called DBSCAN to help identify when a parachute drop occurs in a toy simulated scenario. They believed that this could help operators identify when change has taken place, demonstrating the benefit that AI can have when monitoring a chaotic and changing landscape. Work by Abdelaal begins to look at developing AI for analysists making military

plans (Abdelaal, 2016). They assert that currently the tactics development process is cumbersome and confusing. It also relies too much on insight and not enough on data driven decision making. Their goal is to demonstrate that AI can help sort through different strategies seen in a simulation to help pull out winning methods. They do this for chess and find that using cluster analysis they can find novel behaviors that lead to increased chances of success.

While the limited work surrounding military modeling and simulation demonstrates it's utility, the analytics are often only focusing on high level performance metrics that do not tell a nuanced story of key performance elements (Jung, et al., 2019). Even though the military domain has limited past work looking at more detailed measures of performance, there exists a large deal of literature looking at advanced analytics in the sports and eSprots realm (Parmar, 2018; Almujahed, Ongor, Tigmo, & Sagoo, 2013; Cho, Park, Kim, & Kim, 2017; Joseph, Fenton, & Neil, 2006). Yang et al. looked at predicting the win probability of Dota2 teams (Yang, Qin, & Lei, 2016). They gathered online game data using the standard game plugin and then conducted extensive feature engineering before fitting a model. In their work they highlight the importance of determining the right low level predictive features explaining game outcome before fitting a model. Baio and Blangiardo look at using a Bayesian Network to predict the outcome of soccer games (Biao & Blangiardo, 2010). They build a model off key predictive features like top player scoring and home field advantage. They find that their models with the impactful features included are more accurate at predicting the outcome of a game than standard statistical techniques. Do et al. looks at predicting performance in an online League of Legends game using unsupervised k-means clustering (Do Nascimento, Da Costa, Da Costa Melo, & Marinho, 2017). Their model building process starts by cleaning and analyzing the data to identify the top predictive features. From here they use k-means to find the correct number of clusters that captured behavior variance in teams to identify high vs poor performing strategies. Lee and Ramler look at using support vector machines to help identify patters in higher performing team compositions (Lee & Ramler, 2017). They find a handful of divergent strategies that correlate with an increased win probability against top performing teams. Young et al. looks at using machine learning to help coaches predict team performance (Young, Luo, Gastin, Tran, & Dwyer, 2019). They fit a decision tree and a generalized linear model to the data. They find that they can accurately predict match results, but the decision tree helps coaches identify key performance metrics more easily. Maymin looks at creating nuanced measures to capture how different team member actions impact the team's chance of winning (Maymin, 2021). They use methods like logistic regression and fast frugal trees to determine what derived metrics are the most predictive of team success. They find that their derived metrics explain team losses based on player behavior more strongly than simplistic measures like kills or longevity. Cea looks at the quality and impartiality of the World Cup ranking system using historical data from match winners and losers based on initial ranking (Cea, et al., 2020). They identify the key predictors in the previous model and point to its bias. From here they design a new model using features that are more accurate predictors of team strength that can feed a more truthful prediction model.

Overall, the research into both military applications and sports strongly suggests that to use machine learning plus other forms of analytics it's necessary to understand what features predict the desired outcomes. As a result, to accomplish the vision of Mosaic warfare outlined by the DoD, understanding predictive features is highly important. Understanding these key features will help identify what key metrics the battle manager should focus on and where AI can assist in the process. The work in this paper begins to address the feature exploration need using data from a DCA scenario developed in AFSIM.

METHODS

The initial data analysis strategy is based around answering two main questions: 1) did blue win? 2) if not, why? In order to answer these questions, the team relied on a tiered analytics approach. The first tier of the analytics aims to determine what metrics are important indicators of a scenario's outcome as suggested by background literature. After these are determined, the second tier looks differences between these key factors for blue wins vs losses. These differences will help suggest to battle managers what tactics and strategies they should employ to maximize wins. This section starts off by first describing how the simulation data was generated, it then dives into metrics development, and finally describes the development of a machine learning (ML) model that identifies key performance indicators.

Data Generation and Scenario Development

Table 1: Attributes of Data Categories

	SIMULATION	LARGE FORCE EXERCISE (LFE)
+	Data set diversity	More representative of fight
	Cost effective	Team dynamics captured
	High fidelity recording	Captures operators under strain
-	Quality driven by model	Expensive
	Undesirable data artifacts	Low fidelity capture
	Bound by computational power	Lack of data diversity

The first step towards the analysis goals outlined above involves data collection. Data collection is often an unglamourous but critical task for machine learning based methods. Poor data collection strategies and lack of data set understanding can severely bias model output, limiting the utility of any conclusions drawn. As a result, the team conducted a thorough review of different data collection options and sought to understand the limits of each. Based on the goals of the analysis and the variety of data required for meaningful analysis, the authors identified two main families of data suitable for the work. A high level overview of this analysis is shown in Table 1.

