Military Rank Prediction: Machine Learning Techniques and GenAI Application

Master of Quantitative Economics, UCLA

Yilong Liu

Faculty Advisor: Professor Denis Chetverikov

June 6th, 2025

Contents

1 Abstract

The facial dominance of West Point cadets, as assessed from their graduation portraits, has been shown to correlate with subsequent military rank attainment [Mueller & Mazur, 1996].With recent advances in generative artificial intelligence—particularly the development of models such as Meta's LLaMA 3.0 via Ollama—and modern machine learning techniques, it is now possible to conduct such analyses in a more scalable and cost-effective manner. This study explores the potential of generative AI in social science research by employing AI-generated facial dominance scores as predictive features. We apply a range of machine learning models to evaluate the predictive power of these scores and compare the resulting classification performance. Our findings suggest a nonlinear relationship between facial dominance and future military rank, and demonstrate the substantive contribution of AI-generated facial ratings to model accuracy: excluding these ratings leads to a measurable decline in predictive performance in nonlinear models.

2 Introduction

This study is inspired by the seminal work of Mueller and Mazur (1996), who examined the relationship between the facial dominance of West Point cadets, as assessed from graduation portraits, and their eventual military rank. Building on their influential research, this paper advances the analysis by integrating recent developments in generative artificial intelligence and machine learning techniques. Our objective is to develop an improved and scalable predictive framework for estimating the highest attained rank of U.S. Army personnel using publicly available data.

The primary goals of this paper are twofold. First, we aim to demonstrate the applicability of generative AI, in combination with modern machine learning algorithms, to empirical research in the social sciences. Second, we seek to identify the most effective predictive model for forecasting long-term career outcomes in the military, using facial dominance scores derived from AI and drawing on the framework established by Mueller and Mazur.

The remainder of the paper is structured as follows. Section 3 reviews the relevant literature. Section 4 describes the data sources, preprocessing procedures, and analytical methodologies employed. Section 5 presents the empirical results and key findings. Section 6 discusses the limitations of the study and provides concluding remarks. References and appendices are included at the end of the paper.

3 Literature Review

The literature review proceeds along two distinct but complementary lines. The first track examines the historical development of research on predicting social or behavioral outcomes from facial

features. This line of inquiry dates back to Keating et al. (1981), who found that male individuals were perceived as dominant or submissive based on certain physiognomic traits [Keating, Mazur, & Segall, 1981]. Building on this foundation, Mazur et al. (1984) investigated whether such facial characteristics were associated with social mobility. Analyzing the West Point Class of 1950, they identified a strong correlation between cadets' facial appearance and their military rank while at the academy, along with several weaker associations [Mazur, Mazur, & Keating, 1984].

A major advancement came with Mueller and Mazur (1996), who, leveraging improved data collection and analytical methods, applied ordinal logistic regression to determine whether facial dominance—assessed from neutral-expression yearbook portraits—predicted long-term military promotion [Mueller & Mazur, 1996]. Their results revealed that facial dominance was not significantly associated with mid-career outcomes but was strongly predictive of higher military rank achieved later in life, typically 20 or more years after graduation.

Subsequent studies have extended this line of research into broader domains, examining the relationship between facial features and unethical behavior [Haselhuhn & Wong, 2012], personality inference [Olivola, Eubanks, & Lovelace, 2014], and perceived leadership qualities [Rule & Ambady, 2011]. Despite these contributions, a critical gap remains in the application of generative artificial intelligence (GenAI) to this area. Given the rapid evolution of GenAI technologies, it is both timely and important to assess whether AI-generated facial representations can meaningfully contribute to predictive models of social outcomes. This paper addresses that gap by integrating AI-derived facial dominance measures into machine learning frameworks and evaluating their predictive value and model performance.

The second track of the literature review focuses on the machine learning techniques employed in this study. While Mueller and Mazur (1996) relied on ordinal logistic regression, the high dimensionality and small sample size of our dataset (152 observations and 47 features) necessitate more advanced approaches to variable selection and model regularization. In this context, we draw on the work of Zou, Hastie, and Tibshirani (2006), who introduced L1-regularized and elastic net extensions to principal component analysis (PCA), thereby enhancing the interpretability and sparsity of PCA-based models.

