Naval Information Warfare Center

PACIFIC

TECHNICAL DOCUMENT 3414
FEBRUARY 2022

Improving the Digital Aviation Readiness Technology Engine (DARTE) with Temporal Pattern Attention Mechanisms and Hyper-Deep Ensembles

Dr. Benjamin A. Michlin
Josh A. Duclos
Gary R. Williams
Dean Lee
Dr. Andrew B. Sabater
Dr. Jamal T. Rorie
NIWC Pacific

Naval Information Warfare Center Pacific (NIWC Pacific)
San Diego, CA 92152-5001

This page is intentionally blank.

TECHNICAL DOCUMENT 3414
FEBRUARY 2022

Improving the Digital Aviation Readiness Technology Engine (DARTE) with Temporal Pattern Attention Mechanisms and Hyper-Deep Ensembles

Dr. Benjamin A. Michlin
Josh A. Duclos
Gary R. Williams
Dean Lee
Dr. Andrew B. Sabater
Dr. Jamal T. Rorie
NIWC Pacific

Administrative Notes:
This document was approved through the Release of Scientific and Technical Information (RSTI) process in July 2021 and formally published in the Defense Technical Information Center (DTIC) in February 2022.

NIWC Pacific
San Diego, CA 92152-5001

NIWC Pacific
San Diego, California 92152-5001

A. D. Gainer, CAPT, USN
Commanding Officer

W. R. Bonwit
Executive Director

ADMINISTRATIVE INFORMATION

The work described in this report was performed by the Enterprise Information Systems branch of the Business and Enterprise Information Systems (EIS) Competency, Naval Information Warfare Center Pacific (NIWC Pacific), San Diego, CA. The NIWC Pacific Naval Innovative Science and Engineering (NISE) Program provided funding for this Basic Applied Research project. Further assistance was provided by the Commander, Naval Air Forces (CNAF) Force Readiness Analytics Group (FRAG).

Released by
Donna Williamson, Division Head
Business and Enterprise Information Systems (EIS) Competency

Under authority of
Scott Crellin, Department Head
Command and Control and Enterprise Engineering

ACKNOWLEDGMENTS

Editor: MRM

ACRONYMS

AI	Artificial Intelligence
AMEX	Aviation Maintenance Experience
AMSRR	Aviation Maintenance Supply Readiness Reporting
CNAF	Commander, Naval Air Forces
CNN	Convolutional Neural Network
DARTE	Digital Aviation Readiness Technology Engine
FHRM	Flying Hour Resource Model
FMC	Fully Mission Capable
LSTM	Long Short-Term Memory
LSTMA	Attention Mechanism-Enhanced LSTM
MC	Mission Capable
NLP	Natural Language Processing
NMCM	Not Mission Capable for Maintenance
NMCS	Not Mission Capable for Supply
PMC	Partially Mission Capable
RNN	Recurrent Neural Network
TDA	Topological Data Analysis
TPA	Temporal Pattern Attention
XAI	Explainable Artificial Intelligence

This page is intentionally blank.

CONTENTS

FIGURES

TABLES

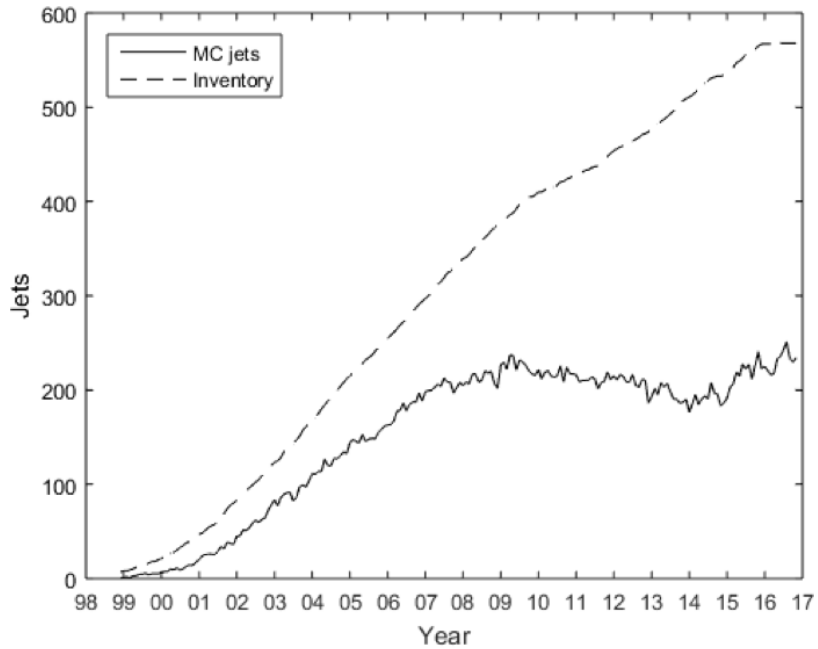
1. INTRODUCTION

The concept of “readiness” in Naval aviation relies on three key aspects: manning, training, and equipment. “Manning” refers to the available manpower including both the enlisted maintainers and pilots. Manning also includes the experience level and specialty of maintainers. “Training” refers to how well the pilots are trained and the quantity of flight hours executed. Finally, “equipment” refers to the necessary aircraft, supplies, etc. Even without a strict definition, it is immediately clear that impairing any of these three key factors would result in decreased readiness for a squadron. Two key metrics are used in the Naval aviation community to measure the readiness of a squadron: mission capable aircraft and flight hour execution. Specifically, a mission capable (MC) aircraft is one which meets or exceeds the minimum requirements to be operated and complete a mission.

Since their introduction in 1999, the total number of Naval FA-18 Super Hornets has increased nearly linearly to a current number of almost 600 aircraft at the time of this writing [1]. There are two variants of Super Hornets – the single-seat E and two-seat F. Additionally, the Navy still uses some of the older FA-18 single-seat C variant Hornets. The FA-18s can be in one of several readiness states at any given time: not mission capable for supply (NMCS), not mission capable for maintenance (NMCM), partially mission capable (PMC), or fully mission capable (FMC). Ideally, the number of mission capable aircraft would scale linearly with the number of total aircraft, but this is not the case. Figure 1 shows the number of FA-18s in inventory and the number of MC aircraft by year. In the late 1990s and early 2000s, the relationship between MC aircraft and inventory was as expected, but near 2007 there is a clear deviation and the number of MC aircraft plateaus. This is well-documented and readiness improvements are increasingly becoming the focus of leadership [4].