The first type of data reviewed, called simulation data, is generated from computer models of an environment. This data is generated often using programs like AFSIM or other virtual modeling methods. This method of generating data provides several benefits as shown in Table 1. The first is flexibility. Since the relative cost of running a scenario is low, a wide variety of different datasets can be collected. This ensures that data used for making battle management recommendations contains the necessary variety of information. In addition, another positive of simulation is the ability to collect more complete and structured datasets. The computer mediated environment means that greater control is kept over how information is produced and recorded. This can make the data more complete and easier to analyze. While simulation is a powerful resource, it does not come without its drawbacks. One of the biggest being that data utility can be highly dependent on simulation fidelity. Often simulations are only as accurate as the models underlying them. Specifically, complex behaviors like pilot tactics or interactions between team members can be challenging to accurately model. Therefore, nuances in the data could be lacking, negatively impacting any insights or recommendations the data might provide. In addition, in order to run complex scenarios intensive computational resources might be required. As a result, often there is a tradeoff required between fidelity and time.

The second type of data reviewed comes from Large Force Exercises (LFE). LFE are an integral part of preparing warfighters for cooperation between different platforms. During these exercises numerous types of aircraft and operator roles have an opportunity to practice operating as a cohesive unit. Often, during these exercises various types of data are collected about actions taken and their outcomes. However, the types of data collected depend heavily on the LFE and operators involved. This variety and non-standardization can make using the data challenging. While challenging to use, the data is as close to real operational conditions as one can get. This means this data contains a wealth of information that is important for any battle management analytics. However, LFE data still contains artifacts and drawbacks that need to be considered. For example, things like called kill shots and operational floors are not representative of an actual fight and need to be addressed accordingly during analysis. In addition, due to the cost of the exercises, LFE data can only contain so many types of situations and configurations. This limitation leads to unbalanced data sets that might not contain the necessary examples required for battle management analysis. Lastly, one of the biggest drawbacks of the LFE data is the types of data collected. Often data from these types of exercises is much lower fidelity than simulator data. Data collection relies on a mix of automated and hand recorded systems. The hand recorded data can be challenging to digitize in quantities required for large scale analytics. In addition, often the automated systems only record specific pieces of information like global position of an aircraft and not other pieces like sensor usage. These missing pieces of information along with the inability to model interactions with new Mosaic components can make creating a robust battle management aid challenging.

Ultimately, the type of insight the battle manager requires will dictate the type of data collected and the level of detail required. In a perfect world the data collected would be a one to one mapping to the operational environment. However, due to collection limitations and tradeoffs this type of data is not possible to gather. As a result, to provide battle management insights to the operator, careful planning and curation of available data is required. Understanding the capabilities and limitations of different data categories can ensure data is appropriate for the analysis being conducted. After considering the benefits and limitations of each data category the authors selected simulation as the data

generation mechanism. However, the authors also consulted extensively with experienced operators to ensure that any insights developed considered limitations of the data.

Figure 1. DCA Scenario

Once the choice to use simulated data was made the authors set to work developing a scenario to run. The authors selected AFSIM to build the scenario. AFSIM is one of the standard high fidelity modeling tools the United States government uses to build and test operational concepts. It's mixture of high fidelity modeling and extensive data collection capability made it a suitable choice for the work. Once the simulation tool was selected the next step involved selecting a scenario type to run. The authors considered several traditional military campaign types, but ultimately selected a DCA scenario, shown in Figure 1, as the initial test case based on expert input and advise. The Air Force Doctrine defines a Counter Air scenario as: "The counterair mission integrated offensive and defensive operations to attain and maintain a desired degree of control of the air and of protection by neutralizing or destroying enemy aircraft and missiles, along with threats to air operations from other domains (United States Air Force, 2019)." The defensive aspect of the counter air mission focuses on protecting blue force targets shown as a triangle from an attacking red force. This scenario was selected because it was complex enough to provide interesting analysis avenues, but manageable enough to allow expert operators to interpret analytics results for clarity during this initial stage of development.

Scenario Level Metrics Development

Military combat simulations can be complex and generate large amounts of data. With so much data and interaction between entities analytical avenues looking at different factors like win/loss, quality of tactics, quality of maneuvers, etc. can be numerous. In order to initially bound the problem, the authors decided to focus on determining what run level metrics impacted blue's ability to win. To do this, different slices of data were encapsulated at an aggregated level using scenario or run level metrics. These metrics provide general insight into various aspects of each scenario run which can be analyzed in aggregate to find common trends and correlations. The authors focused on developing scenario level metrics in three areas. The first being high level performance metrics describing mission outcome and team actions. The second focuses on metrics associated with the quality of actions taken in the scenario like shot quality, weapons utilization, or sensor coverage. The third and final category of metrics developed looked at force compositions between blue and red. The process of developing these metrics is briefly described below.

