Deep Neural Networks

Enriching Leadership Screening and Selection

Robert A. Newton
Robert A. Masaitis

Artificial neural networks mathematically approximate how human neural cells perceive objects. Machine learning has proven valuable as a predictive tool to inform human decisionmakers, although decision authority cannot be completely ceded to algorithmic predictors due to tendencies in such tools to create inequities or promulgate systemic biases based on race, gender, or other measured categories. The deep neural network tested in the study demonstrated a 94 percent accuracy for candidate selection, suggesting the approach could assist Air Force Special Operations Command (AFSOC) during initial sorting. Employing such a model could free senior leaders from spending valuable time reviewing hundreds of records for attributes specified by the command's developmental team. Senior leaders could then better spend collective time applying knowledge of candidates and squadrons to ensure AFSOC selects high-caliber leaders.

Each year, senior leaders in various career fields across the Air Force meet to select officers qualified to fill leadership positions within their respective communities, such as squadron command.*[1] The Air Force Special Operations Command (AFSOC) leadership selection board, Commando Eagle, specifically considers special tactics officers (STOs) at the ranks of senior captain through lieutenant colonel and rated (flying) officers from special operations aircraft at the ranks of senior major through lieutenant colonel.[2]

The current process is labor-intensive, requiring all board members individually review hundreds of officer records. Human capital is AFSOC's stated competitive advantage.[3] Using a deep neural network to score officer records—one that is tailorable to the attributes identified by the command's leadership team—into an initial, rank-ordered list would allow board members more time for deeper discussions about officers on the margins

Lieutenant Colonel Robert A. Newton, USAF, holds a master of science in flight test engineering from the Air Force Test Pilot School and a master of science in operations management from the University of Arkansas.

Brigadier General Robert A. Masaitis, USAF, deputy director of global integration on the Joint Staff, holds a master of science in military national resource strategy and policy from the National Defense University and a master of science in defense analysis from the Naval Postgraduate School.

* The authors would like to acknowledge and thank the Air Force Special Operation Command's (AFSOC) Commando Eagle team, particularly Jeff McMaster and Thomas Outlaw, for the insights and access they afforded the authors throughout this project. Without their thoughtfulness and willingness to challenge conventions, this would not have been possible.

1. Air Force Personnel Center, "2020 AFSOC COMMANDO EAGLE Candidate Selection Board Results," Personnel Services Delivery Memorandum, Joint Base San Antonio-Randolph, TX, June 30, 2020.

2 AFSOC, *AFSOC Strategic Guidance* (Hurlburt Field, FL: AFSOC, 2020), https://media.defense.gov/.

and for the command's officer development efforts. Providing senior leaders this additional time to consider individual leadership placements as AFSOC develops and employs its force improves the long-term health of the command. Moreover, it enables transformational change as the command organizes to compete across the operational spectrum.

Background

Artificial neural networks mathematically approximate how human neural cells perceive objects. Simply put, the neural network processes the input layer through multiple hidden layers, yielding the output layer, which is a classification or score in this case.[4] With small datasets—fewer than 10,000 samples using fewer than 100 input variables—deep neural networks with more than two layers have demonstrated higher accuracy and better generalization in classification/regression applications than many traditional machine-learning methods such as random forest, support vector machine, or shallow neural networks.[5]

In the context of evaluating individuals, machine learning has proven valuable as a predictive tool to inform human decisionmakers.[6] Yet given their "black box" nature, it is inappropriate to cede decision authority completely to algorithmic predictors.[7] Decisionmakers must be aware of the possibility that such tools, due to a lack of direct interpretability, may create inequities or promulgate systemic biases based on race, gender, or other measured categories.[8]

Cognizant of possible shortcomings, the public and private sectors have applied neural networks to synthesize multidimensional human resource data into something more interpretable.[9] The US Army, for example, has experimented with using automated assistance to manage talent in its personnel assignment process through its People Analytics initiative.[10]

---

3. Jürgen Schmidhuber, "Deep Learning in Neural Networks: An Overview," *Neural Networks* 61 (January 2015), https://doi.org/.