Additionally, we incorporate a bootstrap-based kernel ridge regression (KRR) approach to account for nonlinear relationships and uncertainty in model estimation. KRR, a method grounded in regularized kernel regression theory [Cristianini & Shawe-Taylor, 2000], is combined with bootstrapping for both inference and model averaging [Mukherjee et al., 2003, Zhu & Hastie, 2005]. This enables us to assess the stability of predictive performance under resampled data conditions and provides a robust framework for evaluating the contribution of AI-generated features.

4 Data

4.1 Data Sources

The original aim of this study was to replicate and validate the findings of Mueller and Mazur (1996) using the same dataset employed in their analysis. Their data was compiled from three primary sources: (1) *The Howitzer*, West Point's student yearbook, which includes cadet portraits and indirect indicators of characteristics such as height and athletic ability; (2) the academy's annual *Register of Graduates and Former Cadets* (1964, 1980 editions), which provides official records of military promotions; and (3) responses to a questionnaire distributed in 1990 to 539 men from the West Point Class of 1950, for whom current addresses were available at the time.

However, access to these exact sources presents substantial challenges. The current West Point digital library only permits public access to *The Howitzer* editions published before 1949, rendering it infeasible to replicate the original dataset under present constraints. Given the limited accessibility of the original data and the time limitations of this study, we instead constructed a new dataset based on publicly available information.

Specifically, we focused on portraits of 19th-century U.S. Army personnel that are accessible online and often accompanied by detailed biographical information, including years of military service, final attained rank, and awards received. These data were manually collected from individual profile pages listed under the Wikipedia category *19th-century United States Army personnel*. Our initial dataset comprises 175 individuals.

To ensure data quality and minimize potential biases, we excluded images that were blurry, incomplete, or artistic illustrations rather than authentic photographic portraits. A sample of the photo selection criteria and a preview of the resulting dataset are provided in the following.

Table 1: Data Preview

Name	**Rank**	**Years of Service**	**Award**
Frederic Vaughan Abbott	Brigadier General	1875–1920	Army Distinguished Service Medal
Emory Sherwood Adams	Major General	1898–1945	Distinguished Service Medal
Fred C. Ainsworth	Major General	1874–1912	NA
Frank Herman Albright	Brigadier General	1887–1919	NA
William Herbert Allaire Jr.	Brigadier General	1882–1921	Silver Star

(a) Selected Photo

(b) Portrait not in the form of photos

(c) Unclear photo

Example of Photo Selection

4.2 Data Description and Preprocessing

After collecting portraits of generals, we generate the rating based on the dominance level of each general's picture. In the original study, facial dominance was measured by projecting the yearbook graduate portrait in front of 20-40 judges (usually undergraduate classes), who would independently rate faces on a seven-point scale of dominance-submissiveness (1 = very submissive, 4 = neutral,or undecided, 7 = very dominant). Judges were also instructed that a dominant person tells others what to do, is respected, and is often a leader; submissive or subordinate people are not influential or assertive and are usually directed by others. According to the description, we let the generative AI check each portrait 100 times and take the average value as the rating of dominance for this personnel. More details regarding the AI model used in our research and how to deploy it on a personal laptop can be found in the Appendix. Since the original paper did not round up the numbers and chose to use the continous rating as a predictor, this paper respects this typical social science practice and keeps the rating as a continuous numerical variable. The prompt we provide to the AI model (Llama 3.0 from Meta via Ollama) is as follows:

"You are an expert in facial dominance analysis of U.S military portraits. You have read the paper: Facial Dominance of West Point Cadets as a Predictor of Later Military Rank by Ulrich Mueller and Allan Mazur in 1996, and the way they used in rating people, rate how dominant or submissive the person appears on this 7-point scale: 1 = Very Submissive 2 = Moderately Submissive 3 = Slightly Submissive 4 = Neutral 5 = Slightly Dominant 6 = Moderately Dominant 7 = Very Dominant"

Following the generation of AI-based facial dominance ratings, most portraits of generals were successfully evaluated 100 times. To ensure reliability, we excluded all observations with fewer than 10 valid ratings. We then constructed indicator variables for each type of medal awarded during military service. In total, 41 distinct medal types were identified, resulting in 41 corresponding binary indicator variables. This high dimensionality poses a significant modeling challenge, given the relatively small sample size ($n = 152$, $p = 47$).