To develop metrics describing scenario performance the authors began by using the Universal Joint Task List (UJTL). This document details various metrics identified by the military as important measures of success within combat scenarios. While these metrics served as a starting point, they are unfortunately not sufficiently detailed to capture nuanced performance impacts as mentioned in background literature. As a result, additional metrics were included to increase information fidelity at the run level. Metrics were derived using different sources of raw information provided by AFSIM DCA scenario outcome, platform or weapon loadout logs, and shot log results. Some of these metrics were easier to derive than others. The first category of team success based metrics developed from these logs like Loss Exchange Ratio (measure of terminated enemies vs. terminated friendlies), Mission Success (binary measure of whether the mission was successfully executed), and Ratio Shots on Target Hit (percentage of all shots fired on the defended target which successfully hit and damaged the target) were straight forward to calculate from logs. The second category looking at action quality were derived from platform level actions throughout a run. These were more challenging to calculate since time series information needed to be rolled up into a single scenario level number. Metrics developed for action quality included the weapon shot log performance broken down by platform side, type, weapon type, and result. The shot results include things like % of killed intended target, hit intended target, missed, intercepted, and parent platform destroyed while weapon still attached. The breakdown by side & type yields higher fidelity insight at the run level (e.g. % of blue 4G fighter jets killed, % of red bombers killed), providing the battle manager a more nuanced picture of how each platform performs. The third and final category of metrics looked force composition. These metrics described the types of force packages red and blue used in a scenario. These metrics included information like Ratio of Warfighters (number of red fighter aircraft divided by number of blue fighter aircraft), number of platforms for a given side & type, and the percent of platforms killed for a given side & type.

In total, after developing the different categories of metrics there were around 430 different numerical values describing reach run. With this level of fidelity, the problem becomes a combinatorial explosion yielding hundreds of metrics for every run. This level of data creates review and analysis challenges. To combat this and help identify key drivers, a machine learning method was developed to quickly identify what metrics contribute to mission success. Using this insight, decisions can be made to optimize mission performance in real scenarios and ensure that warfighters achieve success.

Random Forest Analysis Method

The large number of metrics developed tells an extensive story about how the scenario unfolds. However, with so many parameters it is challenging to know what ones play a role determining a scenario's outcome. Expert insight is one way to help reduce the number of metrics to look at, however, as evidenced in the background literature there is limited data driven analysis to support these insights. This could lead to missing insights or being steered by expert blindness. In addition, current analytical methods are ill-suited to deal with the combinatorial explosion of variables or they make overly simplistic assumptions about underlying data distributions (Lewinson, 2019; Furrer, 2019). This lack of tools suited to help conduct scale analysis for more complex multifactor simulations puts the battle manager and analytics at a disadvantage because insight generally comes from overly simplified problems designed to fit available analysis tools. Lastly, as new warfighting methods like Mosaic are introduced limited historical knowledge will be available to leverage. As a result, a more data driven method is required to determine at the scale of big data important factors for the battle manager. Thankfully data driven machine learning can help solve this problem in a flexible and scalable way. This section describes the creation of a machine learning classifier for determining important metrics indicating to the battle manager what to factor into their decision making process to maximize win potential.

The goal of the machine learning classifier development was to help pinpoint the key determining features of the scenario. The reason machine learning was selected for feature importance identification rather than traditional statistical means is because: 1) machine learning methods are purely data driven 2) data driven methods make less underlying assumptions about data distributions 3) machine learning analysis is scalable (Furrer, 2019; Lewinson, 2019). Looking at the constraints of explain ability in the form of feature importance the team selected Random Forest classifiers as their ML method. Random Forest classifiers are known as an ensemble method of machine learning. Meaning they fit many models to the data and decide based on the aggregate consensus. This proves powerful because it allows the method to be more fault tolerant when dealing with messy, incomplete, nosy, or biased data than other methods. In addition, the method due to using aggregation is less prone to a concept in machine learning called overfitting where the model too closely fits the data and is not generalizable outside of limited realm where it was trained. Another attractive benefit of Random Forest is that it allows the user to see how important each feature is when making a classification determination. This combination of robustness and explainablity is what led the team to select the model.

Figure 2 shows an overview of this analysis process. The process starts by generating data in AFSIM for the DCA scenario by randomizing different key scenario parameters like number of fighters and weapon loadouts. From these runs a set of 430 metrics described above were calculated for each scenario run. These numeric values of metrics like loss exchange ratio are then fed into a Random Forest classifier. This classifier makes a win or loss prediction based on the unique scenario metrics. From here, based on the importance of each feature when making the win/loss decision, the impact of each

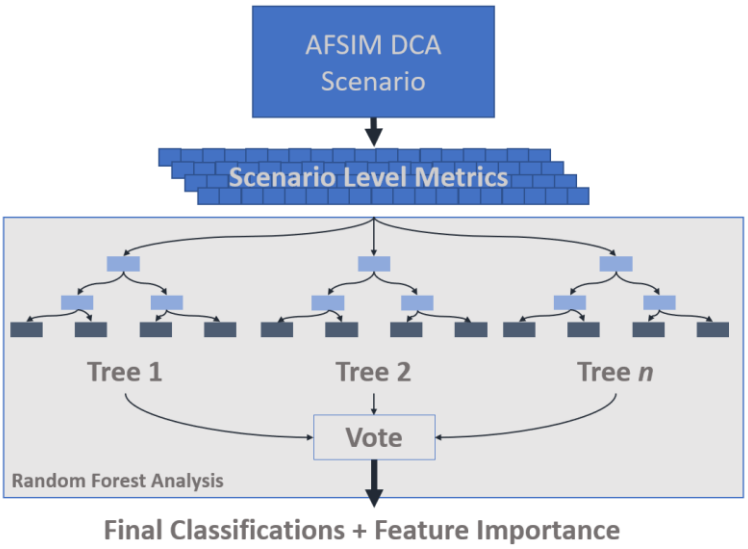

Figure 2. Random Forest Analysis Process

metric can be gauged and selected for further analysis to develop actionable insights for the battle manager. Further analysis after key drivers have been identified can use a multitude of techniques from traditional statistical hypothesis testing to graphical visualizations like box plots. The goal of the Random Forest is to help guild the analysis in the right direction after a large number of metrics are calculated. The Results and Discussion section below walks through several sample use cases that illustrate how this process can play out in practice to help provide battle managers with actionable insights.