4. Shuo Feng, Huiyu Zhou, and Hongbiao Dong, "Using Deep Neural Network with Small Dataset to Predict Material Defects," *Materials & Design* 162 (January 15, 2019), https://doi.org/; and Antonello Pasini, "Artificial Neural Networks for Small Dataset Analysis," *Journal of Thoracic Disease* 7, no. 5 (May 29, 2015), https://doi.org/.

5. Jon Kleinberg et al., "Human Decisions and Machine Predictions," *Quarterly Journal of Economics* 133, no. 1 (2018), https://doi.org/; and Aaron Chalfin et al., "Productivity and Selection of Human Capital with Machine Learning," *American Economic Review* 106, no. 5 (May 1, 2016), https://doi.org/.

6. Michael Luca, Jon Kleinberg, and Sendhil Mullainathan, "Algorithms Need Managers, Too," *Harvard Business Review* 94, no. 1 (2016), https://hbr.org/.

7. Ruha Benjamin, "Assessing Risk, Automating Racism," *Science* 366, no. 6464 (October 25, 2019), https://doi.org/.

8. Eleni T. Stavrou, Christakis Charalambous, and Stelios Spiliotis, "Human Resource Management and Performance: A Neural Network Analysis," *European Journal of Operational Research* 181, no. 1 (August 2007), https://doi.org/.

9. Kristin C. Saling and Michael D. Do, "Leveraging People Analytics for an Adaptive Complex Talent Management System," *Procedia Computer Science* 168 (2020), https://doi.org/.

And to mitigate pitfalls, it is possible to create a tool based on equity and fairness by understanding potential algorithmic biases and maintaining a human-in-the-loop to overcome them.[11]

Air Force Special Operations Command's stated intent to develop human capital and employ automation to realize efficiencies presents an opportunity to leverage machine learning in the Commando Eagle process.[12] Using automation would provide more time for the panel to focus on a smaller group of leadership candidates and hold a richer discussion, theoretically resulting in improved squadron leadership selection matches.

Further, a process that is bias-aware could better inform command leaders of systemic disadvantages to any demographic groups. This article examines the development and testing of a process to automate the labor-intensive portion of the scoring procedures, freeing time for senior leaders to employ their collective experience, candidate knowledge, and judgment—the best use of the command's senior-most human capital—in evaluating potential officer candidates.

The Commando Eagle Panel

The Commando Eagle panel consists of colonels and general officers from across the command—approximately 15 officers in total—who review and score the personnel record of every eligible officer over a period of several days.[13] Relevant core Air Force Specialty Codes include special tactics officers, special operations pilots, combat systems officers, and remotely piloted aircraft operators. Communities represented include special tactics, AC-130, MC-130, U-28, combat aviation advisors, CV-22, nonstandard aviation, data-masked, and remotely piloted aircraft, with the highest representation of eligible officers from the AC-130 and MC-130 communities (73 and 89 officers, respectively, for 2020).

Every Commando Eagle panel member scores each eligible officer's personnel record—a dense collection of duty history, training and performance reports, and decorations received over a career spanning 14 to 18 years—on a scale of 6 to 10. Panel members use criteria defined by AFSOC in the scoring guide provided at the beginning of the selection board regarding the depth and breadth of an officer's experience, education, training, and leadership.[14]

If any panel member scores a record with a difference of two or more points from any other panel member, the "split" in scores is resolved by discussion. The board members

10. David Anderson, Margrét Vilborg Bjarnadóttir, and David Ross, "There Are No Colorblind Models in a Colorful World: How to Successfully Apply a People Analytics Tool to Build Equitable Workplaces" (1st place paper, White Paper Competition, Wharton People Analytics Conference, University of Pennsylvania, Philadelphia, 2021), 10, https://wpa.wharton.upenn.edu/.