Furthermore, although the initial dataset included multiple final rank categories, some categories had very few observations. To maintain statistical power and comparability, we retained only those individuals whose final rank was either Brigadier General or Major General. An additional data-cleaning step was applied to remove rows with missing values.

The resulting dataset consists of 152 observations and 47 columns. These include: a column for the individual's name, a column for final rank and its corresponding binary indicator, a column for the number of years served, a column for the average AI-generated dominance rating, a column indicating the number of valid AI ratings obtained, and 41 binary medal indicator columns. A preview of this dataset is presented in Table 2, and summary statistics for the numerical columns—excluding the medal indicators—are provided in Table 3.

Table 2: Summary of Data After Preprocessing

Name	**Final Rank**	**Years**	**Rank Code**	**Medals**	**Avg. Rating**	**Valid Ratings**
Frederic Vaughan Abbott	Brigadier General	45	0	...	4.12	91
Emory Sherwood Adams	Major General	47	1	...	4.00	92
Fred C. Ainsworth	Major General	38	1	...	4.22	85
Frank Herman Albright	Brigadier General	32	0	...	5.04	67
William Herbert Allaire Jr.	Brigadier General	39	0	...	5.01	83

Table 3: Summary Statistics for Key Variables

Statistic	**Average Rating**	**Years of Service**	**Valid Ratings (n)**	**Number of Medals**
Count	152	152	152	152
Mean	4.63	36.11	84.47	1.87
Std. Dev.	0.50	7.31	9.29	1.91
Min	2.61	1.00	49.00	0.00
25th Pctl	4.24	33.00	80.75	0.00
Median	4.74	39.00	86.00	1.00
75th Pctl	5.00	41.00	91.00	3.00
Max	6.11	50.00	100.00	10.00

4.3 Methodology

To investigate the potential relationship between AI-generated facial rating and military rank, we applied multiple methods and compared their performance based on prediction accuracy. To ensure the robustness of our calculation and check the consistency, every method we discussed is conducted with a bootstrap-based data set to select the model with the best performance.

To begin with, we assume a linear relationship between AI-generated facial dominance level and military rank based on the previous study in 1996. Mueller and Mazur applied a logistic linear regression model in their paper, and we take a similar approach as there are only two types of outcomes in our study. We first fit a binary logistic linear regression model between their military achievement and AI-generated ratings, and then control the number of years generals have served in the army and the number of medals they have achieved. In the following models for this scenario, Y_i is a binary variable representing the type of final rank, and we set $Y_i = 0$ if the person's final rank is Brigadier General and $Y_i = 1$ if the final rank is Major General; x_i in the equation represents the AI-generated rating for facial dominance level; z_1 and z_2 are the two control variables representing the number of years served and the number of medals generals have achieved.

$$Pr(Y_i = 1|x_i) = \frac{exp(\beta_0 + \beta_1 x_i)}{1 + exp(\beta_0 + \beta_1 x_i)}$$

$$Pr(Y_i = 1|x_i) = \frac{exp(\beta_0 + \beta_1 x_i + \gamma_1 z_{i1} + \gamma_2 z_{i2})}{1 + exp(\beta_0 + \beta_1 x_i + \gamma_1 z_{i1} + \gamma_2 z_{i2})}$$

After constructing the dataset, we incorporated all available medal indicators into the analysis. Given the high dimensionality of the medal-related variables, we explored several methods for variable selection. A standard technique for handling high-dimensional data is Principal Component Analysis (PCA). Specifically, we first applied PCA to the matrix of medal indicators, denoted by **M**, in order to reduce dimensionality.

From this, we selected the top k principal components, denoted by $\{m_i\}_{1 \leq i \leq k}$. These components, along with the AI-generated facial dominance rating variable x_i and the number of years served z_1, were used as predictors in a logistic regression model with L1 regularization. It is important to note that while PCA was employed as a preliminary dimension reduction step, the application of L1 regularization was still necessary due to the relatively large number of variables retained after PCA.