RESULTS AND DISCUSSION

The goal of the analysis is to determine if blue won the match and if not why. Answering these questions ultimately will help determine strategies that the battle manager can employ to allocate resources most effectively to accomplish the mission. In order to start answering these questions, however, a sufficient data set is required to capture the variation of parameters in the scenario. Data set generation was carried out using the AFSIM DCA scenario described above. In total four-thousand runs were conducted with different randomized parameters. Table 2 shows the randomized variable names, types, and bounds. To simplify analysis the number of weapons assigned was consistent between a blue or a red fighter type for each run. Meaning all 4th and 5th generation fighters for a side were given the same randomized breakdown of weapons for a run. This simplified initial proof of concept analysis. In addition, the random assignment did not consider maximum limits on loadout configurations for each type of weapon by platform. This type of constrained assignment is saved for future work as the analysis fidelity is scaled.

Table 2. Randomized Variable Bounds

Variable	Variable Type	Values/Bounds
Blue 5G/4G Fighter Logic	Categorical	Low vs Medium Aggression
Red 5G/4G Fighter Logic	Categorical	Low vs Medium Aggression
Number Blue 5G/4G Fighter	Integer	[0 – 16]
Number Red 5G/4G Fighter	Integer	[0 – 16]
Blue - # MRAAM per 5G/4G	Integer	[0 – 4]
Blue - # SRAAM per 5G/4G	Integer	[0 – 4]
Red - # MRAAM per 5G/4G	Integer	[0 – 2]
Red - # SRAAM per 5G/4G	Integer	[0 – 4]
Red - # LACM per 5G/4G	Integer	[0 – 2]

Once these parameters were set and data generated, the first analysis step involved looking at the overall winning percentages for blue using different logic types or 'strategies' that could be selected by the battle manager. Table 3 shows the win rate for each blue aggression logic against the different red aggression logic. Results show that the low blue aggression logic seems to slightly outperform the medium aggression logic. However, from this high of a level determining why blue logic seems to perform better is not possible. Also, these high level metrics do not answer the second guiding question of if blue does not win, why are they losing? In order to help answer these questions, finer grained levels of performance and behavior description are required as suggested by background literature.

While this initial scenario only randomized force composition and weapon loadout component of a DCA scenario, it still contributes many metrics impacting mission outcome. To begin describing how each of these components impact the outcome a list of four hundred plus metrics where developed. Metrics developed focused on aspects such as shot effectiveness against different platforms or logic types. Realistically, some of these metrics may not have a critical impact on scenario outcome especially for a more straightforward scenario like DCA. However, as new types of warfighting strategies are introduced less insight will be available into what factors are important for the battle manager. As a result, methods are required to help determine at scale what the key drivers of mission outcome are. Traditional design of experiment, significance tests, and correlations break down when the number of factors is large. This is why past literature focuses on well-defined problems

Table 3. Blue Win Rate by Logic Type

Blue Aggression Logic	Red Aggression Logic	Blue Win Rate
All	All	69%
Low	All	71%
Medium	All	66%
Low	Low	74%
Low	Medium	67%
Medium	Low	66%
Medium	Medium	63%

Table 4. Use Case One Feature Importance – Force Structure

Metric Name	Analysis One	Analysis Two	Analysis Three
Num_Blue_4G_Fighter	0.020	0.024	0.059
Num_Blue_5G_Fighter	0.073	0.610	---
Num_Red_4G_Fighter	0.033	---	---
Num_Red_5G_Fighter	0.143	---	---
Num_MRAAM_Per_Blue_4G_Fighter	0.009	0.015	0.039
Num_SRAAM_Per_Blue_4G_Fighter	0.012	0.030	0.077
Num_Blue_4G_Fighter_MRAAM	0.023	0.037	0.093
Num_Blue_4G_Fighter_SRAAM	0.025	0.050	0.131
Num_MRAAM_Per_Blue_5G_Fighter	0.009	0.016	0.036
Num_SRAAM_Per_Blue_5G_Fighter	0.012	0.031	0.073
Num_Blue_5G_Fighter_MRAAM	0.022	0.034	0.094
Num_Blue_5G_Fighter_SRAAM	0.025	0.052	0.128
Num_Red_4G_Fighter_LACM	0.094	---	---
Num_LACM_Per_Red_5G_Fighter	0.042	---	---
Num_Red_5G_Fighter_LACM	0.076	---	---
Classification Accuracy	0.843	0.636	0.670

and parameter adjustments to make analysis more manageable. This however sacrifices realism and the potential generalizability of results, hampering the battle managers ability to apply them in real world scenarios.

To combat this problem the authors developed a method relying on data driven machine learning theory to help determine the key metrics that determine whether blue wins or loses a scenario. The result of this analysis is described below. Since data analysis for a scenario like DCA can branch into many directions, the authors selected several use cases to help guide the result presentation. The first use case aims to help the battle manager decide what the most important force configuration parameters, like number of different aircraft or loadouts, are. The second use case looks at helping the battle manager determine how their choice of aggression strategy influences key mission outcome metrics. The third and final use case looks at determining what measures of platform performance are the most important for determining mission outcome.