While the ability to accurately monitor and predict readiness is extremely important, the process non-trivial. Predictive power gives squadrons and decision makers the time and ability to reallocate resources, adjust manning levels, and make smarter decisions before an issue has occurred. By creating machine learning models to predict readiness, rather than intuition and human insights, there is also the possibility of discovering unintuitive insights into squadron operations.

The goal of the Digital Aviation Readiness Technology Engine (DARTE) is to predict FA-18 squadron readiness as measured by monthly MC and quarterly flight hour execution [6]. Further, DARTE provides the ability to monitor readiness as a squadron moves through a quarter. This is accomplished in two steps. First, a deep learning model is created to predict the mean number of mission capable aircraft for each squadron on a monthly basis. This model is then extended and used as an input for a second machine learning model which predicts the flight hour execution of a squadron over a quarter. Additionally, there is an explainable AI (XAI) engine [7] and statistical manning model [8] that accompany the MC model. The architecture of DARTE is shown in Figure 2.



The number of total jets in inventory increases nearly linearly while the number of MC jets plateaus in 2007 [5].

Figure 1. Number of FA-18 jets by year.

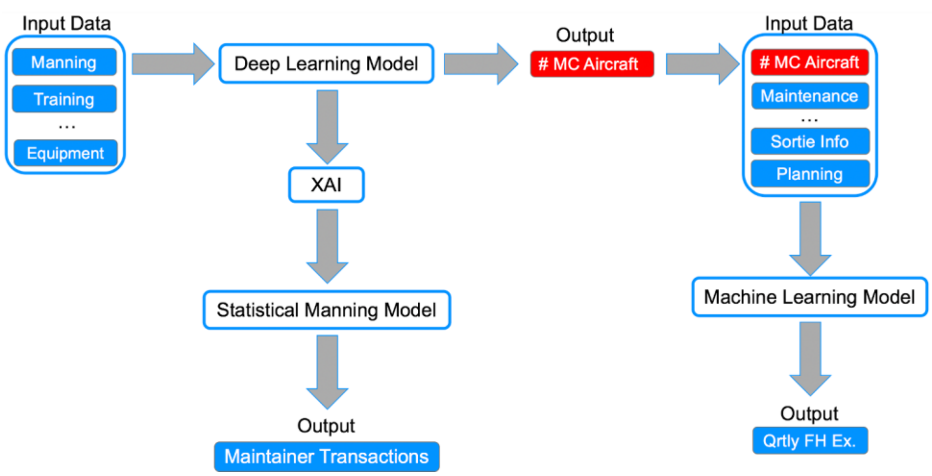


Figure 2. DARTE architecture.

This paper focuses on improving the foundational model of DARTE – the deep learning model predicting the number of mission capable aircraft that a Naval FA-18 squadron will have up to three months in advance. This model influences every other piece of DARTE, so it is crucial that the MC model is accurate, well understood, and robust.

This paper is organized as follows: Section 2 reviews the datasets used along with their conditioning and transformation, Section 3 discusses the creation and results of the base MC model, Section 4 shows the results of the final model, and Section 5 reviews the conclusions and discusses future work.

2. DATA

There are many datasets representing the pillars of readiness (man-train-equip) which are combined to generate a consolidated dataset. Several of these sources are discussed here, but a more in-depth discussion may be found in Reference 6. Manning data contains the officer and enlisted personnel information per squadron per month from the AIRFOR dataset. Additional manning and training data come from the Aviation Maintenance Experience report (AMEX) which contains the experience levels and rates of the aircraft maintainers on a monthly basis. Equipment data include the Aviation Maintenance Supply Readiness Reporting report (AMSRR), which contains daily information on squadron's Mission Capable aircraft (MC), and the Flying Hour Resource Model report (FHRM), which contains information on squadron phase and planned MC (MC entitlement). All raw data is aggregated to the squadron-month level. An overview of these datasets may be seen in Table 1.

Table 1. Most important data sources used for modeling and analysis.

Man	Train	Equipment
AIRFOR	AMSRR	Deckplate (Maint. Reports)
NAVFLIR (Aircrew)	NAVFLIR (Flightlegs, TMR)	AFAST
AMEX	SCIR	FHRM/ADW NAVAIR Reports

Since the only features common to all of the datasets are the squadron and date, these features are used to join the data (although the date ranges do not always overlap). There is an added complication in that some datasets record information on a monthly basis while others record daily, weekly, or on an ad-hoc basis. For those that record daily, weekly, or ad-hoc, aggregate monthly values are created including the mean value, standard deviation, max, min, sum, and count.

This page is intentionally blank.

3. ATTENTION MECHANISM-ENHANCED LONG SHORT-TERM MEMORY NETWORKS

All models and results shown are trained on 70% of the data, validated on 15% of the data, and a 15% holdout dataset is reserved for use in Section 4.

3.1 CHALLENGES ADDRESSED

Due to the complex nature of the data there are unique challenges that must be overcome during model creation. The data were collected over more than a 10-year period and include information for 35 VFA squadrons. Depending on the aggregation level, there are up to 3,000 features, many of which are noisy and/or unnecessary. Because of this, important features and patterns are difficult to uncover and are obfuscated. This is especially true for long-range patterns between features in time.

Since these long-range patterns in time exist and are extremely predictive, 6 months of historical data must be used to make an MC prediction for 3 months in the future. This magnifies the issue of noisy and/or unnecessary data 6-fold as a massive historical record is added to every sample. Although it may seem that traditional recurrent neural networks (RNNs) and more complex long short-term memory (LSTM) networks may be ideally suited for this type of problem, they suffer from the long-range dependency problem [9]. Furthermore, for LSTM networks, because recursive updates are performed in a Markov-like manner it is difficult, if not impossible, for an LSTM network to completely forget a feature-timestep [10]. This is also exacerbated by the long historical record needed for accurate predictions.