11. AFSOC, *Strategic Guidance*.

12. Brandon Webster, Thomas Outlaw, and Nicole Whigham, "2021 SOF Developmental Team Outbrief," AFSOC, June 16, 2021.

13. Robert A. Masaitis, personal experience, January 20, 2021.

discuss the rationale for their scores then adjust them to resolve the disagreement.[15] After the scoring and adjudication process is complete, the board secretariat finds the average of the panel members' scores for each eligible officer and produces a rank-ordered list of the officers reviewed. Based on the projected number of vacant leadership positions, plus a multiplication factor to allow for attrition, AFSOC derives an at-target number of leadership candidates, which becomes the "cut line" number for the rank-ordered list. The command then further considers the officers above the cut line for projected available leadership positions.[16]

With this list of candidates, the panel spends its remaining time discussing the personnel and identifying potential fits for leadership positions. Often, as each member of the panel must review several hundred officer records, only a portion of the conference's final days is available for this nuanced discussion, and the panel does not have sufficient time to discuss every officer.[17]

Developing a Deep Neural Network

To address this deficit, the authors developed a deep neural network using an existing database from within Headquarters AFSOC's personnel system as input to generate a score for each officer's record on a scale of 6 to 10, similar to the score generated by the Commando Eagle board process.

Rated Officer Analysis Report Database

The command maintains and updates entries in the Rated Officer Analysis Report (ROAR) database for all of its officers. The database includes 177 columns of largely categorical data. The authors evaluated each column for unique values, eliminating redundancy in the dataset by using a correlation matrix (see the parametric reduction below). The dataset was otherwise simple to factorize and required minimal preprocessing for use by the deep neural network.

The ROAR database by design captures career details of rated officers. In tuning the network design, the authors found including special tactics officers—who do not hold an aviation rating but are evaluated by the Commando Eagle board—increased average and maximum error in the network. Because of the differences in career timing, community size, and requisite experiences, STOs are often selected for leadership placement at different points in an officer's career length when compared with rated aviators, which creates inconsistency across the two communities during the board scoring process.

---

14 Department of the Air Force, *Officer Promotions and Selective Continuation*, Air Force Instruction (AFI) 36-2501, incorporating through DAF Guidance Memorandum (DAFGM) 2023-01, January 20, 2023, https://static.e-publishing.af.mil/.

15. Webster, Outlaw, and Whigham, "Team Outbrief."

16. Masaitis, personal experience.

In practice, special tactics officers compete for the leadership of special tactics units, which are generally not led by aviators, and the combination of the two communities in a single board process appears to be for ease of process management. The panel usually reviews STO candidates separately after board scoring to ensure sufficiency for projected vacancies.[18] Consequently, the authors elected to remove the STO community from the dataset for this demonstration, applying the neural network to the 332 rated officers.

Additionally, while rated officers reviewed included both majors and lieutenant colonels, eligible majors were senior and soon to be considered for promotion to lieutenant colonel or awaiting their promotion date, based on career timing (14 to 15 years of commissioned service).

Neural Network Design

Utilizing the advantages of deep neural networks while taking care not to create an overly complex model that could overfit, the authors iteratively developed a topology minimizing degrees of freedom and error.[19] As scores were generated on a continuous scale, they used the rectified linear unit (ReLU) activation function in each layer to avoid the vanishing gradient problem with additional layers.[20] Additionally, the ReLU function can accelerate the training speed of deep neural networks compared to traditional activation functions.

The study used mean squared error as a loss function, similar to regression analysis, and mean absolute error as the metric for the model, as this is intuitively interpreted when reporting how close the model was scoring each officer's record. The study also employed root mean square propagation (RMSProp)—a gradient descent optimization algorithm—as the adaptive learning rate method, since RMSProp does not decay the learning rate too quickly and thus prevents convergence.[21] Finally, while the authors set the maximum number of epochs to 5,000, they used early stopping to avoid overfitting with a loss "patience" of 250 epochs.[22]

---

17. Masaitis.

18. Feng, Zhou, and Dong, "Deep Neural Network"; and Nikolai Nowaczyk et al., "How Deep Is Your Model? Network Topology Selection from a Model Validation Perspective," *Journal of Mathematics in Industry* 12, no. 1 (December 2022): 1, https://mathematicsinindustry.springeropen.com/.

19. Gregory Naitzat, Andrey Zhitnikov, and Lek-Heng Lim, "Topology of Deep Neural Networks," *Journal of Machine Learning Research* 21, no. 184 (2020), https://arxiv.org/.

20. Fangyu Zou et al., "A Sufficient Condition for Convergences of Adam and RMSProp," in *2019 Institute of Electrical and Electronic Engineers (IEEE)/Computer Vision Foundation (CVF) Conference on Computer Vision and Pattern Recognition* (Long Beach, CA: IEEE, 2019), https://doi.org/.