Starting with use case one, the goal is to help the battle manager determine the structure of the force they are using in the DCA scenario. To help answer that question only the parameters known before the battle are used to feed the Random Forest classifier. The first analysis case shown in Table 4 looks at using the parameters known before mission start with perfect knowledge of both red and blue. Using this perfect knowledge, the classifier can predict the outcome of the scenario over eighty percent of the time.

Looking at the feature importance for each variable, the initial condition most predictive of blue winning or losing is the number of red 5G fighters. This indicates that this factor is an important consideration for the battle manager to pay attention to when preparing for a mission. Contextualizing this result indicates that from a battle manger perspective, red's 5th generation fighter stealth likely plays a role helping conduct a mission. This could possibly be countered by the battle manager by deploying more sensor coverage or enhanced radar. While this knowledge of red's 5th generation fighters impactfulness is helpful, it is of limited tactical use since rarely are adversaries force structures known ahead of time. As a result, the next two sets of features provided to the

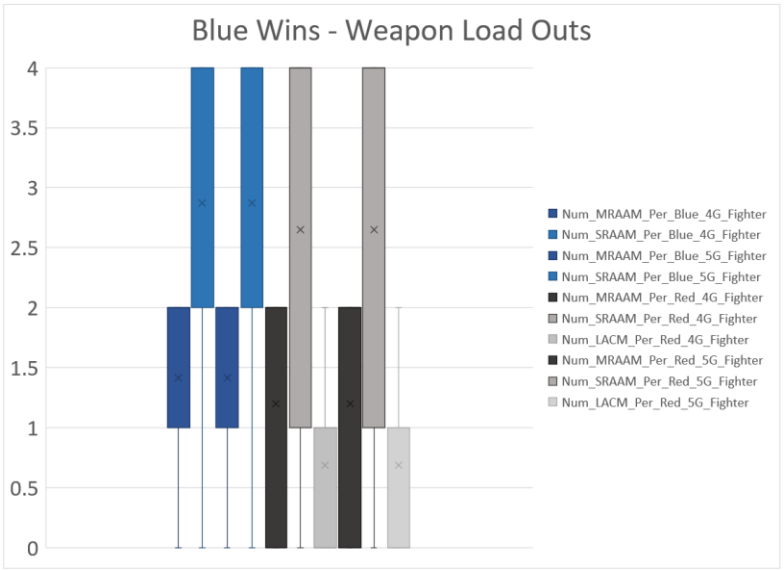

Figure 3. Blue Win Weapon Load Outs

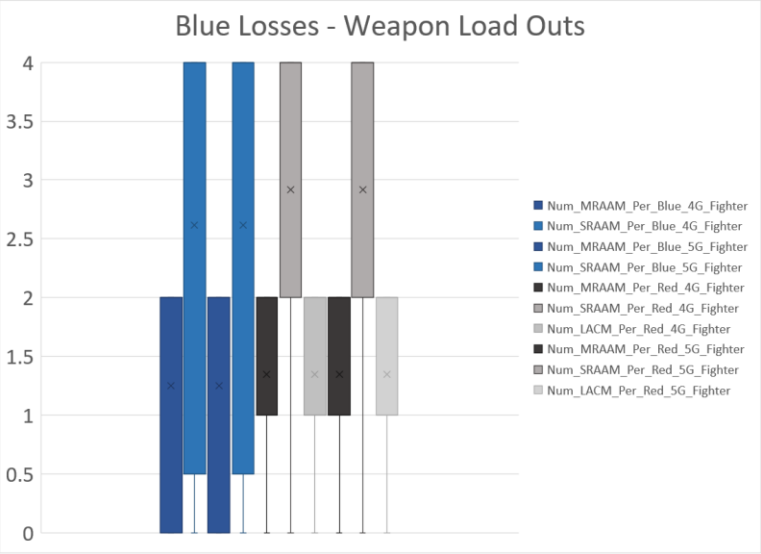

Figure 4. Blue Losses Weapon Load Outs

Random Forest classifier look at parameters only blue can control and know before mission start.

Looking at analysis case two, the number of blue 5G fighters dominates the Random Forest feature importance. However, the classification accuracy drops into the sixty percent range. This indicates that the classifier is having a harder time determining the outcome of the scenario from this more limited subset of information. This is to be expected since reducing the information provided to the classifier gives it a less complete picture of the initial conditions. This reduction in accuracy also suggests that as expected, the red force package impacts blue's chances of winning. While this reduction in classifier accuracy is unfortunate, it does not mean that the method unvaluable. In fact, if the battle manager can use the insight provided to select a better force structure using nothing but initial configuration settings to increase their chances of winning a conflict this is advantageous tactically. The last feature set shown in Table 4 removes the blue 5th generation fighter from the features provided to the Random Forest prediction algorithm. This was done to allow other salient factors to be used by the classifier. Removing the dominating feature slightly increases the classification accuracy. However, most importantly it shows that the number of short range air-to-air missiles (SRAAMs), are the most important factor for determining a blue win or loss.