3.2 METHODOLOGY

To address the first two challenges, the noisy and/or unnecessary features, and the long-range dependency problem, attention mechanisms may be used to enhance the LSTM nodes. Attention mechanisms allow the network to focus on what is important and truly disregard all else. This mechanism is similar to how a “forget gate” works in a traditional LSTM, but attention mechanisms allow for more complicated and precise forgetting. This allows the model to ignore false patterns in the data and noisy and/or unnecessary features, prompting the network to perform a more robust analysis of what remains.

Figure 3 shows how an attention mechanism may be used to augment a convolution neural network (CNN) in order to fill in the underlined word in the caption. It is clear that the model is focusing on the people in order to locate the woman, and then looks for what she is throwing - a frisbee. The model also successfully ignores the grass and most of the background image. Figure 4 shows another example of an attention mechanism applied to an LSTM for time-series data. A dataset with three features is shown and the goal is to predict the future cost of gasoline given the price of crude oil and lumber. As expected, gasoline and crude-oil are highly correlated, and lumber is an unnecessary feature that has a false correlation. A vanilla LSTM or RNN based model will actually perform better at predicting the price of gasoline if the lumber feature is removed, but it is unable to ignore or forget the feature on its own. However, the attention mechanism-enhanced long short-term memory network (LSTMA) is able to successfully ignore the lumber feature and make an accurate prediction. This type of attention mechanism is called a Temporal Pattern Attention (TPA) [10].

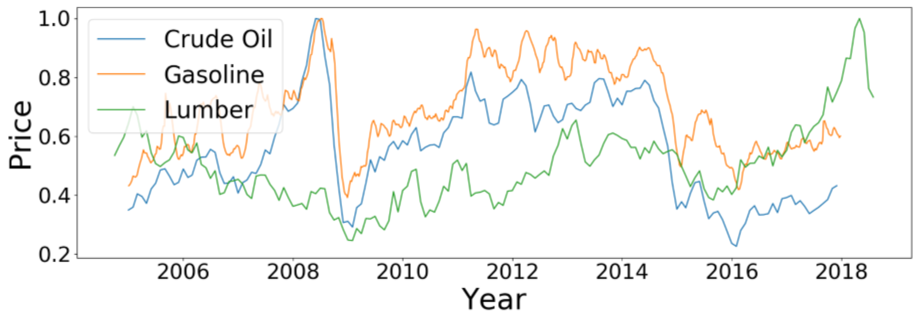
An attention mechanism as applied to a CNN. The highlighted areas show what the model is paying attention [11].

Figure 3. Attention Mechanisms for CNNs.

Temporal pattern attention mechanisms build on LSTM units by using a CNN to convolve the hidden states from every time step at once. This means that the importance of a pattern does not diminish with increasing distance – solving the long-range dependency problem. Also, TPAs have an additional activation function that allows for completely dropping and forgetting an unnecessary and/or noisy feature. Details on this attention mechanism may be found in Reference 10.

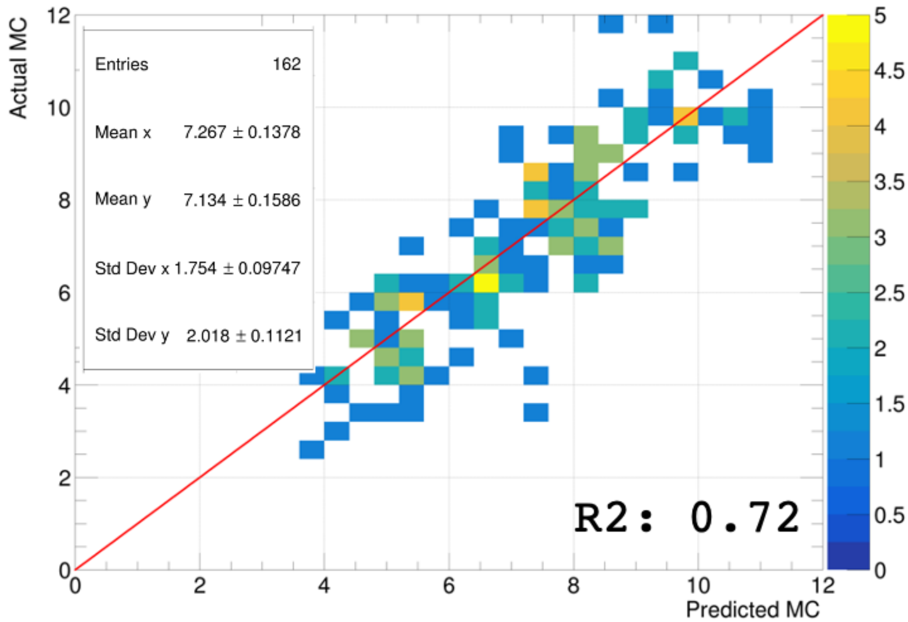
3.3 RESULTS

The results of implementing the LSTMA and predicting the number of mission capable aircraft that a Naval FA-18 squadron will have 3 months in advance are shown in Figure 5. This plot shows a 2D histogram of the actual vs predicted MC with a red line showing a perfect prediction. Figure 6 shows the same information in a different way with the prediction residual shown directly. The residual has a Gaussian fit with a χ^2 consistent with an excellent Gaussian fit, a mean nearly consistent with zero indicating very little bias in the model, and a sigma of slightly below 1. Because of the sigma, about 70% of the time MC predictions for 3 months in the future are accurate within 1 aircraft. In terms of performance metrics, this LSTMA model has an r^2 value of 0.72.