21. Lavanya Shukla, "Designing Your Neural Network: A Step by Step Walkthrough," Towards Data Science, September 23, 2019, https://towardsdatascience.com/.

Model Results and Discussion

Parametric Reduction

Among the ROAR database entries for officers considered by Commando Eagle for command positions, 23 of the 177 columns had no unique values. This number included the historical columns no longer in use and columns that would signal ineligibility and were therefore unnecessary to consider: specifically, whether an officer was currently a commander, had an unfavorable information file, and had a processed date of separation from active duty. Other columns found through the use of a correlation matrix were determined to be redundant. An example of this is drone experience, a binary value intuitively correlated with a nonzero entry in unmanned aerial vehicle type, also captured more broadly by values like aircraft last flown.

After considering these redundancies and nonunique columns, the authors eliminated a total of 102 columns as inputs for the neural network. They also removed demographic data as inputs to later use to evaluate the model for bias according to gender, race, and Hispanic/Latino designation. The remaining 75 columns from the ROAR database were inputs in the neural network, followed by two hidden layers each with 53 neurons, and a single output neuron for the predicted score. The study authors arrived at this topology through experimentation, increasing the number of neurons to increase accuracy but without adding so many degrees of freedom to cause overfitting.[23]

Model Accuracy

The authors evaluated the proposed network using a test set of 83 officer records not used in training—25 percent of the dataset. The mean absolute error of the model in this test set was 0.145 points with a maximum error of 0.666 points. The network converged in 1,915 epochs, stopping early to prevent overfitting.

Figure 1 shows the relationship of actual scores from the test set with the predicted scores from the model. As previously mentioned, projected leadership vacancies drive the number of officers selected as candidates, and in 2020, this was 160. While the minimum actual score for candidates selected was 8.230 points, the minimum predicted score required to ensure 160 officers selected was 8.182. With both the actual and predicted officers selected as command candidates, the authors generated the confusion matrix for the test set shown in table 1. From this confusion matrix, they determined the accuracy of the model to be 0.940, precision and specificity both 1.000, and sensitivity as 0.884—based on the five false negatives and zero false positives in the test set.

The vertical dashed line in figure 1 represents the minimum score required for an officer to be on the list and considered as a candidate for squadron command (8.230 points), while the horizontal dashed line represents the minimum score generated by the neural

22. Naitzat, Zhitnikov, and Lim, "Deep Neural Networks."

network (8.182 points). The cutoff score for both ensured 160 rated officers were on the list (43 of whom were in the test set), based on forecasted personnel requirements. The solid black line represents a perfect fit (actual scores equal to predicted scores), and the red arrow denotes the maximum error in the test set (0.666 points).