$$\mathbf{m}^{\text{PCA}} = \mathbf{M}\mathbf{W}_K$$

Note: W_K contains the top K eigenvectors of $\boldsymbol{M}^\top \boldsymbol{M}$

$$\hat{\boldsymbol{\beta}} = \arg\min_{\boldsymbol{\beta}} \left\{ \sum_{i=1}^{n} \left[-Y_i \log \pi_i - (1 - Y_i) \log(1 - \pi_i) \right] + \lambda \|\boldsymbol{\beta}\|_1 \right\}$$

$$\pi_i = \frac{\exp\left(\beta_0 + \beta_1 x_i + \gamma_1 z_{i1} + \boldsymbol{\gamma}^\top \mathbf{m}_i^{\text{PCA}}\right)}{1 + \exp\left(\beta_0 + \beta_1 x_i + \gamma_1 z_{i1} + \boldsymbol{\gamma}^\top \mathbf{m}_i^{\text{PCA}}\right)}$$

Inspired by [Zou, Hastie, & Tibshirani, 2006], we also applied the L1-regularized PCA (Sparse PCA) to conduct variable selection and computed the prediction accuracy based on ratings, number of years served, and PCA-derived components. The process is similar to the regular PCA method, and the only difference is that the Sparse PCA process also adds the L1-penalty into the generation of principal components.

$$\mathbf{w}_k = \arg\max_{\mathbf{w}} \left\{ \mathbf{w}^\top \boldsymbol{\Sigma} \mathbf{w} - \lambda_{\text{spca}} \|\mathbf{w}\|_1 \right\} \quad \text{subject to } \|\mathbf{w}\|_2 = 1$$

$$\mathbf{m}_i^{\text{SPCA}} = \mathbf{M}_i \mathbf{W}_K$$

$$\hat{\boldsymbol{\beta}} = \arg\min_{\boldsymbol{\beta}} \left\{ \sum_{i=1}^{n} \left[-Y_i \log \pi_i - (1 - Y_i) \log(1 - \pi_i) \right] + \lambda \|\boldsymbol{\beta}\|_1 \right\}$$

$$\pi_i = \frac{\exp\left(\beta_0 + \beta_1 x_i + \gamma_1 z_{i1} + \boldsymbol{\gamma}^\top \mathbf{m}_i^{\text{SPCA}}\right)}{1 + \exp\left(\beta_0 + \beta_1 x_i + \gamma_1 z_{i1} + \boldsymbol{\gamma}^\top \mathbf{m}_i^{\text{SPCA}}\right)}$$

After PCA related methods, we also reconsidered logistic linear regression model with manual grouping method. By dividing medals into four types based on their historical credibility and origins, we fit the logistic linear model with final ranks and facial ratings, meanwhile controlling the number of years served in the army, and the number of medals earned for each type. This model takes the similar form with the previous logistic linear model, and the only difference is in the control variables. We add z_2, z_3, z_4 and z_5 to respectively represent the number of medals for each type. The manual classification of medals can be seen in Table 4.

Table 4: Manual Grouping of Military Medals by Tier

Tier	Included Medals
Tier 1 (Valor Awards)	Medal of Honor, Distinguished Service Cross, Navy Cross, Silver Star, Bronze Star
Tier 2 (High Merit and Sacrifice)	Distinguished Service Medal, Legion of Merit, Purple Heart, Oak Leaf Cluster, Public Welfare Medal
Tier 3 (Service and Campaign Medals)	Civil War Campaign Medal, Indian Campaign Medal, Mexican Border Service Medal, Philippine Campaign Medal, Philippine Congressional Medal, Cuban Pacification Medal, Spanish Campaign Medal, Spanish War Service Medal, World War I Victory Medal, War Merit Cross
Tier 4 (Foreign and Diplomatic Honors)	Companion of the Order of the Bath (UK), Croix de Guerre, Czechoslovak War Cross 1918, French Legion of Honor (Commander), Grand Cordon of the Order of the Sacred Treasure, Honorary Knight Commander, Legion of Honor, Order of Leopold II (Belgium), Order of the Black Star (Commander), Order of La Solidaridad (Panama), Order of Leopold, Order of Saints Maurice and Lazarus, Order of St Michael and St George, Order of Wen Hu, Order of the Bath, Order of the Crown, Order of the Dragon of Annam, Order of the Rising Sun, Order of the Star of Africa

After examining models grounded in linear assumptions, we turn to the possibility that the relationship between military rank outcomes and AI-generated facial dominance ratings may be inherently nonlinear. To explore this, we begin by applying the bootstrap method to estimate prediction accuracy using the Random Forest algorithm.