One key feature to note about the Random Forest analysis is that it only shows what features are important. It does not inform how they are important. Contextualized, it does not tell the battle manager whether to increase or decrease the number of SRAAMs on platforms to provide a better chance of winning. In order to answer that question, auxiliary analysis is needed. This analysis can be performed in many ways. The key is though that the Random Forest helps sort through metrics to indicate what ones require further analysis. For this use case visual inspection of box plots was used to help the battle manager determine assignment of SRAAMs. Figure 3 shows the boxplot of weapon load outs for blue wins. Figure 4 shows the load outs of blue and red fighters for blue losses. Upon visual inspection it's clear that in general more SRAAMs are assigned to blue platforms when wins occur. This gives the battle manger concrete insight into how to assign SRAAMs to their platforms at mission outset to increase their chances of winning.

Table 5. Use Case Two Random Forest Feature Importance

Metric Name	Force Structure Analysis All	Force Structure Analysis Low	Force Structure Analysis Medium
Num_Blue_4G_Fighter	0.059	0.062	0.067
Num_Blue_5G_Fighter	---	---	---
Num_MRAAM_Per_Blue_4G_Fighter	0.039	0.042	0.047
Num_SRAAM_Per_Blue_4G_Fighter	0.077	0.064	0.075
Num_Blue_4G_Fighter_MRAAM	0.093	0.111	0.094
Num_Blue_4G_Fighter_SRAAM	0.131	0.121	0.130
Num_MRAAM_Per_Blue_5G_Fighter	0.036	0.040	0.040
Num_SRAAM_Per_Blue_5G_Fighter	0.073	0.071	0.078
Num_Blue_5G_Fighter_MRAAM	0.094	0.088	0.084
Num_Blue_5G_Fighter_SRAAM	0.128	0.127	0.124
Classification Accuracy	0.670	0.731	0.556

In addition, this could indicate to the battle manager that SRAAMs should be prioritized over other types of missiles when loading out platforms. Although, follow up simulations would have to be conducted to determine the point of diminishing return. Lastly, it's worth noting that the box plots show some differences in other weapon distributions for red and blue platforms. However, from the Random Forest analysis it's evident that these differences are not key factors necessitating attention. This shows the power of the machine learning based analysis to help identify only key features, thus saving time and resources by directing analysis towards worthwhile avenues.

The second use case aims to help the battle manger see how different logic methods, low aggression vs medium aggression, impact force selection methodology. Table 5 shows the results of the Radom Forest analysis for the different logic aggressiveness levels. Results indicate that the SRAAMs are still the most important factor. However, looking at the classification accuracy indicates that when the battle manager selects the medium aggressive logic the outcome of the scenario is harder for the Random Forest classifier to predict using only initial force configuration. This indicates that there is something else influencing blue's ability to win when selecting medium levels of aggression. Results like this suggest more metrics are required to capture the nuance of using this control type. After talking to a former pilot, sensor coverage is believed to be one of the missing elements not captured in current metrics. To test this theory, however, metrics quantifying sensor coverage throughout the scenario are needed. This result shows the method can be used not only to help indicate key factors, but also to help develop new metrics when current ones do not describe the performance. New metric development could be especially helpful when less understood methods of warfighting are introduced like new Mosaic concepts.

Table 6. Use Case Three Random Forest Feature Importance

Metric Name	Blue Low Logic	Blue Medium Logic	All
Force_Ratio	0.100	0.111	0.115
Loss_Exchange_Ratio	0.061	0.098	0.085
Ratio_Red_4G_Fighter_LACM_Deployed	0.018	0.022	0.022
Ratio_Red_5G_Fighter_LACM_Deployed	0.015	0.007	0.013
Classification_Accuracy	0.828	0.798	0.852

The final use case three aims to help the battle manager determine what measures of performance within the scenario determine win vs loss. This will help the battle manger best identify areas of improvement or metrics to watch while the battle unfolds. Table 6 shows the top four features for each analysis case. These top four features were picked out of a field of over four hundred metrics developed for the scenario. From the importance's one can see that the force ratio is consistently the most important indicator of blue win vs loss. The force ratio represents the number of blue fighters divided by the number of red. A positive force ratio means that blue has the advantage. Again, while the Random Forest tells what is important, it does not indicate how to drive the metric to impact mission success.

In order to determine how the Force Ratio should be adjusted, again, auxiliary methods of analysis are required. Figure 5 shows the box plot comparison of wining vs losing force ratios. From the plots it's apparent that blue wins more often when they are out matching their opponents. However, the box plots show that for the winning runs there is a subset performing well even though they are outmatched by red. This fact suggests that there might be force configurations that perform well when outmatched. This is an aveue for future analysis.