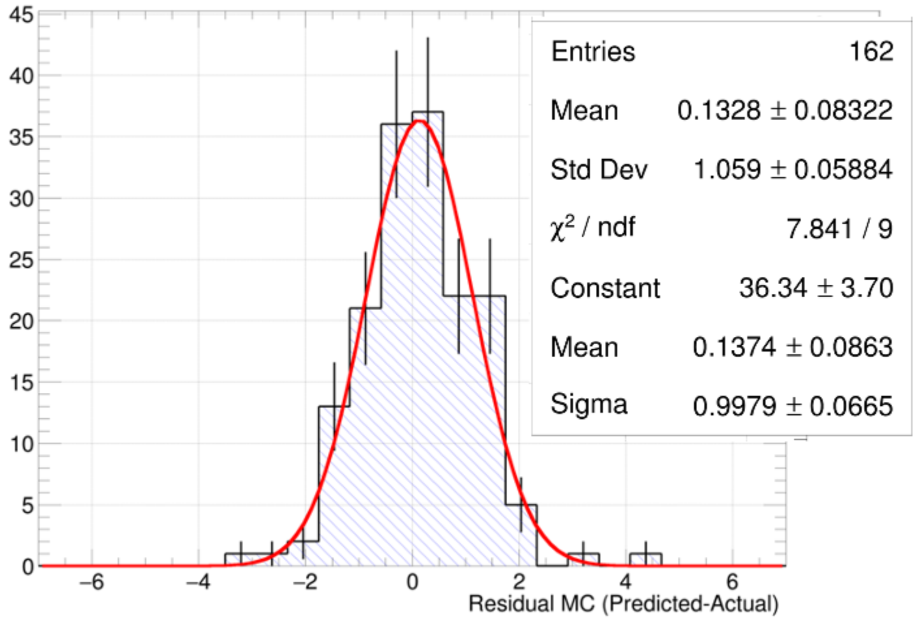
An attention mechanism is applied to an LSTM with time-series data. Lumber is an unnecessary feature with false correlations that hurt the performance of vanilla LSTMs and RNNs [10].

Figure 4. Temporal Pattern Attention.



A 2D histogram of the actual vs predicted MC for the LSTMA MC model. The red line shows a perfect prediction, and the model has an r-squared value of 0.72.

Figure 5. 2D Histogram for LSTMA Model.



The residual is fit to a Gaussian.

Figure 6. LSTMA Model Prediction Residual.

4. HYPER-DEEP ENSEMBLE

All models and results shown in this Section use the same training, validation, and test sets as the LSTMA in Section 3.

4.1 CHALLENGES ADDRESSED

In a final challenge, if predictions are to be trusted and acted upon it is required that there is a robust prediction uncertainty estimation. By leveraging prediction uncertainty estimation, decision makers may know how much an individual prediction should be trusted and acted upon to allow for better-informed decisions to be made.

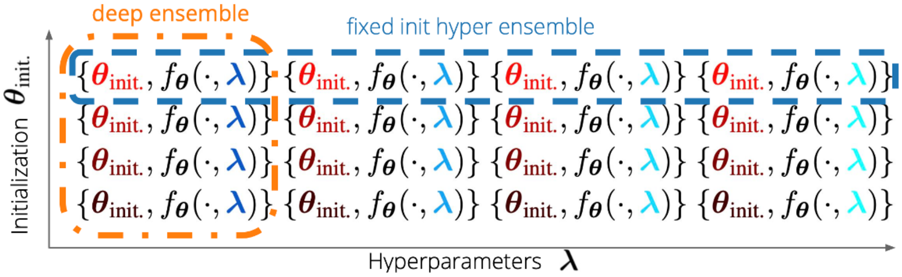
One way to improve performance and allow for prediction uncertainties to be calculated is to create an ensemble. This Section uses the LSTMA MC model as the basis for an ensemble.

4.2 METHODOLOGY

Deep ensembles were developed in 2017 to estimate uncertainties and improve model performance by running network architecture and hyperparameters with many different weight initializations [12]. This methodology relies on a non-convex loss space which leads to each trained model converging to a slightly different, but complementary, solution. The output of a deep ensemble is derived by combining the many model predictions as a mixture of Gaussian distributions. The uncertainty may then be obtained in a straightforward way by using sigma.

In addition to diversifying the weight initializations, in 2020 deep ensembles were improved by further varying the models in the ensemble by using different combinations of hyperparameters [13]. This adjustment resulted in improved performance and a more robust uncertainty estimation. Also, for a single model, the choice of hyperparameters for a production deep learning model is a non-trivial task. By diversifying the ensemble over hyperparameter combinations, the burden of choosing a single combination is alleviated.

To create a hyper-deep ensemble, a library of models must be created. First, many models are trained via a random search with both random weight initialization and hyperparameters. Then, a subset of these models is chosen by a greedy ensemble construction algorithm [14]. For each selected model, the hyperparameters are fixed and the model is retrained with many different weight initializations. Figure 7 shows the creation of this hyper-deep library of models where each column shows a deep ensemble wherein the hyperparameters are fixed and the initial weight initializations are changed. Each row shows a fixed weight initialization with different combinations of hyperparameters used. The library of models is then used for a final greedy ensemble construction algorithm. This final ensemble step completes the model, allowing predictions to be made and uncertainty to be calculated.

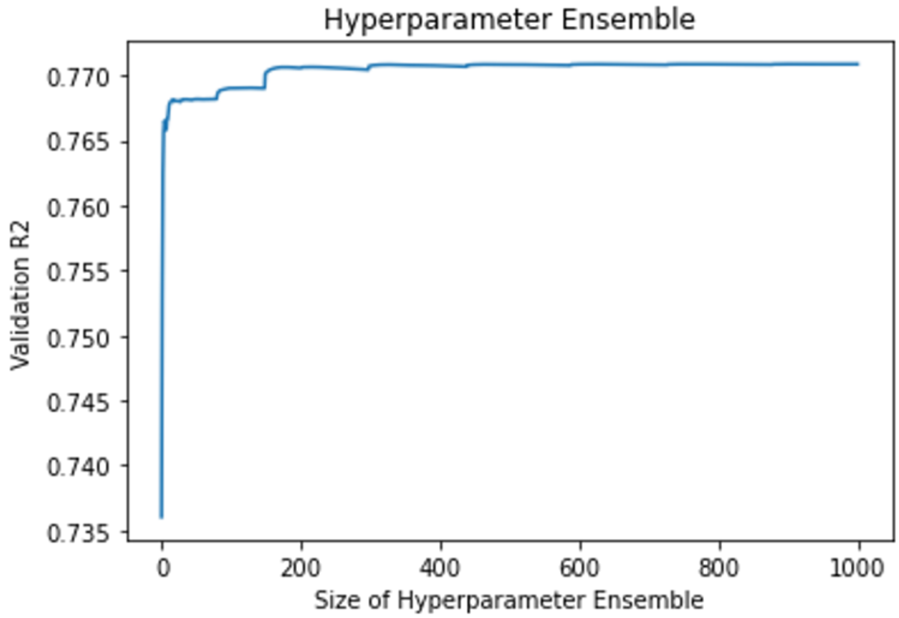


Each column shows a deep ensemble wherein the hyperparameters are fixed and the initial weight initializations are changed. Each row shows a fixed weight initialization with different combinationsof hyperparameters used [13].