To further assess the impact of modeling assumptions, we compare predictive performance under linear and nonlinear frameworks by implementing Kernel Ridge Regression (KRR), which introduces nonlinearity through kernel functions while maintaining the regularization structure of ridge regression. This allows for a direct comparison of predictive accuracy under differing functional forms.

Finally, to pursue the highest achievable predictive performance, we employ Extreme Gradient Boosting (XGBoost), a highly efficient and scalable implementation of gradient-boosted decision trees. Widely used in both classification and regression tasks, XGBoost offers strong predictive capability and robustness, making it a suitable candidate for capturing complex relationships in our dataset.

To investigate whether AI-generated facial dominance rating plays a significant role in the prediction model, we generated summary tables for each linear model and examined whether the coef-

ficient for rating is statistically significant. However, for non-linear scenarios, it is usually hard to discover the actual form of the model as many machine learning models are considered black boxes only suitable for prediction. We discuss the significance of AI-generated rating by comparing the prediction accuracy of the full model and the same model that only excludes the rating factor.

5 Result and Discussion

5.1 Results based on Linear Assumptions

We begin our empirical analysis by applying a basic logistic linear regression model to examine the relationship between AI-generated facial dominance ratings and the final ranks attained by generals. To account for potential confounding factors, we subsequently include control variables representing objective characteristics, specifically the number of years served and the total number of medals awarded.

The regression results are presented in Table 5. As a robustness check, we also estimate a probit regression model using the same set of covariates. The corresponding results are reported alongside for comparison.

Table 5: Logit and Probit Estimates of Promotion to Major General (Dependent Variable: *rank_code*)

	Logit (1)	**Probit (2)**	**Logit + Controls (3)**	**Probit + Controls (4)**
Rating	-0.5982*	-0.3771*	-0.5684	-0.3391
	(0.345)	(0.215)	(0.368)	(0.223)
Number of Years Served			0.0474*	0.0280*
			(0.026)	(0.015)
Number of Medals			0.3086***	0.1873***
			(0.102)	(0.060)
Observations	152	152	152	152
Pseudo R^2	0.015	0.015	0.088	0.088
Log-Likelihood	-103.77	-103.75	-96.04	-96.11
LLR p-value	0.076	0.075	0.000	0.000

Standard errors in parentheses.

$^{***}p < 0.01$, $^{**}p < 0.05$, $^{*}p < 0.10$

As shown in Table 5, the AI-generated facial dominance rating appears to be a weak predictor of military promotion when used in isolation. However, its statistical significance diminishes once additional objective covariates—namely, years of service and the number of medals earned—are included in the model. Among these factors, the number of medals emerges as the most statistically significant predictor of final military rank.

To further evaluate the predictive utility of the rating variable, we compare the performance of logistic regression models estimated on bootstrapped samples, with and without the inclusion of the AI-generated rating. The results, summarized in Table 6, indicate that incorporating the rating variable yields a modest improvement in average prediction accuracy of approximately 2%. Overall, the logistic model achieves an average prediction accuracy of around 67%.

These results of linear models underscore the importance of properly handling the medal-related variables in constructing the prediction models. However, it comes to our attention that each type of medal is different based on its credibility and origins. Therefore, it is natural to consider that different medals should have a different weight as a predictor in the model and some medals may not be important enough to be considered as parts of predictors. To address this problem, we then applied both regular PCA with L1-penalty for logistic regression and Sparse PCA also with L1-penalty to select variables among medals, and compare the prediction accuracy.

Table 7 and Table 8 display our findings from regular PCA and Sparse PCA. For regular PCA variable selection, we added L1-penalty in logistic regression after generating principal components. Only results that explained total variances from 60 % to 80% are included, results from higher threshold (i.e. 90% and 95%) are not available as too many components were generated and it led to severe perfect collinearity. We find that the while the prediction accuracy continues to increase as the threshold increases, adding rating into the linear model as a predictor does not necessary increase the model performance and may add unnecessary noise as the standard deviation increases. For Sparse PCA, we compare results between different numbers of components retained, and the numbers range from 5 to 12 based on practical experience. The results show that while rating hardly contributes as a predictive factor when the number of components is small ($n \leq 7$), models including rating as a predictor outperform others when the number is larger ($n \in [8,12]$), and including rating generally adds more stability to the model.