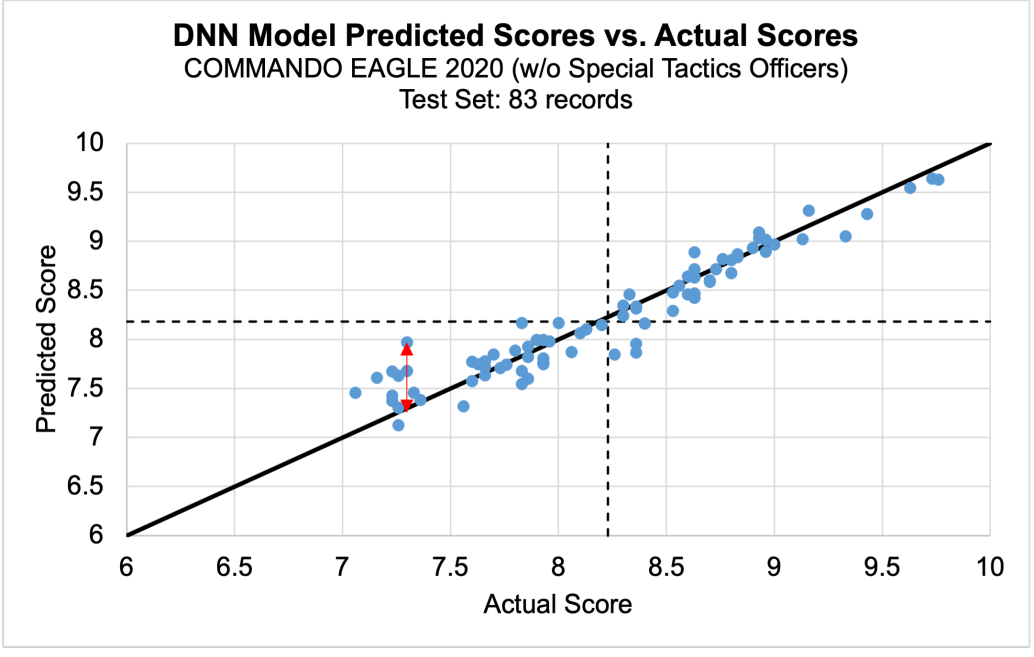

Figure 1. Predicted scores from the deep neural network versus actual scores from the Commando Eagle board

Table 1. Confusion matrix of the 83 records in the test set with actual Commando Eagle 2020 results versus the predicted top officers from the neural network

		Predicted with Neural Network	
	Population = 83	Selected	Not Selected
Actual Board Results	Selected	True Positive = 38	False Negative = 5
	Not Selected	False Positive = 0	True Negative = 40

Given the limited dataset, the authors repeated the fitting of the deep neural network model for over 100 iterations to account for the small size of the randomly selected test set. All but one of the iterations converged, while the remaining iteration was divergent and terminated early after 258 epochs. The maximum error for 99 of the 100 runs was 0.920 points or less with 68 runs less than 0.749 points. The mean absolute error for 99 runs was 0.185 points or less with 95 runs less than 0.175 points, suggesting the use of a deep neural network for Commando Eagle is a repeatable process.

Importance of Factors

Demographic Factors

While the gender, race, and Hispanic/Latino designation factors were not provided as inputs to the deep neural network, the authors tested their significance on the scores using analysis of variance (ANOVA). The dataset, however, was unbalanced demographically in favor of white males who did not identify as Hispanic or Latino (209 of the 332 officers). For example, only 13 females were evaluated, and only two of those females were nonwhite. Only eight officers identified as Hispanic or Latino, and none of those eight were female. As a result, the authors aggregated race to white versus nonwhite for statistical power and did not evaluate intersections with Hispanic/Latino designation.

After the aggregation of race to white versus nonwhite due to sample size, none of these factors were found significant in the actual data or the predicted model nor was the interaction of gender and race (admittedly limited with only two nonwhite females). These results are encouraging in that neither the actual scores nor the model scores appear biased for or against reported demographics. But a dataset with more representation from groups other than male, white, and not Hispanic or Latino would provide greater confidence in this conclusion, particularly with more females.

Nondemographic Factors

The authors also examined the nondemographic data provided to the network to determine significance in the final score assigned to an officer. In the model-predicted scores and actual scores, the study found significant factors in early promotion selection (below-the-zone promotion), year group, weapons school, community, aircraft, core Air Force Specialty Code, professional military education (PME), PME method, academic degree, duty command level, and commissioning source. The authors also looked for interactions informed by the scoring guidance the panel received, with significant interactions between community, aircraft, and Air Force Specialty Code, as well as PME and PME method.

Next, the authors ran regression to determine the relative importance of individual categorical factors in the dataset. Because these categorical data are "one-hot" encoded, they could use the regression coefficients as indicators of importance; that is, a larger coefficient represented a more important factor. The significance of commissioning sources and the communities to which officers belonged revealed by the ANOVA were noteworthy, as these do not necessarily indicate officer performance but are correlated with higher scores. The model predicts officers commissioned through the US Air Force Academy to have slightly better scores on average than officers commissioned through the Reserve Officer Training Corps, the reference population.

The most negative coefficient of the factors found significant was the performance of officers commissioned through the Air National Guard. Additionally, the positive coefficients of every community listed suggest that the reference community, AC-130, on

average, received poorer scores than peers from other communities. Even when considering the different aircraft variants of the AC-130 flown by this community—the interaction mentioned above—the differences were generally not large enough to overcome the performance of the other communities that included remotely piloted aircraft, data-masked, combat aviation advisors, U-28, MC-130, and CV-22.