Figure 7. Hyper-Deep Library.

4.3 RESULTS

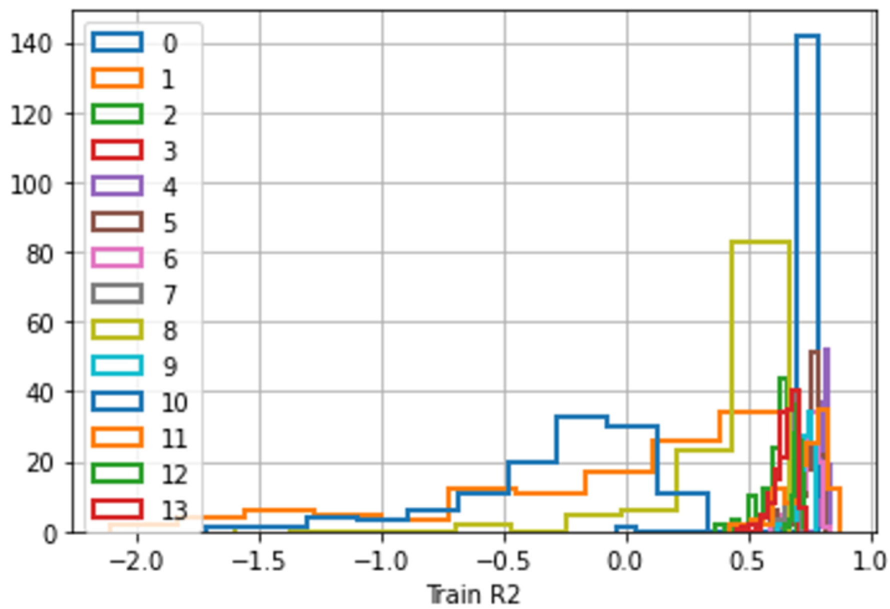
In the creation of the hyper-deep library, 231 LSTMA models predicting the mean number of mission capable aircraft that a Naval FA-18 squadron would have 3 months in advance were trained. Each model had a random combination of hyperparameters and weight initialization. Then, a subset of these models was chosen by a simple greedy ensemble construction algorithm. The results of this construction are shown in Figure 8. Here, the r-squared value of the validation set is shown versus the size of the constructed ensemble (where an ensemble prediction is considered as the mean value of all the predictions of models within the ensemble). The model performance increases from the single best model with an r-squared of 0.735 to 0.770 for the ensemble. The ensembles were created with replacement which means that not every model is chosen once, and this allows the ensemble to weight each constituent model individually and yield a more accurate result. For example, with an ensemble size of 1000 models, one model is chosen over 200 times and others are not chosen at all. In fact, only 14 different combinations of hyperparameters and initializations were selected for any ensemble. Of these 14 selected combinations, there are two loss functions, 8 with MAE and 6 with MSE, two optimizers, 3 with Adam and 11 with SGD, and 3 different network architectures.



Ensembles are created via a greedy construction algorithm with replacement. The performance increases from an r-squared of 0.735 to 0.770.

Figure 8. Hyperparameter Ensemble Construction.

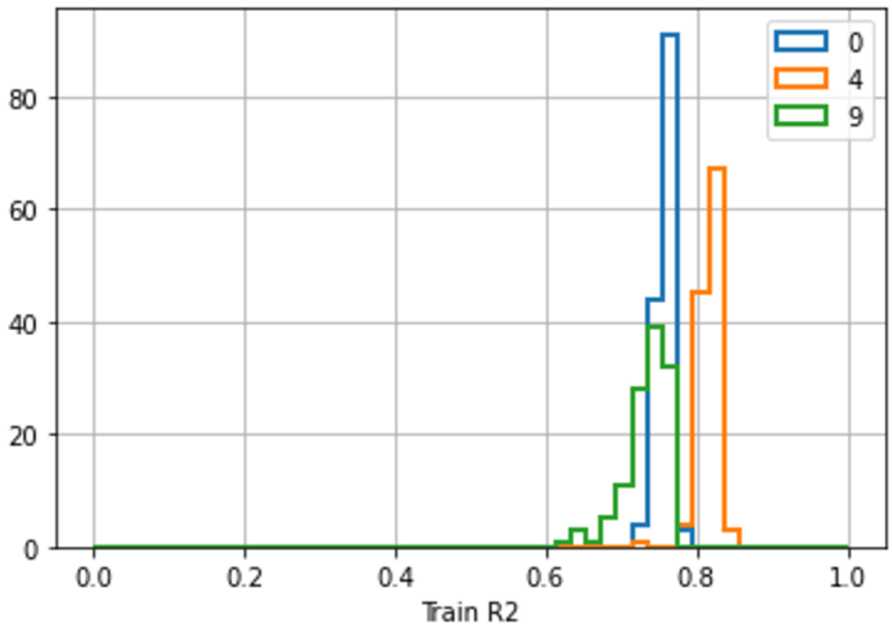
Following the methodology in Section 4.2 the next step is to then retrain each selected hyperparameter combination for many different weight initializations. Each of the 14 selected hyperparameter combinations was re-trained with 40-160 different weight initializations for a hyper-deep library of 1731 models. The training performance of each combination is shown in Figure 9. Although it is clear many of the models in the library perform well, some perform very poorly. Since these 14 hyperparameter combinations were found to be the best of 231 options and selected for inclusion in the hyperparameter ensemble we can conclude that in their initial training, in the previous step, they had a lucky random weight initialization. This highlights the importance of training each hyperparameter combination over many different weight initializations - without following this procedure, if a single well-performing model had been selected for production, the results may have been catastrophic.



A histogram showing the training performance of eachof the 14 hyperparameter combinations.

Figure 9. Hyper-Deep Library.

Since it is apparent that not all of the hyperparameter combinations perform well over all random weight initializations the library is culled before the next ensemble construction. Specifically, only the three best-performing hyperparameter combinations are retained, with many weight initializations, for a total of 382 trained models in the final library. Figure 10 shows the training performance for the final hyper-deep library of models. It is interesting to note that the hyperparameter combinations of the three best models are more consistent. There are still 2 different optimizers, Adam and SGD, but there are now only 2 network architectures, and a single loss function – MAE.