Table 6: Bootstrapped Prediction Accuracy and Coefficient Estimates (200 Resamples)

	Logit w/ Rating	**Probit w/ Rating**	**Logit w/o Rating**	**Probit w/o Rating**
Accuracy (Mean)	0.665	0.666	0.643	0.641
Accuracy (SD)	0.039	0.042	0.043	0.038
Accuracy (Median)	0.664	0.668	0.641	0.645
95% Interval	[0.605, 0.730]	[0.598, 0.737]	[0.579, 0.711]	[0.586, 0.698]
Coef: Years Served	0.365	0.226	0.363	0.207
(SE)	(0.204)	(0.115)	(0.219)	(0.105)
Coef: Rating	-0.275	-0.189	–	–
(SE)	(0.184)	(0.117)	–	–
Coef: Num Medals	0.621	0.374	0.617	0.372
(SE)	(0.204)	(0.132)	(0.225)	(0.118)

Table 7: Bootstrapped Prediction Accuracy of PCA-Logit Models With and Without Rating

Threshold	Components Retained	Includes Rating	Mean Accuracy	Std. Dev.
60%	4	Yes	0.713	0.030
		No	0.730	0.035
70%	6	Yes	0.728	0.040
		No	0.730	0.032
80%	9	Yes	0.755	0.048
		No	0.752	0.040

Table 8: Bootstrapped Prediction Accuracy with Sparse PCA Components (Alpha = 1.0)

SPC Components	Rating Included	Mean Accuracy	Std. Dev.
5	Yes	0.712	0.045
5	No	0.708	0.045
6	Yes	0.722	0.049
6	No	0.721	0.040
7	Yes	0.721	0.049
7	No	0.727	0.048
8	Yes	0.747	0.040
8	No	0.736	0.034
9	Yes	0.762	0.039
9	No	0.746	0.058
10	Yes	0.765	0.038
10	No	0.739	0.049
11	Yes	0.740	0.066
11	No	0.755	0.044
12	Yes	0.764	0.070
12	No	0.741	0.079

Following the discussion of PCA-based dimensionality reduction techniques, we next consider an alternative approach by manually grouping medals according to their credibility and origin. We categorize the medals into four tiers and incorporate these groupings into a logistic linear regression model. In this specification, we regress final military rank on the AI-generated facial dominance rating, controlling for the number of years served and the count of medals earned within each medal tier. The regression results are presented in Table 9.

Compared to previous logistic regression models, this specification yields more consistent and statistically significant results. Interestingly, our findings suggest that, at higher levels of military hierarchy, facial dominance may have an adverse effect on promotion likelihood—contrary to the positive association reported in Mueller and Mazur's (1996) study.

Among the four medal tiers, the second and third tiers exhibit the strongest statistical significance in predicting final rank. This is plausible, as these tiers typically correspond to decorations awarded for direct contributions to combat operations and service performance. In contrast, fourth-tier medals tend to represent diplomatic honors exchanged between nations, and are therefore less reflective of individual military achievement. First-tier medals, although highly prestigious, are rarely awarded, resulting in a very limited number of observations and a lack of statistical significance in the regression.

In terms of predictive performance, the average accuracy of the model under this grouped-medal specification increases from 66.5% to 72.2% when evaluated using the bootstrap method. Excluding the AI-generated rating from the model leads to an approximate 2% decline in prediction accuracy, further supporting its marginal contribution to predictive power.

Table 9: Logit and Probit Regression Estimates with Tiered Medal Counts and Rating

Variable	**Logit Coef.**	**Logit p-value**	**Probit Coef.**	**Probit p-value**
Number of Years Served	0.0530	0.036	0.0330	0.029
rating	−0.5596	0.005	−0.3502	0.003
num_medal_tier1	−0.4619	0.213	−0.3261	0.153
num_medal_tier2	1.0456	0.003	0.6811	0.002
num_medal_tier3	1.6095	0.038	1.1452	0.049
num_medal_tier4	0.0604	0.714	0.0372	0.711

5.2 Results based on Nonlinear Assumptions

After considering different models under linear assumption, we keep exploring the possibilities of better prediction under the nonlinear scenario. Table 10 displays our findings with Kernel Ridge Regression models, and it is found that in both linear and nonlinear cases, prediction accuracy is significantly increased after adding rating as a predictor, and especially the radial basis function

model ('RBF', the nonlinear model) out beats the performance of linear model in prediction accuracy, indicating the nonlinear relationship between military ranks and facial ratings. The highest prediction accuracy is achieved by bootstrap-based KRR nonlinear model and is at 86.5%.