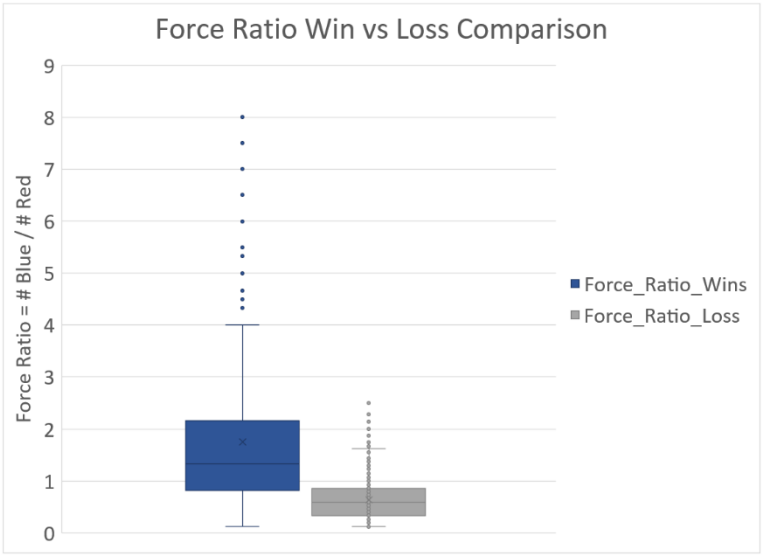

Figure 5. Force Ratio Box Plot

Overall, the results presented demonstrate the Random Forest feature imporance method's utility at helping determine imporant key features driving scenario outcomes. This insight can help the battle

manager determine what metrics and features to focus on in an otherwise complex scenario. While the scenario presented made a number of assumptions to keep the analysis managable, it showed it's ability to pinpoint key deciding factors. This is the first step towards validating the methods use for more complex scenarios where the driving features are not known like in new Mosaic warfare conects. Ultimatley, the method helps demonstrate how the warfighting community can use data driven machine learning methods to help quantify and benchmark performance in a flexible robust way.

CONCLUSION AND FUTURE WORK

The United States military is increasingly exploring the use of artificially intelligent agents. This emerging AI technology is a key component in future Mosaic warfare strategies outlined by DARPA. However, to ensure seamless orchestration of missions, great care needs to be taken to mesh battle managers and AI agents. To do this involves helping battle managers understand how impactful different decisions and strategies are. Unfortunately, generating this level of insight for complex battle scenarios is challenging due to data collection, metrics development, and key performance indicator identification challenges. This work described development of a DCA scenario used to generate data for key performance indicator identification. Over 430 metrics were developed to describe performance within the scenario. These metrics where used to help answer the questions: 1) does blue win? and 2) if not why? A machine learning based Random Forest classifier was used to identify the most impactful performance indicators. From here auxiliary analysis using box plots helped determine which way to drive different measures, such as equipping the blue fighters with more short range missiles. Analysis showed that the method can help battle managers identify what metrics to use to make decisions like what force structure to use, what weapons to use on those platforms, and what performance metrics to use for helping increase mission success.

Moving forward the work will focus on increasing fidelity of the simulation, increasing the types of scenarios studied, and incorporating different types of trained AI agent logic into the analysis. Increasing simulation fidelity task will focus on adding more variety to the elements that are randomized. Currently only the number of air platforms and their weapon loadouts are being manipulated. However, this is only a component of the DCA scenario. Other platforms like surface to air missiles (SAM) and bombers also contribute to mission success. Adding in the ability to investigate these other elements will help increase the value of the insights provided to the battle manager. In addition, DCA is only one of many types of military missions. Beginning to look at these moving forward and applying the same analysis method will help provide helpful insight in other areas of the battle space. Lastly, looking at how different AI agents interact with the battle manager and the scenario will help execute the Mosaic vision for future warfighting in a human centric way.

REFERENCES

Abdelaal, M. A. (2016). *METHODOLOGY FOR DETERMINING CRITICAL DECISION POINTS THROUGH ANALYSIS OF WARGAME DATA.*

Alkire, B., Kim, Y., Berry, M., Blancett, D., Inamdar, N., Lingel, S., . . . Williams, W. A. (2020). *Enhancing Assessments of Space Mission Assurance.* RAND.

Almujahed, S., Ongor, N., Tigmo, J., & Sagoo, N. (2013). *Sports Analytics - Designing A Decision-Support Tool for Game Analysis Using Big Data.*

Biao, G., & Blangiardo, M. A. (2010). Bayesian hierarchical model for the prediction of football results. *Journal of Applied Statistics, 27*(2), 253-264.

Bryan, C., Patt, D., & Schramn, H. (2020). *MOSAIC WARFARE MOSAIC WARFARE EXPLOITING ARTIFICIAL INTELLIGENCE AND EXPLOITING ARTIFICIAL INTELLIGENCE AND AUTONOMOUS SYSTEMS TO IMPLEMENT AUTONOMOUS SYSTEMS TO IMPLEMENT DECISION-CENTRIC OPERATIONS DECISION-CENTRIC OPERATIONS.* Center for Strategic and Budgetary Assessments.

Cea, S., Duran, G., Guajardo, M., Saure, D., Siebert, J., & Zamorano, G. (2020). An analytics approach to the FIFA ranking procedure and the World Cup final draw. *Annals of Operations Research, 286*, 119-146.

Cho, H., Park, H., Kim, C.-Y., & Kim, K.-J. (2017). Investigation of the Effect of 'Fog of War' in the Prediction of StarCraft Strategy Using Machine Learning. *Computers in Entertainment, 1*(2).

Congressional Research Service. (2020). *Artificial Intelligence and National Security.* United States Government.