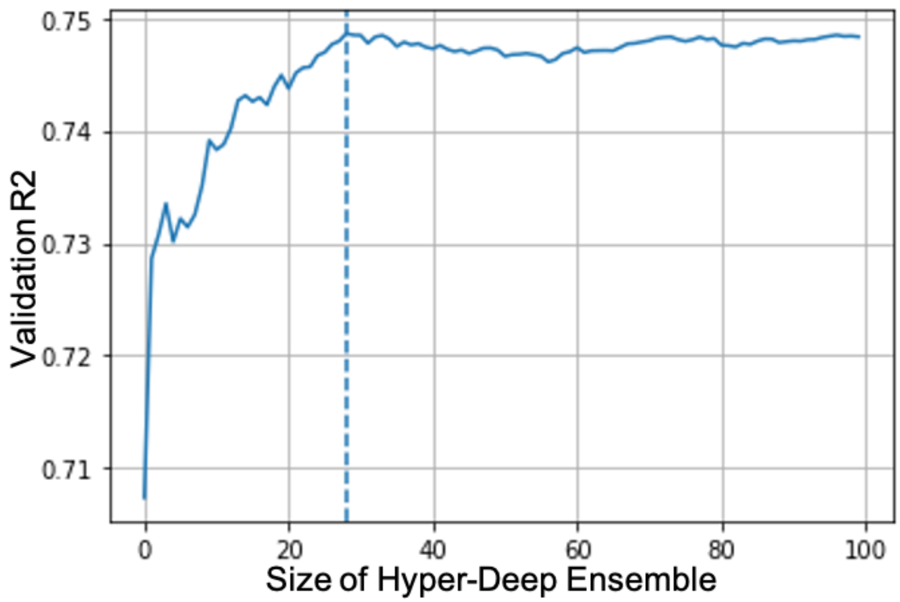


A histogram showing the performance of the 3 bestperforming hyperparameter combinations.

Figure 10. Culled Hyper-Deep Library.

A final greedy construction algorithm is run on the 382 models in the hyper-deep library to produce the final resulting ensemble. The validation set results of this process are shown in Figure 11. This ensemble is much more diverse than the previous step's ensemble. Where the last greedy ensemble construction algorithm only selected 14 different hyperparameter combinations, the hyper-deep ensemble has selected 70 unique models for an ensemble size of 100. However, the optimal ensemble size is found to be 29 models as shown by the dashed line in Figure 11. The increased diversity of models selected for the ensemble improves performance by combining many complementary results, making predictions more robust.

With the final hyper-deep ensemble constructed, Figure 12 shows the first look at the hold-out set ensemble results. The plot displays the test performance vs the ensemble size and a dashed line showing the ensemble size of 29 models as determined on the validation set. Aside from a spike at 2 models, the selected ensemble size is also the optimal value for the test set. Note that even though the performance on the test set is better with fewer models it is still not optimal to select too few models for the ensemble because that will provide a less robust uncertainty calculation in Section 4.4. It is also important to note that, on the test set, the hyper-deep ensemble performs better than any individual model by about 6.5%.



The validation r-squared vs the size of the hyper-deep ensemble. The optimal ensemble size of 29 models is shown with the dashed line.

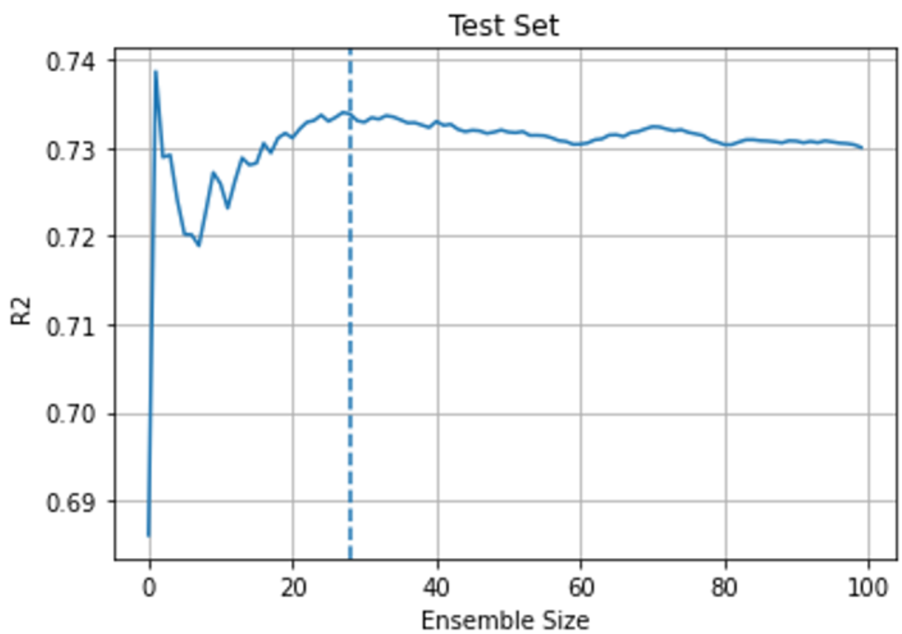
Figure 11. Hyper-Deep Ensemble Construction (Validation Set).

4.4 UNCERTAINTY ESTIMATION

With the hyper-deep ensemble created in Section 4.3, instead of an individual prediction the output of inference on a new sample is a distribution of predictions from the 29 models in the ensemble. Predictions follow a Gaussian distribution, which then provides a straightforward method of uncertainty estimation via the sigma parameter from a fit. Figure 13 shows two prediction distributions on the test set with Gaussian fits. Figure 13(a) has a very consistent distribution with low sigma indicating a low uncertainty estimation, while Figure 13(b) has a very wide distribution with a higher sigma indicating a large uncertainty estimation. As expected, both distributions show a fit consistent with a Gaussian distribution. Rather than determining an overall model-level uncertainty, calculating uncertainty in this way for individual predictions allows for the model results to be carefully interpreted and acted upon on a prediction-by-prediction basis.