Table 10: Bootstrapped Prediction Accuracy of Kernel Ridge Regression Models

Kernel Type	**Includes Rating**	**Mean Accuracy**	**Std. Dev.**
RBF (Nonlinear)	Yes	0.865	0.027
	No	0.818	0.025
Linear	Yes	0.821	0.032
	No	0.804	0.038

Inspired by the indication of nonlinear relationship, we also tested Random Forest Model and XGBoosting model, and the results are displayed in Table 11. From results we find that , the Random Forest model increases its prediciton accuracy from 82.5% to 88.1%, and the XGBoosting model increases its accuracy from 88.1% to 95.4%. Both models show substantial increase in prediction accuracy and the model stability as the standard deviation decreases after adding rating into the model, and it is within our expectation that the XGBoosting model outperforms Random Forest model as XGBoosting is famous for its high prediction accuracy in machine learning . These results are consistent with previous discussion about significant nonlinear relationship between facial rating and military ranks, and it suggests that rating carries meaningful nonlinear signal that is better captured by Kernel Ridge Regression model, Random Forest model and XGBoosting model than linear models.

To summary this section, while AI-generated facial dominance rating appears to have an adverse effect on military promotion, its signifiance is better captured by nonlinear machine learning models, suggesting further investigation on the interaction between facial features and other objective factors in the future research.

Table 11: Bootstrapped Prediction Accuracy of Random Forest and XGBoost Models

Model	**Includes Rating**	**Mean Accuracy**	**Std. Dev.**
Random Forest	Yes	0.881	0.017
	No	0.825	0.031
XGBoost	Yes	0.954	0.015
	No	0.881	0.024

6 Limitation and Conclusion

6.1 Research Limitation

Before any conclusion regarding findings of this point, it is worth pointing out the limitation of this paper and discuss possible ways of improvement in the future. The first limitation of this paper is the difference between prepared dataset and original dataset from 1996' study. While we tried to generate facial dominance level rating through AI in our research and controlled the number of years served and medal related factors, the original data contains more detailed information for each general apart from those information in our study. For example, the number of years it takes for one West Point Cadet to be promoted from one specific position to the next level. What is more, the original dataset contains information for the entire class 1950 of West Point Cadets, while our data only contains personal information of the US generals born in 19th century, which may explain the reason why our estimation of facial features is different from the original paper. Therefore, it would be only appropriate to say we discuss the significance of AI-generated rating in military career prediction insteading of saying our result can be used to compare with the original study, and there is still a gap between our research from validating the original result in a rigorous and comprehensive way.

Secondly, while this paper applies many different methods investigating the relationship between facial dominance rating and military achievement, the small size of dataset has been a main concern and make the prediction result from machine learning method such as XGBoosting, Random Forest, and KRR less reliable even with bootstrapped method.

The third limitation is related to interpretation. While the prediction results are satisfying and the significance of AI-generated rating is proved by comparing results from the same model with and without the rating predictor, we are still not sure about the actual form of the best model as we can only ensure the nonlinear relationship between rating and military rank.

The last but equally important limitation is the strategy used in considering medal columns. While results from logistic regression with manual grouping are significant, a better grouping can be achieved if we can consider the importance of medals based on the exact reason why the medal was awarded to the corresponding personnel with more details of personal information collected.

6.2 Conclusion

This paper is inspired by Mueller and Mazur's amazing work in 1996, proposes the use of generative AI to replace the human judge rating process, and conducts a comprehensive analysis between different machine learning models to discuss the relationship between rating predictor and the best possible prediction accuracy. It provides support in using AI-generated analysis result in social science research, and has shown the potential of generative AI in increasing the efficiency and

lowering the cost for social science research.

In general, methods like regular PCA, Sparse PCA, Kernel Ridge Regression, Random Forest, and XGBoosting are discussed and compared. The results show that for army personnel who have reached a certain level such as general, facial dominance level appears to have a negative, and significant impact on their final military rank and the relationship between them appears to be nonlinear as all nonlinear models have achieved a high prediction accuracy, and displayed the importance of facial rating as a predictor.