Not surprisingly, based on the knowledge of the scoring criteria, the negative coefficients associated with completing professional military education in correspondence and only having completed Squadron Officer School, versus intermediate developmental education—Air Command and Staff College or equivalent—disadvantaged officers on average. The early promotion selection was the reference population—22 officers in the dataset—and not being selected below the zone had a negative coefficient. This makes sense if one intuitively assumes officers selected for early promotion perform at higher levels and have higher board scores on average.

Conclusion

A deep neural network yielded a command candidate list with 94 percent accuracy and precision, a mean absolute error in scores of 0.145, and a maximum absolute error of 0.666 points. Even with a small dataset with which to train and test, the authors found these results repeatable. In examining for bias among demographic groups, neither the board nor the model exhibited biases, but these results were based on analysis with limited representation from nonwhite and female officers. Including data from multiple years' boards would further increase a network's capacity to model board scores and examine for demographic bias.

That the neural network closely predicted the scores of the actual board suggests AFSOC could find efficiencies in their decision-making/command selection processes by supporting the Commando Eagle panel with a deep neural network. The goal of the Commando Eagle board is to identify the command's most capable officers for future leadership roles. This board and its companion process, the major command developmental team panels, ultimately provide vectors for officers' careers and professional advancement, yet the process of scoring hundreds of individual records currently consumes much of both panels' time.

A hybrid approach, where a deep neural network provides decision support, could give time back to decisionmakers that they can use to create more meaningful, nuanced vectors for the officers evaluated. Further, if future boards manually scored only a subset of 10 to 20 percent of candidates, the command could validate the model while still benefiting from a less labor-intensive process than the current board practice.

To be clear, this proposed process, while bias-aware, does not remove bias from the board. Instead, it provides a method with which to detect potential biases from the human decisionmakers it is attempting to model. Similarly, it is not a replacement for personal knowledge of officers' experiences, skills, behaviors, and talents. This level of insight comes from the human decisionmakers the process intends to support.

Future Work

Air Force Special Operations Command has held subsequent Commando Eagle boards and added additional screening events to its command candidate selection process since generating the data for this review. Adding the actual results from those boards to create a larger dataset with potentially more diverse categorical options as inputs could build an even stronger model for future use. Moreover, it could incorporate any new variables derived from additive screening events such as physical, cognitive, and psychological evaluations. Beyond that, validating the model using current year results trained off previous years' data would further demonstrate utility to the command. Through this practice, senior leaders could become more accustomed to interacting with and applying machine learning, adapting these techniques as needed.

As demonstrated by the authors' removal of special tactics officers from the dataset to reduce model error for aviation-rated personnel, other personnel databases may be more appropriate for nonrated officers. Experimenting with other available personnel databases that capture enough significant details for a wider array of career fields could allow the support from neural networks on a larger scale.

The authors intentionally did not include any natural or human language interpretation in this analysis, looking instead for categorical predictors in officers' records. The language of officer performance reports is often cryptic, particularly regarding stratification statements (e.g., "#1/XX Majors") as guidelines for such statements are inconsistent year after year. A fully informed model for assigning value and weight to stratification statements would be a daunting undertaking in itself. Analysis that included these human language data—potentially scrubbing for word clusters based on previous successful raters and commanders—could enrich a decision-support tool, providing even more nuanced information to decisionmakers in a process the command is already seeking to improve.[24]

Nevertheless, the model's ability to achieve 94 percent accuracy without these data indicates as valid the study's assumption that these discriminators are not independent of other indicators such as below-the-zone promotion and PME method, and that categorical data can adequately provide an initial stratification to the board. ✈

23. AFSOC Public Affairs, "AFSOC Launches Improved Command Screening Process," AFSOC (website), February 10, 2022, https://www.afsoc.af.mil/.

Disclaimer and Copyright

The views and opinions in Air & Space Operations Review (ASOR) are those of the authors and are not officially sanctioned by any agency or department of the US government. This document and trademarks(s) contained herein are protected by law and provided for noncommercial use only. Any reproduction is subject to the Copyright Act of 1976 and applicable treaties of the United States. The authors retain all rights granted under 17 U.S.C. §106. Any reproduction requires author permission and a standard source credit line. Contact the ASOR editor for assistance: asor@au.af.edu.