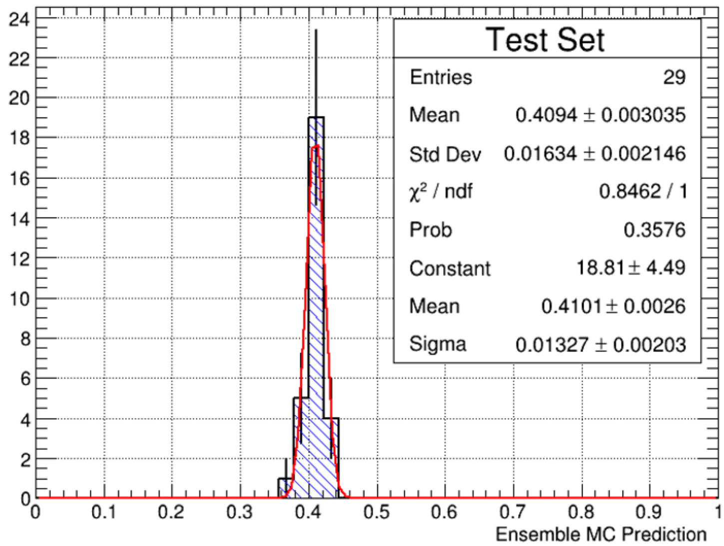
The final step is to calibrate the uncertainty estimations. While it is clear that the uncertainty estimation from sigma gives an indication of how well the predictions from the models in the ensemble agree, it does not indicate how the predictions compare to the actual target value. To calibrate the uncertainty, the sigma values are multiplied by two coefficients. The coefficients are calculated such that about 68% of validation set values fall within one band, and about 95% of validation set values fall within the outer band. The same calibrated coefficients are applied to the test set. This method preserves the information from the hyper-deep ensemble prediction distributions while correlating the results to an actual value and allows for traditional interpretation of the results.

When the same uncertainty calibration coefficients are applied to the test set the results are consistent with those from the validation set with about 64% of the test set predictions falling within $\pm 1\sigma$, and about 93% of the test set predictions within $\pm 2\sigma$. The hyper-deep ensemble predictions, uncertainty, and actual values for the test set are shown in Figure 14. In Figure 14(b) it is apparent that, as expected, the uncertainty estimations are smaller when the predictions are accurate.

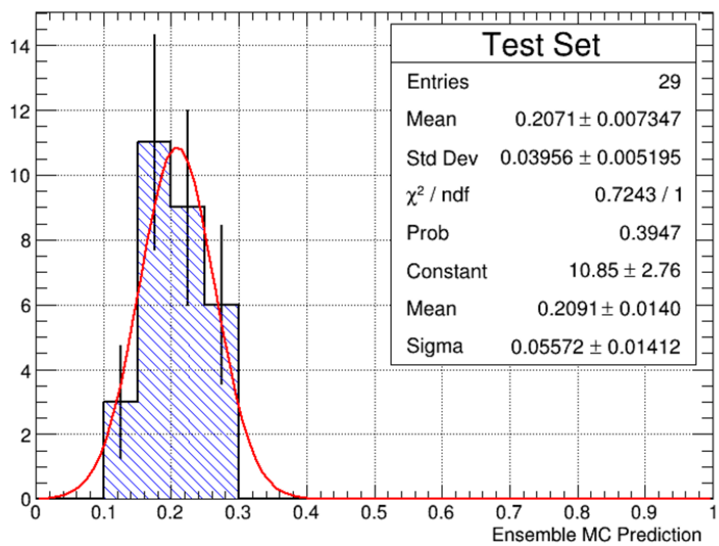


The test set r-squared vs the size of the hyper-deep ensemble. The optimal ensemble size, as identified with the validation set, of 29 models is shown with the dashed line.

Figure 12. Hyper-Deep Ensemble Construction (Test Set).

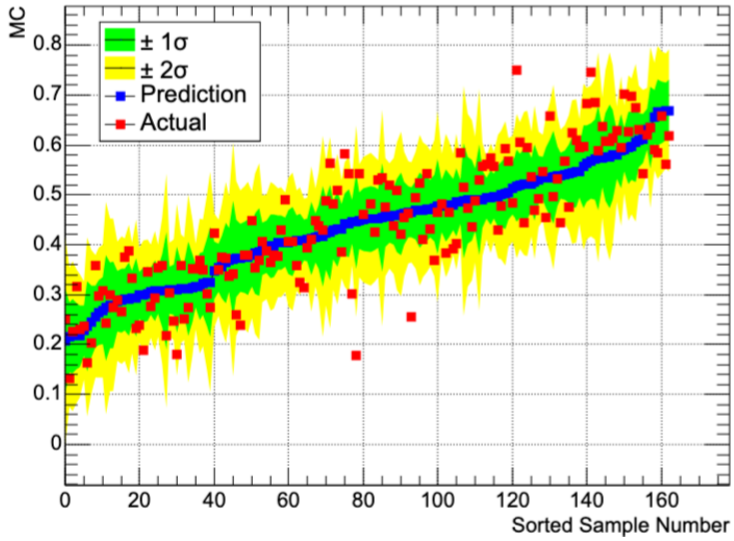


(a) The prediction shows low uncertainty.

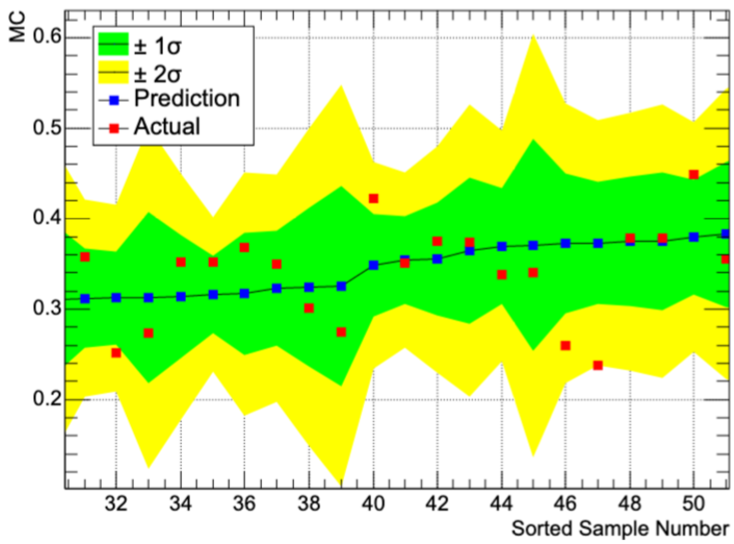


(b) The prediction shows high uncertainty.