Despite the contributions of this study, several limitations remain due to constraints in data availability and time. To more comprehensively address the research question, future studies should aim to collect more detailed and expansive datasets. In particular, researchers are encouraged to obtain access to the original data used in Mueller and Mazur's study—either by visiting the West Point campus archives or by directly contacting the original authors for assistance in validating and extending their findings.

With a larger and richer dataset, future work could incorporate additional control variables, improve model specification, and enhance both the validity and predictive power of the analysis. Moreover, advanced econometric techniques such as the instrumental variables (IV) approach may be employed to address potential endogeneity concerns, provided the requisite data is available. Researchers may also benefit from experimenting with alternative generative AI models and systematically comparing their predictive performance across different model architectures and specifications.

References

Cristianini, N., & Shawe-Taylor, J. (2000). *An introduction to support vector machines and other kernel-based learning methods*. Cambridge University Press.

Eberhardt, J. L., Davies, P. G., Purdie-Vaughns, V. J., & Johnson, S. L. (2006). Looking deathworthy: Perceived stereotypicality of Black defendants predicts capital-sentencing outcomes. *Psychological Science*, 17(5), 383-386.

Haselhuhn, M. P., & Wong, E. M. (2012). Bad to the bone: Facial structure predicts unethical behaviour. *Proceedings of the Royal Society B: Biological Sciences*, 279(1728), 571-576.

Keating, C. F., Mazur, A., & Segall, M. H. (1981). A cross-cultural exploration of physiognomic traits of dominance and happiness. *Ethology and Sociobiology*, 2(1), 41-48.

Mazur, A., Mazur, J., & Keating, C. (1984). Military rank attainment of a West Point class: Effects of cadets' physical features. *American Journal of Sociology*, 90(1), 125-150.

Mueller, U., & Mazur, A. (1996). Facial dominance of West Point cadets as a predictor of later military rank. *Social Forces*, 74(3), 823-850.

Mukherjee, S., Tamayo, P., Rogers, S., Rifkin, R., Engle, A., Campbell, C., & Mesirov, J. P. (2003). Estimating dataset size requirements for classifying DNA microarray data. *Journal of Computational Biology*, 10(2), 119–142. https://doi.org/10.1089/106652703321825928

Olivola, C. Y., Eubanks, D. L., & Lovelace, J. B. (2014). The many (distinctive) faces of leadership: Inferring leadership domain from facial appearance. *The Leadership Quarterly*, 25(5), 817-834.

Rule, N. O., & Ambady, N. (2011). Face and fortune: Inferences of personality from Managing Partners' faces predict their law firms' financial success. *The Leadership Quarterly*, 22(4), 690-696.

Todorov, A., Mandisodza, A. N., Goren, A., & Hall, C. C. (2005). Inferences of competence from faces predict election outcomes. *Science*, 308(5728), 1623-1626.

Zhu, J., & Hastie, T. (2005). Kernel logistic regression and the import vector machine. *Journal of Computational and Graphical Statistics*, 14(1), 185–205. https://doi.org/10.1198/106186005X25620

Zou, H., Hastie, T., & Tibshirani, R. (2006). Sparse Principal Component Analysis. *Journal of Computational and Graphical Statistics*, 15(2), 265–286. https://doi.org/10.1198/106186006X113430

7 Appendix: Introduction of Ollama and Llama3.0

While people are surprised at the fast development of Generative AI and imagine its potential in helping with research and many other tasks, it may not be easy for people with 0 or relative less experience in computer science and artificial AI to deploy and apply Generative AI in their daily research. Ollama, as a user-friendly interface for running large language model locally, can be a valuable tool for researchers, developers, and anyone who wants to experiment with language models. It is easy to download from Ollama's Official Website `https://ollama.com/` which provides detailed intruction in helping users deploying LLMs locally.

After downloading Ollama, it offers a wide range of LLMs for people to choose. While it has made it a lot easier for people to deploy LLMs, some LLMs are still too large in size to be deployed on a personal working laptop. For example, you can now deploy the latest Llama4.0 model developed by Meta company, but it would need 64GB of RAM to operate even with Ollama's help. Therefore, we finally choose an older version, and also the most capable openly available model Llama3.0 to conduct our research. More details about properties of Llama3.0 model can be found on Meta's website:`https://ai.meta.com/blog/meta-llama-3/`.