Connors, C. (2015). *AGENT-BASED MODELING METHODOLOGY FOR ANALYZING WEAPONS SYSTEMS.*

DARPA. (2018). *DARPA Tiles Together a Vision of Mosaic Warfare.* Retrieved from https://www.darpa.mil/work-with-us/darpa-tiles-together-a-vision-of-mosiac-warfare

Deptula, L., & Penney, U. (2019). *Mosaic Warfare.* Retrieved from Air Force Magazine: https:www.airforcemag.com-article-mosaic-warfare

Do Nascimento, F. F., Da Costa, I. B., Da Costa Melo, A. S., & Marinho, L. B. (2017). Profiling successful team behaviors in league of Legends. *WebMedia 2017 - Proceedings of the 23rd Brazillian Symposium on Multimedia and the Web* (pp. 261-268). Association for Computing Machinery, Inc.

Drachen, A., Seif El-Nasr, M., & Canossa, A. (2013). Game Analytics - The Basics. In *Game Analytics* (pp. 13-40).

Furrer, D. (2019). *Interpretability Linear Models vs. Random Forests by Jeremy Howard*. Retrieved from Medium: https://fulowa.medium.com/interpretability-linear-models-vs-random-forests-by-jeremy-howard-30733508de2b

Gordon, S. E. (2018). *A STOCHASTIC AGENT APPROACH (SAA) FOR MISSION EFFECTIVENESS.* Thesis: Georgia Institute of Technology.

Grana, J. (2021). *FINDINGS ON MOSAIC WARFARE FROM A COLONEL BLOTTO GAME.* NIELSEN BOOKDATA.

Grooms, G. B. (2019). *ARTIFICIAL INTELLIGENCE APPLICATIONS FOR AUTOMATED BATTLE MANAGEMENT AIDS IN FUTURE MILITARY ENDEAVORS.* Thesis: Naval Post-Gradaute School.

Gulden, T. R., Lamb, J., Hagen, J., & O'Donoughue, N. A. (2021). *Modeling Rapidly Composable, Heterogeneous, and Fractionated Forces: Findings on Mosaic Warfare from an Agent-Based Model.* RAND.

Hanlon, N., Garcia, E., Casbeer, D., & Pachter, M. (2018). AFSIM Implementation and Simulation of the Active Target Defense Differential Game. *AIAA SciTech.* AIAA.

Heinze, C., Gross, S., & Pearce, A. (1999). Plan recognition in military simulation: Incorporating machine learning with intelligent agents. *Proceedings of IJCAI-99 Workshop on Team Behaviour and Plan Recognition*, (pp. 53-64).

Hodickỳ, J., Procházka, D., Baxa, F., Melichar,, J., Krejčík, M., Křížek, P., . . . Drozd, J. (2020). Computer assisted wargame for military capability-based planning. *Entropy, 22*(8).

Jensen, B., & Paschkewitz, J. (2019). *Mosaic Warfare: Small and Scalable are Beautiful.* Retrieved from War On the Rocks: https://warontherocks.com/2019/12/mosaic-warfare-small-and-scalable-are-beautiful/

Joseph, A., Fenton, N., & Neil, M. (2006). Predicting football results using Bayesian nets and other machine learning techniques. *Knowledge Based Systems*.

Jung, W., Marin, M., Lee, K., Rabelo, L., Lee, G., & Noh, D. (2019). A Proposed Methodology on Weapon Combat Effectiveness Analytics Using Big Data and Live, Virtual, or/and Constructive Simulation. *SAE Technical Papers* (pp. 357-374). SAE.

Lee, C. S., & Ramler, I. (2017). Identifying and evaluating successful non-meta strategies in league of legends. *ACM International Conference Proceeding Series.* Association for Computing Machinery.

Lewinson, E. (2019). *Explaining Feature Importance by example of a Random Forest*. Retrieved from Towards Data Science: https://towardsdatascience.com/explaining-feature-importance-by-example-of-a-random-forest-d9166011959e

Maymin, P. Z. (2021). Smart kills and worthless deaths: ESports analytics for League of Legends. *Journal of Quantitative Analysis in Sports, 17*(1), 11-27.

O'Donougue, N. A. (2021). *DISTRIBUTED KILL CHAINS: Drawing insights for mosaic warfare from the immune system and from the navy.* RAND Corporation.

Parmar, M. (2018). KABADDI Analytics. *Analytics*.

Ratnoo, A., & Shima, T. (2012). Guidance strategies against defended aerial targets. *Journal of Guidance, Control, and Dynamics*, 1059-1068.

Rushing, J., Tiller, J., Tanner, S., & Mcdowell, D. (2004). *Augmenting Wargame AI with Data Mining Technology.*

Sweetser, A., & Bexfield, J. (2020). *Lessons from the MORS Workshop on Advancing Campaign Analytics on JSTOR.* Military Operations Research Society.

United States Air Force. (2019). *AIR FORCE DOCTRINE PUBLICATION 3-01 COUNTERAIR OPERATIONS.* United States Government.

Yang, Y., Qin, T., & Lei, Y.-H. (2016). Real-time eSports Match Result Prediction. *30th Conference on Neural Information Processing Systems.*

Young, C. M., Luo, W., Gastin, P., Tran, J., & Dwyer, D. B. (2019). The relationship between match performance indicators and outcome in Australian Football. *Journal of Science and Medicine in Sport, 22*, 467-471.