Figure 13. Hyper-Deep Ensemble Prediction Distributions.



(a) Complete test set results.



(b) Zoomed-in results show behavior of predictions and uncertainties.

Figure 14. Hyper-Deep Ensemble Results with Prediction Uncertainty Estimation.

This page is intentionally blank.

5. CONCLUSIONS AND FUTURE WORK

State-of-the-art deep learning and ensemble techniques have been used to accurately predict the number of mission capable aircraft that a Naval FA-18 squadron will have up to 3 months in the future. Specifically, attention mechanisms were used to enhance a long short-term memory network to improve an existing deep learning model, and a hyper-deep ensemble was used to further improve model performance and derive robust prediction uncertainty estimations.

Future work includes incorporating all results and insights into an interactive dashboard for use by Commander, Naval Air Forces (CNAF) and the Naval aviation community. Further model improvements are also being made by utilizing Natural Language Processing (NLP) and Topological Data Analysis (TDA) for feature construction. Additionally, the remaining pieces of DARTE are also undergoing improvements including the XAI engine and the flight hour prediction model. Finally, although predictions further into the future have a drop in prediction performance, there is an emphasis for increasingly granular predictions up to the daily level.

This page is intentionally blank.

REFERENCES

1. D. Larter, “The US Navy’s fight to fix its worn-out super hornet fleet is making way,” 2018.
2. V. Insinna, “Boeing super hornet program gets second life through future sales and upgrades,” 2018.
3. D. Larter, “Flush with cash, the Navy bores in on aviation readiness amid a crisis,” 2018.
4. Mehta, “Mattis orders fighter jet readiness to jump to 80 percent - in one year,” 2018.
5. D. T. Blasius and D. D. Doll, “The rise of man hours per flight hour,” tech. rep., CNA, 2018.
6. Michlin, R. Chang, R. Cruz, J. Duclos, D. Lee, V. Siu, and C. Yetman, “Predicting fa-18 squadron readiness and quarterly flight hour execution using machine learning,” tech. rep., Naval Information Warfare Center Pa- cific San Diego United States, 2019.
7. S. M. Lundberg and S.-I. Lee, “A unified approach to interpreting model predictions,” in Advances in Neural Information Processing Systems 30 (I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vish- wanathan, and R. Garnett, eds.), pp. 4765–4774, Curran Associates, Inc., 2017.

This page is intentionally blank.

INITIAL DISTRIBUTION

84310	Technical Library/Archives	(1)
53424	B. Michlin	(1)
71740	J. Duclos	(1)
53424	G. Williams	(1)
53424	D. Lee	(1)
71780	A. Sabater	(1)
53424	J. Rorie	(1)

Defense Technical Information Center
Fort Belvoir, VA 22060–6218 (1)

This page is intentionally blank.

REPORT DOCUMENTATION PAGE

Form Approved
OMB No. 0704-01-0188

1. REPORT DATE *(DD-MM-YYYY)*
February 2022

2. REPORT TYPE
Final

3. DATES COVERED *(From - To)*

4. TITLE AND SUBTITLE

Improving the Digital Aviation Readiness Technology Engine (DARTE) with Temporal Pattern Attention Mechanisms and Hyper-Deep Ensembles

5a. CONTRACT NUMBER

5b. GRANT NUMBER

5c. PROGRAM ELEMENT NUMBER

5d. PROJECT NUMBER

5e. TASK NUMBER

5f. WORK UNIT NUMBER

6. AUTHORS

Dr. Benjamin A. Michlin
Josh A. Duclos
Gary R. Williams
NIWC Pacific

Dean Lee
Dr. Andrew B. Sabater
Dr. Jamal T. Rorie
NIWC Pacific

7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)

NIWC Pacific
53560 Hull Street
San Diego, CA 92152–5001

8. PERFORMING ORGANIZATION REPORT NUMBER

TD-3414

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

NIWC Pacific
53560 Hull Street
San Diego, CA 92152-5001

10. SPONSOR/MONITOR'S ACRONYM(S)

NIWC PAC NISE

11. SPONSOR/MONITOR'S REPORT NUMBER(S)

12. DISTRIBUTION/AVAILABILITY STATEMENT

DISTRIBUTION STATEMENT A: Approved for public release. Distribution is unlimited.

13. SUPPLEMENTARY NOTES

14. ABSTRACT

The Digital Aviation Readiness Technology Engine (DARTE) provides unprecedented predictive readiness capabilities for the Naval FA-18 fleet. DARTE focuses on discovering actionable insights in relation to predicting two key readiness metrics: the number of mission capable (MC) aircraft and flight hours. Recent DARTE efforts have focused on improvements including the adoption of cutting edge artificial intelligence (AI) and deep learning techniques such as temporal pattern attention mechanism-enhanced long short-term memory (LSTMA) networks, hyper-deep ensembles for enhanced performance, and improved uncertainty estimation and robustness. Hyper-deep ensembles and attention mechanisms have been shown to provide state-of-the art results in industry and academia. Furthermore, their improved uncertainty estimation provides decision makers with an increased level of confidence that allows for better, smarter decisions.

.

15. SUBJECT TERMS

DARTE; DART-E; AI; Artificial Intelligence; ML; Machine Learning; Analytics; Aviation; Readiness; LSTM; LSTMA; Mission Capable; MC; Deep Learning; DL; Digital Aviation Readiness Technology Engine

16. SECURITY CLASSIFICATION OF: a. REPORT	b. ABSTRACT	c. THIS PAGE	17. LIMITATION OF ABSTRACT	18. NUMBER OF PAGES	19a. NAME OF RESPONSIBLE PERSON
U	U	U	SAR	36	Benjamin A. Michlin
					19B. TELEPHONE NUMBER *(Include area code)* (619) 553-2925

Standard Form 298 (Rev.
Prescribed by ANSI Std. Z39.18

This page is intentionally blank.

DISTRIBUTION STATEMENT A: Approved for public release. Distribution is unlimited.

Naval Information Warfare Center Pacific (NIWC Pacific)
San Diego, CA 92152-5001